Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Maestría en Ciencias
en Ciencias de la Computación

Caracterización de arbitraje financiero en el mercado de divisas

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de Maestro en Ciencias

Presenta:

Héctor Luna Armendáriz

Ensenada, Baja California, México

2021

Tesis defendida por

Héctor Luna Armendáriz

y aprobada por el siguiente Comité

Dr. Israel Marck Martínez Pérez
Director de tesis

Dr. Hugo Homero Hidalgo Silva

Dr. Ubaldo Ruíz López

Dr. Miguel Ángel Alonso Arévalo

Dr. Israel Marck Martínez Pérez
Coordinador del Posgrado en Ciencias de la Computación

Dra. Rufina Hernández Martínez
Directora de Estudios de Posgrado

Resumen de la tesis que presenta Héctor Luna Armendáriz como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la Computación.

Caracterización de arbitraje financiero en el mercado de divisas

Resumen aprobado por:

Dr. Israel Marck Martínez Pérez
Director de tesis

El mercado de divisas o FOREX (*Foreing Exchange Market*) es un gran negocio que permite operar a favor o en contra de una moneda desde una plataforma virtual. Los usuarios desarrollan estrategias que buscan predecir tendencias y patrones de cada una de las divisas, para así invertir dinero y obtener ganancias. Una de estas estrategias es el arbitraje triangular, la cual consiste en realizar múltiples compras y ventas simultaneas en el mercado para beneficiarse de las diferencias de precios. Esta estrategia ha mostrado un buen desempeño, pero breves oportunidades debido a la volatilidad de los precios en el mercado. Basado en esta herramienta, se propone caracterizar su comportamiento en combinación con otras estrategias, utilizando indicadores técnicos clásicos para adquirir propiedades, evaluar su efectividad y encontrar más oportunidades para operar con técnicas de aprendizaje máquina. En el presente trabajo de investigación se analiza el par de divisas NZDUSD como par principal, ya que es un par poco estudiado en la literatura, utilizando 10 años de datos históricos (2008-2018) en periodos de minuto. El par NZDUSD se creó por medio de EUR, AUD, GBP, CAD, CHF, y JPY. Además, se modeló cada par como series de tiempo. Posteriormente cada par generado se caracterizó por medio de 33 indicadores técnicos. Los puntos en las series de tiempo se etiquetaron de tres maneras distintas: el primero con el objetivo de encontrar oportunidades de arbitraje triangular, el segundo para determinar el tipo de operación en las oportunidades encontradas con el primer etiquetado y un tercero que combina los dos experimentos anteriores por medio de clasificación multiclase. En cada experimento se detectaron distintos grupos de características, probándose su eficiencia con distintos clasificadores mediante cinco métricas. Los resultados obtenidos muestran que existen al menos cuatro oportunidades de arbitraje triangular al día. Finalmente, desde la generación de características hasta el entrenamiento de algoritmos, el proceso se ejecuta en menos de medio minuto, identificando oportunidades en un lapso de cinco minutos, tiempo suficiente para operar en el mercado.

Palabras clave: mercado de divisas, divisa, arbitraje, arbitraje triangular, precio, tendencia, volatilidad, volumen, indicador, eficiencia del mercado, estrategia de inversión, par de divisas, par sintético, umbral, algoritmo, series, series sintéticas

Abstract of the thesis presented by Héctor Luna Armendáriz as a partial requirement to obtain the Master of Science degree in Master in Computer Science.

Characterization of financial arbitrage in foreign exchange market

Abstract approved by:

Dr. Israel Marck Martínez Pérez
Thesis Director

The foreign exchange market or FOREX is a great business that allows you to trade for or against a currency from a virtual platform. Users develop strategies that seek to predict trends and patterns of currencies to invest money and obtain profits. One such strategy is triangular arbitrage, which consists of making multiple simultaneous purchases and sales to benefit from price differences. This approach has shown a good performance but few opportunities due to the volatility of market prices. Based on this tool, it is proposed to characterize its behavior using classical technical indicators to acquire properties, evaluate their effectiveness, and find more opportunities to operate in the market with machine learning techniques. In this research work, the NZDUSD currency pair is analyzed as the main pair using ten years of historical data (2008-2018) at a one-minute timeframe. The NZDUSD was created using EUR, AUD, GBP, CAD, CHF, and JPY. Additionally, each synthetic pair was modeled as time series and characterized with 33 technical indicators. Data points were labeled in three different ways: the first with the objective of finding triangular arbitrage opportunities, the second to determine the type of operation in the opportunities found with the first labeling, and a third combining the two previous experiments through multiclass classification. Different groups of characteristics were detected in each experiment, and their efficiency tested with different classifiers using five metrics. The results obtained show that there are at least four triangular arbitrage opportunities per day. Finally, from the generation of characteristics to algorithms' training, the process is executed in less than half a minute, detecting arbitrage opportunities in five minutes, having enough time to operate the market effectively.

Keywords: forex market, currency, arbitrage, triangle arbitrage, price, trend, volatility, volume, indicator, market efficiency, investment strategy, currency pair, synthetic pair, threshold, algorithm, series, synthetic series

Dedicatoria

A mis padres por su apoyo incondicional. A mis compañeros por su ayuda, en especial a Rafael, Diego, Tadeo y Mónica. Y aquellos seres amados que estuvieron a lo largo del trayecto mientras emprendí y culminé este camino.

Agradecimientos

Al Centro de Investigación Científica y de Educación Superior de Ensenada por la oportunidad brindada.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de maestría/doctorado. No. de becario: 18264353.

A mi director de tesis, el Dr. Israel Marck Martínez Pérez por su fe en mi y dirección.

Tabla de contenido

Página

Tabla de contenido

Tabla de contenido

Lista de figuras

Figura Página

Lista de figuras

Figura Página

Lista de figuras

Figura Página

Lista de figuras

Figura Página

Lista de tablas

Lista de tablas

Lista de tablas

Lista de tablas

Capítulo 1. Introducción

El Mercado de Divisas es por mucho, el mercado financiero más grande del mundo; estando por encima del mercado de *acciones*, *opciones* y *futuros*. A diferencia de otros mercados financieros, no tiene una ubicación física donde se realizan las operaciones (Mavrides, 1992). Es un mercado de venta libre y se basa en sistemas de telecomunicaciones. Las características únicas de este mercado es que se trata de un mercado global, es continuo las 24 horas del día, los participantes del mercado están ubicados en todo el mundo y se comercializa utilizando los sistemas de telecomunicación más avanzados que puede ofrecer la tecnología. A lo largo de los años, este mercado ha sufrido importantes cambios estructurales. Los cambios que se están produciendo se atribuyen a los desarrollos institucionales, así como a los avances tecnológicos. Esos cambios tuvieron un impacto significativo en varias aspectos del mercado: la forma en que se realiza el comercio, la competencia, el tamaño del mercado y la eficiencia del mismo. Por lo tanto, es necesario proporcionar una actualización del entorno en el que se determinan los precios de las divisas. El cambio más importante en los últimos años es la evolución de un mercado cruzado independiente en el que las tasas no monetarias se determinan en función de la libre oferta y las fuerzas de la demanda. Lo anterior permitió ofrecer más productos y más oportunidades a los comerciantes y, por lo tanto, también aumentó el volumen y su negociación en el mercado. Cada vez que el valor monetario de cualquiera de los tipos de cambio de libre comercio cambia, como resultado de las fuerzas de suministro y demanda, existe una posibilidad potencial para que el comerciante opere sin riesgo. Aquí nace la idea de arbitraje aplicado en monedas o en cualquier otro producto.

1.1. Planteamiento del problema

De manera informal, el arbitraje es una estrategia del mercado de divisas en la que se realizan compras y ventas simultáneas de pares de divisas (Moosa, 2003). La definición formal se encuentra en la sección 2.2.2 Un par de divisas está conformado de dos divisas, como el *NZD* (dólar neozelandes) y *USD* (dólar americano). Al formar el par se tiene el *NZDUSD* que se lee *dólar australiano-dólar americano* (para más detalle de los acrónimos de las divisas veáse los Apéndices). Las oportunidades de arbitraje surgen cuando los tipos de cambio son inconsistentes, o bien, no coinciden los precios cruzados en el mercado, violando un equilibrio. Por ejemplo, dadas dos divisas cualquiera X y Y, el arbitraje (en este caso "de dos puntos") se da en las casas de cambio cuando se vende una cantidad de X a través de Y y se compra de vuelta Y a través de X por la misma cantidad (Fig. 1).

Figura 1. Arbitraje de dos puntos. La compra y venta de una misma divisa con alguna divisa particular tiene como resultado arbitraje de dos puntos.

Ahora bien, si se hacen más transacciones (compras y ventas de divisas), se tiene arbitraje múltiple (lo cual se detalla en el siguiente capítulo). Normalmente, uno no terminaría con la misma cantidad de dinero que empezó al hacer estas operaciones (compra y venta de dinero) por distintos factores (como el cobro por transacción o *slippage*, por ejemplo). Sin embargo, el arbitraje financiero nos permite aprovechar estas deficiencias para nuestro beneficio gracias al equilibrio del mercado. El equilibrio del mercado (Olivera, 1991) es una situación que se da cuando los precios que éste ofrece, a aquellas personas que compran o consumen un bien o servicio, pueden adquirir las cantidades que deseen, y los que ofrecen ese bien o servicio, pueden vender todas sus existencias. A su vez, si llega a haber alguna discrepancia entre precios al realizar operaciones, esta tiende a converger al precio original como se muestra en la Figura 2 por la misma eficiencia y equilibrio del mercado. Así mismo el arbitraje nos permite ver estas diferencias y aprovecharlas para hacer operaciones en el mercado de manera estratégica, en la Figura 3 se muestra como visualizar estas las diferencias. La gráfica se encuentra normalizada para observar que tanto se distancia el precio respecto al original a través del tiempo.

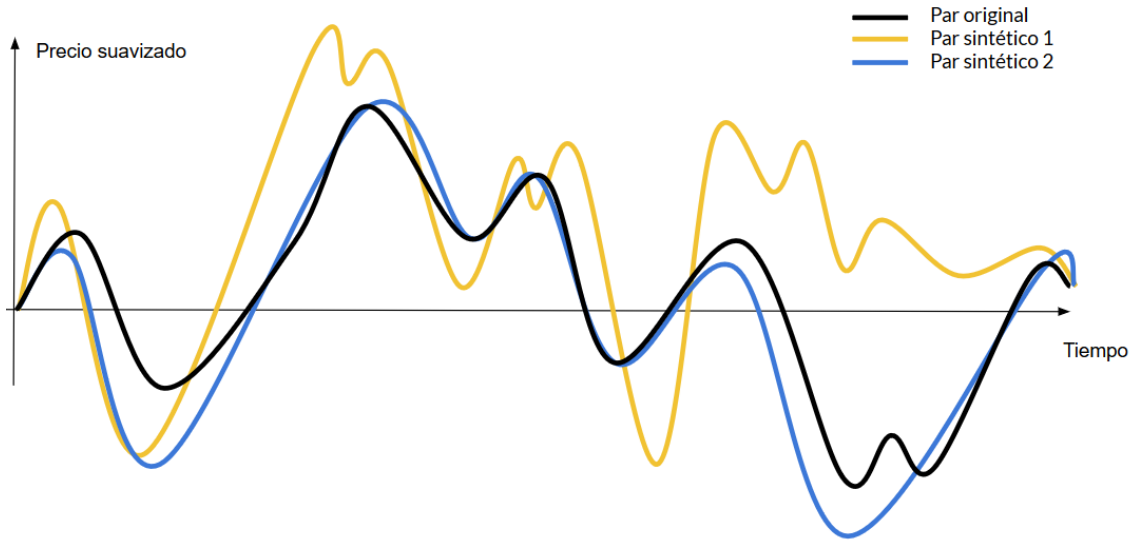

Figura 2. Divergencia y convergencia de precios en el mercado entre una divisa original y una divisa creada a partir de arbitraje (o divisa sintética).

Para aprovechar esta diferencia de precios en el mercado y operar de manera efectiva, lo ideal es comprar barato y vender caro. En la Figura 3 podemos ver dos puntos marcados en los pares sintéticos, estos puntos indican momentos ideales para realizar operaciones. La operación en el punto alto corresponde a la venta, y la operación en el punto bajo corresponde a la compra. Si analizamos esta estrategia gracias a la eficiencia del mercado, ambos precios sintéticos convergerán al precio original, pero para llegar a ese punto se pasarán por cambios. Al tener la eficiencia del mercado a favor, cuando se esté perdiendo en una operación, el dinero perdido se compensará con la ganancia de la otra operación (comúnmente conocido como estrategia de *cobertura*).

En la misma figura se ilustra la existencia de umbrales, dichos umbrales son márgenes de error. O bien, indican a partir de cuando es más seguro operar en el mercado, ya que la incidencia de máximos y mínimos relativos entre pares sintéticos no siempre será precisa. Los umbrales se definen tomando en cuenta conceptos como el *spread*, *slippage*, precios máximos y precios mínimo (para más detalle, véase la Sección 2.2.9).

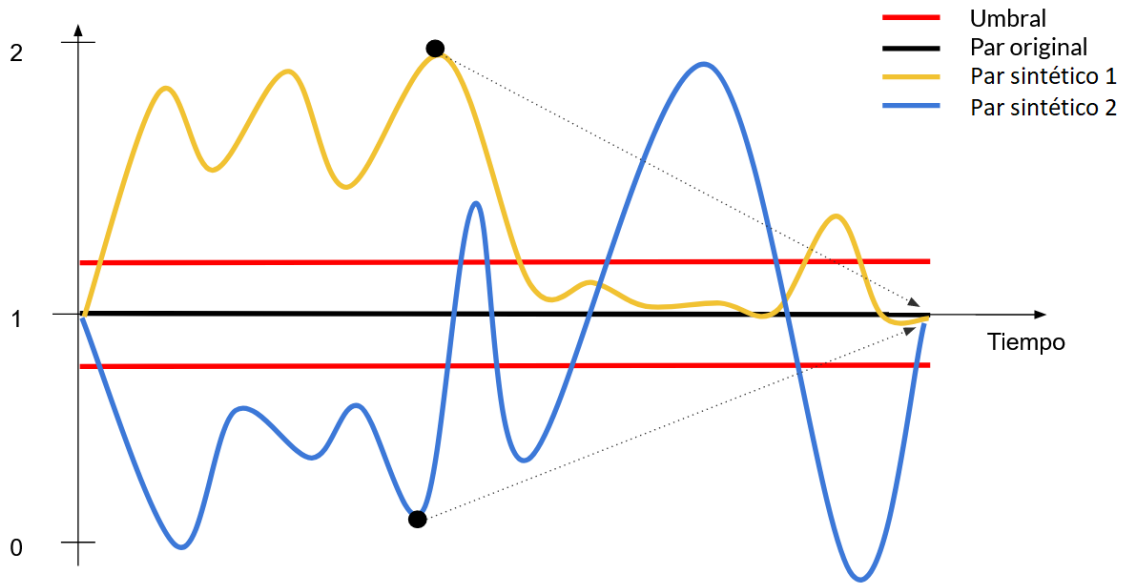

Figura 3. Gráfico de series sintéticas normalizado respecto al par original. Las curvas en la gráfica representan los precios de diferentes pares (sintéticos como el original) y la divergencia o convergencia entre estos creado a partir de arbitraje (o divisa sintética) con el precio normalizado.

El arbitraje estuvo en su mayor auge en la década de los noventas y principios de los dos mil (Aiba *et al.*, 2002), pero poco a poco fue perdiendo popularidad ya que su detección se volvió poco eficiente. Posteriormente se buscó predecir este fenómeno, pero dado el avance tecnológico, esta técnica sólo resultó viable para grandes empresas con un gran poder de cómputo (Cui y Taylor, 2020; Akram *et al.*, 2008; Aiba *et al.*, 2002; Wang *et al.*, 2008; Mavrides, 1992). El tipo de arbitraje estudiado mayormente fue el arbitraje de tres puntos (o también conocido como arbitraje triangular), particularmente el par de divisa EURUSD. Sin embargo, existe una publicación reciente demostrando la existencia de oportunidad de arbitraje que vincula el problema de la detección de arbitraje de *k*-divisas con la identificación del ciclo de longitud máxima en un grafo construido a partir de la matriz de tipo de cambio de precio de venta más alta (Cui y Taylor, 2020).

En esta tesis, se busca corroborar que el arbitraje triangular doble (vease el subcapítulo 2.2.12, para más detalles de este concepto) es eficaz con la tecnología actual en el par NZDUSD. El par de divisas más estudiado es el EURUSD (Popovic y Durovic, 2014; Cui y Taylor, 2020; Akram *et al.*, 2008; Aiba *et al.*, 2002; Wang *et al.*, 2008; Mavrides, 1992), sugiriendo que tiene menos errores. Así que en este trabajo se estudiará un par menos habitual, el NZDUSD, dadas las sugerencias de expertos en el tema. Para esto es necesario encontrar características o propiedades del arbitraje triangular, determinar cuando hay oportunidades de operar el par (como la duración), establecer qué serie sintética de divisas es conveniente trabajar y precisar un umbral de operación.

1.2. Motivación

La información recaudada en la literatura muestra que el *arbitraje triangular* ha sido estudiado considerablemente. Sin embargo, en los últimos años se ha optado por investigar otro tipo de estrategias, como las *estrategias de alta frecuencia* (Ito *et al.*, 2012), que radica en operar en intervalos muy cortos de tiempo. Las oportunidades de arbitraje tienen una duración breve (Aiba *et al.*, 2002), cuestión de segundos(Ito *et al.*, 2012), los cuales ocurren en ciertos periodos del día y tratan de ser frecuentemente predichos (Akram *et al.*, 2008). El enfoque de las investigaciones a lo largo de estos años (2000-2019) ha sido determinar porqué existe este fenómeno, la razón de su corta duración, su viabilidad y con que instrumentos es conveniente operarlo (Aiba y Hatano, 2006; Aiba *et al.*, 2002; Cui y Taylor, 2020). Es fundamental analizar el arbitraje por la relación que existe entre distintas monedas, esto lleva a querer investigar que características tiene el arbitraje (sobre todo el *doble triangular*), donde fue considerado eficaz (Wang *et al.*, 2008), pero no se dieron detalles al respecto.

1.3. Objetivo general

Caracterizar el arbitraje triangular usando la divisa NZDUSD, para encontrar oportunidades de compra/venta y las condiciones con las cuales se pueda operar sin alto riesgo incrementando las oportunidades de inversión.

1.3.1. Objetivos específicos

1. Desarrollar una base de datos de series sintéticas tomando como par de divisas original el NZDUSD.
2. Determinar qué series sintéticas divergen más respecto al par original para llevar a cabo arbitraje triangular.
3. Obtener características del umbral de operación en las series sintéticas que conduzcan a encontrar oportunidades de arbitraje viables.
4. Clasificar las series sintéticas de acuerdo a su viabilidad para operar en el mercado con arbitraje triangular.

1.4. Preguntas de investigación

- ¿Qué características tiene el arbitraje triangular en el mercado de divisas?
- ¿Cuánto dura la oportunidad de arbitraje triangular?

1.5. Resultados esperados

Encontrar una duración promedio del arbitraje triangular con un umbral de operación estable y con amplitud poco volátil en relación al NZDUSD. Bajo esta condición, estaría casi asegurada la existencia de al menos una serie sintética operable con riesgo mínimo. Además, se espera determinar alguna relación entre el volumen y el arbitraje triangular que de indicios de oportunidades en el mercado para poder generalizar el procedimiento y aplicarlo a cualquier par de divisas.

1.6. Metodología propuesta

1. Creación de series sintéticas
    a) Diseño de base de datos sintética.
    b) Obtención de series de tiempo de relacionados con NZD y USD: AUDNZD, AUDUSD, EURNZD, EURUSD, GBPNZD, GBPUSD, NZDCAD, NZDCHF, NZDJPY, NZDUSD, NZDSGD, USDCAD, USDCHF, USDJPY, USDSGD.
    c) Limpieza de datos.
    d) Desarrollo de series sintéticas en base al arbitraje triangular.
    e) Normalización de series sintéticas respecto al par original NZDUSD.
2. Creación de umbral y etiquetado
    a) Diseño de un umbral para las series sintéticas.
    b) Desarrollo del umbral (límite superior e inferior).
    c) Etiquetado de base de datos con respecto al umbral de operación.
3. Obtención y selección de características
    a) Búsqueda de indicadores técnicos para implementar en las series sintéticas.

   b) Diseño e implementación de indicadores técnicos y sus variaciones respecto al etiquetado.

   c) Implementación de indicadores técnicos para obtener características.

   d) Selección de características.

4. Experimentación

   a) Corroborar que el umbral de ganancia es eficaz.

   b) Verificación de eficiencia de los puntos de inflexión obtenidos.

   c) Eliminación del mercado con las características obtenidas.

5. Análisis

   a) Análisis de resultados arrojados con las características obtenidas.

   b) Comprobación de los puntos de compra/venta obtenidos, para determinación de viabilidad, recurrencia y nivel de riesgo.

Capítulo 2. Marco Teórico

En este Capítulo se presentan conceptos económicos que se manejan o están implicados en el trabajo. Posteriormente, se exponen trabajos anteriores relacionados con el arbitraje.

2.1. Conceptos básicos económicos

2.1.1. Pips

Un *pip* es en realidad un acrónimo de *porcentaje en punto* (en inglés: *percentage in point*) (Mahmoodzadeh y Gençay, 2014). Un *pip* es el movimiento de precio más pequeño que puede hacer un tipo de cambio basado en la convención del mercado. La mayoría de los pares de divisas tienen un precio de cuatro decimales y el cambio más pequeño es el último (cuarto) punto decimal. Un *pip* es el equivalente a 1/100 de 1 por ciento o un punto base. Por ejemplo, el movimiento más pequeño que puede hacer el par de divisas USDCAD es $0.0001 o un punto base.

2.1.2. Broker

Un *broker* es una persona o una sociedad que se dedica a operar en el mercado financiero, en este caso en el Forex, realizando operaciones para sus clientes bajo las órdenes de aquellos (Nuti *et al.*, 2011). En español, el término *broker* equivale a Corredor (corredor financiero, corredor de bolsa, corredor de Forex, etc.). Por tanto, el *broker* es el elemento autorizado para intermediar entre el mercado de divisas y el inversor. Este último paga unas comisiones por los servicios de intermediación, que corresponden al pago del *know-how* o saber hacer del *broker* y de su licencia para operar en el mercado. Como otros mercados financieros, en el Forex sólo pueden intervenir directamente los agentes y corredores autorizados.

2.1.3. Spread

El significado de spread en el mercado de divisas es la diferencia entre los precios de demanda y oferta medido en *pips* (Goodhart *et al.*, 2002). Para los instrumentos de Forex que

cotizan al quinto decimal (por ejemplo, GBPUSD: 1.3245 dólares) 1 *pip* es igual a un incremento de precio de 0.00010. Para los instrumentos de Forex que cotizan al tercer decimal (por ejemplo, USDJPY: 101.522 yenes) 1 *pip* es igual a un incremento de precio de 0.010. Para los índices 1 *pip* es igual a un incremento de precio de 1.0, lo que también se llama un punto del índice. Desde el punto de vista de un *broker* en línea, el spread de operación es una de sus principales fuentes de ingresos, junto con comisiones y tarifas de intercambio. El spread puede ser fijo o variable, la mayoría de los *brokers* en línea ofrecen *spreads* de operaciones variables.

2.1.4. Slippage

El *slipagge* es la diferencia de precio que puede ocurrir entre el momento en que se coloca una orden de compra/venta y su ejecución real en el mercado (Mahmoodzadeh y Gençay, 2014). Un *slipagge* en Forex, por ejemplo, puede ocurrir cuando una orden de compra de EURUSD se coloca en 1.2300 y se ejecuta en 1.2305. En este caso, el operador sufre un deslizamiento negativo del precio de 5 *pips*. Esto se debe a que desde que la orden se coloca en la plataforma de operaciones y se ejecuta en el mercado puede pasar un periodo muy corto de tiempo durante el cual el precio puede moverse. Este movimiento de precio entre el precio solicitado y el precio obtenido se denomina *slipagge*.

2.1.5. Mercado eficiente, volatilidad y riesgo

Un *mercado eficiente*, es aquel mercado en el que el precio de manera completa e instantánea refleja toda la información disponible, relevante y de manera gratuita (Wang *et al.*, 2008). Por tanto, en un mercado eficiente, los operadores tienen un porcentaje de riesgo altísimo, y no pueden explotar la información disponible, pero esto es solo un supuesto. Existe evidencia de que el mercado no es así de perfecto (Cui y Taylor, 2020; Dow, 1998; Akram *et al.*, 2008). El arbitraje se aprovecha de tres paridades (o uso de tres pares), de la tasa de cambio en un instrumento y su diferencia. En otras palabras, el arbitraje triple o triangular es el aprovechamiento de la interacción entre tres instrumentos y el desequilibrio que tiene el mercado.

Dicho desequilibrio, en el mercado de divisas, es la *volatilidad* de una moneda (Baillie y Bollerslev, 1991). La volatilidad es un concepto que ayuda a medir la incertidumbre de un mercado o valor concreto cuando invertimos en el mercado. Desde el punto de vista del inversor,

hablar de divisas volátiles suele significar que estos están sujetos a fluctuaciones “violentas”. De hecho, se puede dividir el concepto en dos, la concepción técnica y la psicológica. Es decir, técnicamente podemos hablar de la dispersión de los precios respecto a una línea regresiva, mientras que desde el punto de vista subjetivo puede significar que la expectativa o percepción sobre un valor no es estable y fiable a mediano plazo.

Esto indica que existe *riesgo* al operar en el mercado, el cual se define como riesgo de pérdida debido a los factores que afectan a todo un mercado. Sin embargo, en el mercado de divisas, surge cuando existe una volatilidad en los tipos de cambio. Las empresas globales pueden estar expuestas al riesgo cuando realizan negocios debido a coberturas imperfectas (importación en una moneda diferente a la propia, deuda en moneda extranjera, etc.).

Por ejemplo, supongamos que un inversionista estadounidense tiene inversiones en China. El retorno realizado se verá afectado al cambiar las dos monedas. Supongamos que el inversor tiene un retorno realizado del 50 % de la inversión en China, pero el yuan chino se deprecia 20 % frente al dólar estadounidense. Debido al cambio en las monedas, el inversor solo tendrá un retorno del 30 %. Este riesgo puede mitigarse mediante la cobertura con fondos cotizados en el mercado de divisas.

2.1.6. Análisis fundamental

El análisis fundamental de los mercados financieros es una de las herramientas más utilizadas por los inversionistas, tanto principiantes como profesionales, especialmente los que operan a largo plazo (Yazdi y Lashkari, 2013). Trata de analizar el contexto en el que se mueven los mercados y usar la información recabada para tomar decisiones al operar. El análisis fundamental se centra en el estado general de la economía e investiga diversos factores macroeconómicos tales como: los datos del empleo, PIB, relación geopolítica, entre otros, y su impacto en activos financieros relacionados (Moosa, 2003). Por ejemplo, el análisis fundamental en Forex investiga qué publicaciones o eventos pueden provocar movimientos de precios en las divisas para estar preparado y aprovechar oportunidades de inversión. El objetivo final de realizar un análisis fundamental es descubrir el verdadero valor de un activo, compararlo con el precio actual y encontrar una oportunidad de operación (Goodhart *et al.*, 2002).

2.1.7. Análisis técnico

El análisis técnico es el estudio de los mercados financieros que se basa en datos, bolsa gráfica, patrones de precios y tendencias de las cotizaciones (Yazdi y Lashkari, 2013). El análisis fundamental, se centra más en el estudio del contexto económico, político y social, mientras que el análisis técnico es puramente matemático y algorítmico, basándose siempre en patrones y datos pasados. Así, este tipo de estudio se adapta mejor a las posiciones y operaciones a corto plazo. Las operaciones pueden durar desde meses, horas, hasta minutos, mientras que las oportunidades en el mercado son rápidas, por lo que no existe mucho tiempo de análisis (Yazdi y Lashkari, 2013; Moosa, 2003). Por ello, siempre es mejor contar con una serie de herramientas que faciliten nuestra toma de decisiones, sobre todo si puede ser preconfiguradas antes de iniciar nuestra sesión. Ahí entra el análisis técnico.

2.1.8. Indicadores técnicos

Sin duda, la importancia del análisis técnico radica en la utilización de numerosas herramientas y señales que hacen del operar en el mercado algo mucho más automatizado (Dávila y Herrera, 2015). Los indicadores técnicos son una de las principales herramientas utilizadas y existen una gran variedad que ayuda a interpretar los precios con la finalidad de acertar en la decisión de una inversión por medio de gráficos. Debemos tener en cuenta que la variable más importante que tiene una divisa es su precio, y por tanto, todos los indicadores técnicos van a ir retrasados con respecto a este. Los indicadores técnicos pueden clasificarse en cinco categorías generales según su función (Vajda, 2014; Yong *et al.*, 2015; Yazdi y Lashkari, 2013):

- Los *indicadores de tendencia*, como su nombre lo indica, son indicadores que nos muestran cuando un mercado se encuentra en tendencia (ya sea a la alza o baja). El propósito principal de este tipo de indicador es sugerir si debería buscar entrar en una posición larga o corta. Los indicadores de tendencia que se usan en este trabajo son los siguientes: *Moving Average Convergence Divergence (MACD), Average Directional Movement Index (ADX), Vortex Indicator (VI), Trix (TRIX), Mass Index (MI), Commodity Channel Index (CCI), Detrended Price Oscillator (DPO), KST Oscillator (KST), Ichimoku Kinkō Hyō (Ichimoku),* y *Parabolic Stop And Reverse (Parabolic SAR).*
- Los *indicadores osciladores* tienen como característica el moverse en un rango de valores, es decir oscila entre ellos. A su vez, marcan zonas de sobre compra y sobre venta, indicando al inversor cuando el precio puede darse la vuelta. Los osciladores son usados

mayormente cuando el mercado se encuentra lateralizado. Indicadores osciladores usados en este trabajo: *Money Flow Index (MFI), Relative Strength Index (RSI), True strength index (TSI), Ultimate Oscillator (UO), Stochastic Oscillator (SR), Williams %R (WR), Awesome Oscillator (AO), Kaufman's Adaptive Moving Average (KAMA)*, y *Rate of Change (ROC).*

- *Indicadores de volatilidad*. La volatilidad mide qué tan grandes son las subidas y bajadas de un par de divisas en particular. Cuando el precio de una moneda fluctúa velozmente hacia arriba y hacia abajo, se dice que tiene una alta volatilidad. Mientras que un par de divisas que no fluctúa tanto se dice que tiene baja volatilidad. Es importante tener en cuenta qué tan volátil es un par de divisas antes de abrir una operación, por lo que podemos tenerlo en cuenta al elegir el tamaño de nuestra operación y los niveles de stop y límite. Indicadores de volatilidad usados en este trabajo: *Average True Range (ATR), Bollinger Bands (BB), Keltner Channel (KC),* y *Donchian Channel (DC).*
- Los indicadores de volúmenes son aquéllos en cuyos cálculos se utilizan los volúmenes. Para el mercado de divisas, por volumen se entiende el número de *pips* (cambio del precio) que han aparecido durante un intervalo de tiempo. El mercado Forex está descentralizado, por lo que no es posible contar todos los contratos y sus tamaños como se hace en los mercados de valores. Como resultado, los inversionistas de Forex utilizan indicadores de volumen. Indicadores de volumen usados en este proyecto: *Accumulation/Distribution Index (ADI), On-Balance Volume (OBV), Chaikin Money Flow (CMF), Force Index (FI), Ease of Movement (EoM, EMV), Volume-price Trend (VPT), Negative Volume Index (NVI),* y *Volume Weighted Average Price (VWAP).*
- Existen un sin fin de indicadores técnicos que no cumplen con las características de las categorías anteriores, incluso hay otras categorías no menos importantes. Sin embargo, para fines de este trabajo, se englobaran aquellos que no se encuentran en las categorías anteriores como *otros indicadores*. Los indicadores usados en este proyecto son: *Daily Return (DR), Daily Log Return (DLR),* y *Cumulative Return (CR).*

2.2. Estrategias de inversión

En finanzas, una estrategia de inversión es un conjunto de reglas, comportamientos y procedimientos, diseñados para orientar a un inversor en la selección de una cartera de valores (Miner, 2008; Moosa, 2003). Generalmente la estrategia se diseña en torno a la desventaja riesgo-retorno de los inversores: algunos inversores prefieren maximizar los retornos esperados de la inversión a través de activos de riesgo, otros prefieren reducir al mínimo el riesgo,

pero la mayoría selecciona una estrategia en algún punto intermedio.

Existen distintos tipos de estrategias (Miner, 2008; Vidyamurthy, 2004; Dávila y Herrera, 2015), por ejemplo: el *scalping*, consiste en generar una gran cantidad de ganancias pequeñas en el menor tiempo posible; la *posicional*, que es opuesta al *scalping*, ya que los operadores deben mantener sus posiciones durante varias semanas o meses (y en algunos casos años); y las estrategias neutrales, las cuales requieren operar en el mercado sin dirección.

2.2.1. Estrategias neutrales

Las estrategias neutrales son estrategias que tienen un retorno no correlacionado con el rendimiento del mercado (Vidyamurthy, 2004). Sin importar las decisiones que se tomen, o si el mercado sube o baja, tanto en los buenos tiempos como en los malos, la estrategia neutral del mercado funciona de manera constante, y los resultados generalmente se logran con una menor volatilidad. Este resultado deseado se logra mediante el intercambio de portafolios neutrales, donde los portafolios son un conjunto de instrumentos o pares que se trabajan, como el cambio euro-dólar, dólar-yen, etc.

Existen distintas estrategias neutrales, como: las *coberturas*, que se refieren a aquellas estrategias donde sin importar la dirección que se tome en el mercado, se cubre la opuesta con otra operación; *operación por pares*, que se refieren a operar más de dos divisas a la vez de manera estratégica; y el *arbitraje*.

2.2.2. Arbitraje

Se ha hablado vagamente del arbitraje con anterioridad en este documento, pero antes de pasar a la definición formal y matemática, será más fácil de entender si se conceptualiza por medio de algo que se viva en la vida cotidiana.

Tomando como referencia el peso mexicano, el primer paso se lleva a cabo cuando se busca obtener alguna otra divisa (como el dólar estadounidense). En esta instancia cualquiera se dirigiría a alguna casa de cambio (llámese A a esta casa de cambio) para vender pesos y comprar dólares. Ahora bien, esta acción de compra-venta se considera una operación. Después se realiza una segunda operación en algún otro centro financiero (llámese B a este segundo centro financiero). Supongamos que la venta del dólar y la compra del peso difieren en A y

B. Al realizar estas dos acciones, dependiendo de los precios que tengan en *A* y *B*, puede que se tengan más o menos pesos respecto a la cantidad inicial. Esta serie de acciones se le llama arbitraje (particularmente arbitraje de dos puntos), ya que estamos aprovechando las diferencias de precios entre los establecimientos A y B.

Ahora bien, para arbitraje triangular, se tiene que agregar una tercera moneda (tómese como tercera moneda el euro). Si se realizan las operaciones:

1. Venta de peso y compra de dólar.
2. Venta de dólar y compra de euro.
3. Venta de euro y compra de peso.

Con esta serie de tres operaciones también se estaría llevando a cabo arbitraje (particularmente *arbitraje triangular* o de *tres puntos*). De hecho, si se generaliza realizándose *n* operaciones con *n* monedas distintas, esto se denomina *arbitraje múltiple* o *arbitraje n-ésimo*.

Formalmente, el arbitraje se define generalmente como capitalizar una discrepancia en los precios cotizados, desencadenado por la violación de una condición de equilibrio (fijación de precios) (Moosa, 2003). A menudo el arbitraje se describe como una operación sin riesgo, en el sentido de que todas las variables de decisión se conocen cuando se toma la decisión, pero el proceso invariablemente implica riesgo.

Es importante estudiar estas operaciones porque proporcionan los mecanismos por los cuales se mantienen las condiciones de equilibrio. De hecho, la condición de no arbitraje se toma para definir el precio de equilibrio del (de los) activo (s) subyacente (s), y por lo tanto, el estudio del arbitraje se reduce al estudio de la determinación de precios en los mercados financieros, que es un elemento crucial de la economía financiera.

2.2.3. Arbitraje de dos puntos

También conocido como arbitraje espacial, de localización o arbitraje de dos monedas (Moosa, 2003). En otras palabras, supongamos que hay dos centros financieros, A y B, dos monedas, *x* e *y*, y que (por el momento) no hay costos de transacción, impuestos y un margen de cero oferta. Sea $S(x/y)$ el tipo de cambio al contado entre *x* e *y*, medido como el precio (en términos

de x) de una unidad de y, entonces existe arbitraje si

$$S_A(x/y) \neq S_B(x/y). \tag{1}$$

Para simplificar, se evitará el uso de (x/y). Ahora bien, si la condición en la Ecuación 1 es violada, entonces existe la posibilidad de que

$$S_A > S_B \tag{2}$$

indicando que y es más barato en B que en A. Aquí la operación conveniente es comprar y en donde es más barato (B) y venderlo donde es más caro A. De tal manera que la ganancia es representada con π, donde

$$\pi = S_A - S_B. \tag{3}$$

Nótese que cuando $S_A = S_B$, significa que no hay ganancia, entonces no importa donde se compre o venda y.

2.2.4. Fianzas en el arbitraje de dos puntos

Supóngase que los centros financieros A y B cobran una fianza por transacción, ya que el centro financiero (como las casas de cambio) tienen que hacer negocio. En este caso la Ecuación 3 se ve afectada de la siguiente manera

$$\pi = S_A - S_B - (\beta_A + \beta_B), \tag{4}$$

donde β representa la ganancia de cada uno de los centros financieros. Esto nos indica que para tener una ganancia se tiene que cumplir lo siguiente:

$$S_A - S_B > (\beta_A + \beta_B), \tag{5}$$

esto quiere decir que la diferencia de las transacciones debe ser mayor a la fianza que se paga al centro financiero. En caso de existir una relación entre el volumen de transacción y la fianza, la relación se representará con la siguiente ecuación:

$$\pi = S_A(1-\beta_A) - S_B(1+\beta_B), \tag{6}$$

indicando que existe una ganancia ($\pi > 0$) cuando

$$S_A > S_B \frac{(1+\beta_B)}{(1-\beta_A)}. \tag{7}$$

Al tener fianzas continuas, el umbral de ganancia se mueve como se muestra en la Figura 4.

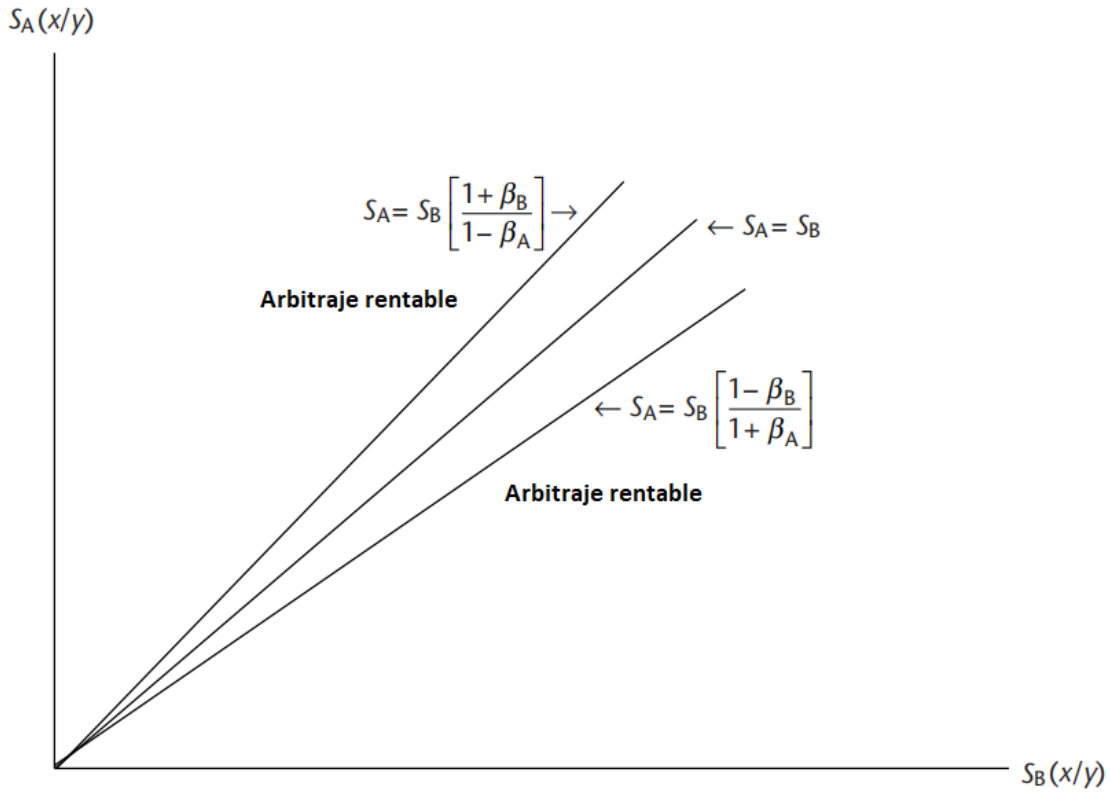

Figura 4. Arbitraje de dos puntos en presencia de fianzas continuas (imagen tomada y adaptada de *International financial operations* por Moosa (2003)).

2.2.5. Impuestos en el arbitraje de dos puntos

Supongamos que existe un impuesto α en relación a la ganancia obtenida. La relación ganancia-operación, representada en la Ecuación 3, se ve afectada como sigue

$$\pi = (1-\alpha)(S_A - S_B). \tag{8}$$

2.2.6. Ambos casos

Al tener impuestos y fianzas en el arbitraje de dos puntos, al combinar las ecuaciones 6 y 8, se obtiene la siguiente relación:

$$\pi = S_A(1-\alpha) - S_B(1-\alpha) - (\beta_A + \beta_B). \tag{9}$$

2.2.7. Arbitraje triangular

El arbitraje triangular asegura la consistencia entre tres monedas; no habría ninguna diferencia si usted compra (vende) libras esterlinas directamente con dólares, o si compra (vende) euros con dólares primero, y luego compra (vende) esterlinas con euros.

El arbitraje triangular implica un tipo de cambio negociado a dos precios diferentes, un precio directo y un precio indirecto (Mavrides, 1992). La forma de ganar dinero es comprar la moneda cuyo valor sea menor y vender la que tenga el precio (valor) más alto. Además, no es necesario tener diferentes centros financieros.

Particularmente estamos creando una divisa a partir de otras dos, por lo cual a esta divisa creada la llamaremos *divisa sintética*. (Moosa, 2003). Suponga tres monedas *x, y, z*, y tres posibles existentes tasas de cambio: $S(x/y)$, $S(x/z)$ y $S(y/z)$. Dado que estamos en este caso tratando con tres tipos de cambio, recurriremos al tipo de cambio original $S(x/y)$, por lo cual, los tipos de cambio son consistentes si

$$S(x/y) = \frac{S(x/z)}{S(y/z)}. \tag{10}$$

En general, si la condición mostrada en la Ecuación 10 es violada entonces se puede obtener una ganancia moviéndose en una dirección particular.

Por ejemplo, suponiendo que se realizan las siguientes operaciones

1. Se compra y con x
2. Se compra z con y
3. Se compra x con z

Entonces, se adquiriría una ganancia cuando se cumple que

$$S(x/y) > \frac{S(x/z)}{S(y/z)}, \tag{11}$$

o bien

$$\pi = \frac{S(x/z)}{S(y/z)S(x/y)} - 1. \tag{12}$$

En Figura 5 se puede apreciar el efecto del arbitraje triangular frente la oferta y demanda de algún instrumento particular.

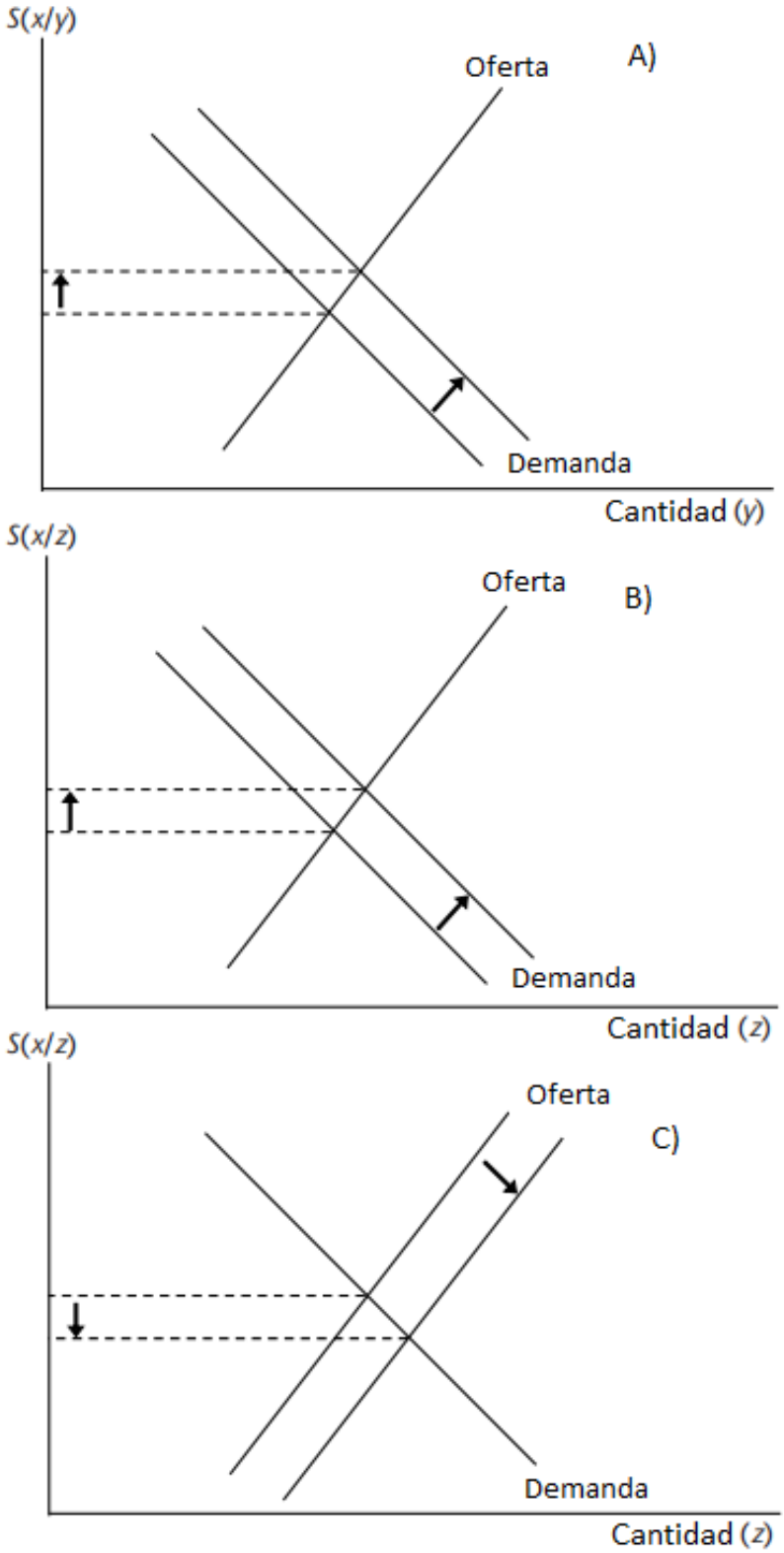

Figura 5. El efecto del arbitraje de tres puntos: El precio S(x/y) con demanda creciente (A), el precio S(x/z) con demanda creciente (B) y el precio S(x/z) con oferta decreciente (C) (imagen tomada y adaptada de *International financial operations* por Moosa (2003)).

2.2.8. Arbitraje múltiple

Siguiendo el mismo razonamiento, lo anterior podría generalizarse para n monedas, con lo cual obtendríamos la siguiente ecuación:

$$S(x_1/x_n) = S(x_1/x_2)S(x_2/x_3)...S(x_{n-2}/x_{n-1})S(x_{n-1}/x_n). \tag{13}$$

Si la Ecuación 13 es violada, tenemos oportunidades de arbitraje. Nótese que si $n = 2, 3$ esto se sigue cumpliendo (Moosa, 2003).

Observe que la parte izquierda de la igualdad es la *divisa original* $(x_1 - x_n)$, mientras que la parte de derecha de la igualdad es la serie de divisas que forman la *divisa sintética* $x_1 - x_n$. En un mercado eficiente, estas teóricamente son iguales. Para referirse a la serie de divisas que forman la *divisa sintética*, se le denominará *serie sintética*.

La ecuación de ganancia se obtiene de la Ecuación 13, quedando como sigue

$$\frac{S(x_1/x_n)}{S(x_1/x_2)S(x_2/x_3)...S(x_{n-2}/x_{n-1})S(x_{n-1}/x_n)} - 1 = \pi. \tag{14}$$

Por simplicidad, se llamará arbitraje múltiple a el *arbitraje n-ésimo*, con $n \in \mathbb{Z}^+ \setminus \{1, 2, 3\}$. Mientras que para el arbitraje de $n = 2, 3$ se usarán sus nombres propios, *arbitraje de dos puntos* y *arbitraje triangular*, respectivamente.

2.2.9. Cobertura

Una cobertura es una estrategia de inversión neutra para reducir la exposición existente al riesgo, denominado también *hedge* o *hedging* (Dash y NS, 2013). En Forex consiste en abrir, de manera estratégica, posiciones adicionales para protegerse contra los cambios adversos en los mercados de divisas. La cobertura en sí se basa en comprar o vender instrumentos financieros para compensar sus posiciones actuales y, así, reducir el riesgo de su exposición. La mayoría de los inversores intentarán encontrar maneras de limitar el riesgo potencial inherente a la exposición y crear una estrategia de cobertura es solo una de las muchas que se pueden emplear (Kočenda y Moravcová, 2019). Los inversores pueden decidir cubrir sus posiciones en Forex como método de protección contra las fluctuaciones en los tipos de cambio. A pesar de

que no existe un modo de eliminar los riesgos completamente, el beneficio de emplear una estrategia de cobertura reside en que puede ayudar a reducir las pérdidas o a limitarlas a una cantidad deseada. La cobertura de divisas es ligeramente diferente a la cobertura en otros mercados, ya que el mercado de Forex en sí es volátil por naturaleza. A pesar de que algunos inversores en divisas no se decantan por cubrir las posiciones en Forex porque creen que la volatilidad es solo una parte de operar en Forex, depende solo de la cantidad de riesgo que se esté dispuesto a asumir (Dash y NS, 2013). Si se cree que el valor de un par de divisas está a punto de caer, pero que la tendencia cambiará finalmente, una estrategia de cobertura puede ayudar a reducir las pérdidas a corto plazo al tiempo que protege los beneficios a largo plazo.

2.2.10. Cobertura simple

Una estrategia de cobertura de Forex simple consiste en abrir una posición contraria a una operación actual (Alexander, 2008). Por ejemplo, si ya se tenía una posición larga (o bien, la compra de un par de divisas que puede incrementar su valor en un futuro) en un par de divisas, se podría decidir abrir una posición corta (o venta de un par de divisa que puede decrementar su valor en un futuro) en el mismo par, y esto se conoce como cobertura directa. A pesar de que el beneficio neto de una cobertura directa es cero, de este modo se mantiene la posición inicial en el mercado listo para el momento en el que la tendencia se invierte (Kočenda y Moravcová, 2019). Si no se cubre la posición, cerrar la operación significa aceptar cualquier pérdida, pero si se opta por cubrirla, esto permite ganar dinero con una segunda operación a medida que el mercado se mueve en su contra al principio. Algunas plataformas no ofrecen la oportunidad de cubrir posiciones de manera directa y simplemente ofrecen compensar las dos posiciones.

2.2.11. Cobertura de varias divisas

Otra estrategia común de cobertura para Forex implica escoger dos pares de divisas que estén relacionadas de manera positiva, tales como los pares GBPUSD y EURUSD para después, abrir posiciones u operaciones en ambos pares pero en direcciones contrarias (Dash y NS, 2013; Kočenda y Moravcová, 2019). Por ejemplo, si se abre una posición corta en el par EURUSD, pero se decide cubrir la exposición al USD abriendo una posición larga en el par GBPUSD. Si el euro cae frente al dólar, la posición larga en la divisa GBPUSD haría incurrir en pérdidas, pero estas se compensan con el beneficio que habría arrojado la posición del par

EURUSD. Si el dólar cae, su cobertura compensaría cualquier pérdida que obtenga con su posición corta. Es fundamental recordar que cubrir más de un par de divisas implica otros propios riesgos. En el ejemplo anterior, a pesar de que se cubre la exposición al dólar, también hay una breve exposición a la libra y una larga al euro (Dash y NS, 2013). Si la estrategia de cobertura funciona, el riesgo se reduce y se obtienen beneficios. Con una cobertura directa, se tendría un balance neto equivalente a cero, pero con una estrategia de cobertura para varias divisas, existe la posibilidad de que una posición dé más beneficios que las pérdidas que arroja otra posición. Si no funciona, pueden ser varias las posiciones con las que se incurra en pérdidas.

2.2.12. Arbitraje triangular con cobertura

El concepto de arbitraje triangular con cobertura (o arbitraje doble triangular) que se manejará a lo largo de este trabajo se definirá como *llevar a cabo una operación de cobertura simple, que en vez de dos operaciones sencillas cubriéndose entre sí, se realiza por medio de dos arbitrajes triangulares que se cubren entre sí.*

2.2.13. Umbral de operación

La definición de umbral de operación, o denominado sencillamente como umbral, depende generalmente de la técnica que se esté utilizando para trabajar en el mercado de divisas. En este caso, se define el umbral de operación como un margen de seguridad en el cual si se está dentro del margen, la probabilidad de ganar alguna operación es baja dado que las series sintéticas involucradas están cerca de su precio original. Si se está fuera del margen, entonces hay mejores oportunidades de operación. Los conceptos involucrados para determinar el umbral de operación son spread, slipage, promedio de precio de entrada y salida por periodo, y 10 *pips* como seguro de operación.

2.3. Antecedentes de la estrategia de arbitraje triangular y sus variantes

Como se mencionó previamente, el arbitraje es el aprovechamiento de la inestabilidad del mercado, lo cual propició su análisis. En 1998, James Dow (Dow, 1998) analizó la relación de

invertir con operaciones especulativas, el uso de coberturas y las consecuencias económicas de introducir un nuevo valor al mercado de divisas por medio de modelos estadísticos centrados en los efectos de la liquidez del mercado cruzado. Las conclusiones de este trabajo fueron las siguientes: Introducir un nuevo valor, instrumento o seguridad, en el modelo aumentará la incidencia de la actividad de arbitraje al realizar alguna operación de cobertura menos riesgosa.

Hasta ese momento, las oportunidades de arbitraje duraban relativamente poco (desde segundos hasta unos pocos minutos) (Aiba *et al.*, 2002; Akram *et al.*, 2008; Hsu *et al.*, 2011), por esto muchas investigaciones se centraron en encontrar las oportunidades para operar con dicha técnica en el mercado por medio de modelos estocásticos (Aiba *et al.*, 2002; Aiba y Hatano, 2006), obteniendo distintas características de la técnica como: frecuencia con la que se podía operar, tamaño de las ventanas de oportunidad, y duración del arbitraje. Sin embargo, no lograban detectar oportunidades de arbitraje a corto plazo en una variedad de mercados financieros. Solamente se reafirmaba que la duración era larga como para explotarla, pero corta para que pasara desapercibida. Por esta razón, se introdujo el concepto de Condición de Paridad por tasa de Interés (o CIP por sus siglas en inglés), que ayuda a obtener costos de transacción, sus implicaciones para los cálculos de ganancias y pérdidas derivadas del arbitraje. Se analizó si las características de las oportunidades de arbitraje rentables variaban según el ritmo y volatilidad del mercado (Akram *et al.*, 2008). Los resultados que obtuvieron agregan detalles, pero las conclusiones respecto al arbitraje no cambiaron.

Tiempo después, las ventanas de oportunidad fueron reducidas gracias a la evolución de la tecnología y grandes cambios estructurales que tuvo el mercado (Wang *et al.*, 2008; Ito *et al.*, 2012). Por esto, la comunidad empezó a desarrollar herramientas con el objetivo de predecir cuándo aparecerían las oportunidades para operar con arbitraje (Wang *et al.*, 2008; Aguilar-Rivera *et al.*, 2015). Entre ellas se encuentra: Financial GP-2, una herramienta interactiva que encuentra oportunidades de arbitraje usando datos diarios (Wang *et al.*, 2008; Ito *et al.*, 2012); el Evaluador de Inversión Evolucionario Dinámico de Datos (EDDIE, por sus siglas en inglés), que se basa en las decisiones tomadas por un experto para evaluar la información relevante (Wang *et al.*, 2008; Ito *et al.*, 2012).

La herramienta EDDIE resultó prometedora (Tsang *et al.*, 2005), pero no en el mercado de divisas, sino con futuros y opciones. Por lo cual, se desarrolló EDDIE—ARB, una especialización de EDDIE, para el pronóstico de oportunidades de arbitraje en el mercado de divisas. Como herramienta, EDDIE—ARB fue diseñada para permitir que economistas y computadoras trabajaran juntos para identificar variables independientes relevantes. Fue capaz de identificar reglas en el mercado con alta precisión y superó otros instrumentos. Esto estableció a EDDIE—ARB como una herramienta para encontrar oportunidades de arbitraje y atrajo la atención de tanto a eco-

nomistas expertos, como a científicos en computación. La última versión de EDDIE—ARB logró generar predicciones de forma fiable. Tales predicciones ayudaron a evitar arbitraje ingenuo, donde los datos están fuera de la muestra.

También se propuso un acercamiento por algoritmos genéticos para operaciones con múltiples pares (Aguilar-Rivera *et al.*, 2015). Esta aproximación usó Medias Móviles y Bandas Bollinger para estimar la media más grande de stocks y sus límites. El algoritmo genético implementado optimizaba los parámetros de Medias Moviles y las Bandas Bollinger. La novedad de este enfoque se presenta de tres maneras diferentes: cubre lapsos de tiempo no incluido en otros artículos de revisión, cubre problemas no considerados por otros, se compara y analiza el alcance cubierto por referencias pasadas y nuevas. Los resultados concluyeron el interés por los métodos que usaron como medio Medias Móviles y Bandas Bollinger ya que los problemas han cambiado con el tiempo. Sin embargo, los algoritmos genéticos siguieron siendo el enfoque más popular en la literatura en esa temporada. Por otro lado, hay combinaciones de problemas y métodos de solución que no se investigaron y quedaron por desarrollarse.

Otra aproximación utilizado fue usando la combinación de cinco pares distintos para buscar oportunidades de arbitraje en el EURUSD (Wang *et al.*, 2008), ya que se decía en ese entonces que las oportunidades de arbitraje triangular eran inexistentes. Este trabajo concluyó que es eficaz la técnica, pero por su liquidez, es uno de los pares más cuidados por los bancos y los mismos inversionistas en el mercado. Se probó la ineficiencia del mercado y pronosticaron oportunidades de inversión por medio de arbitraje. En los experimentos usaron modelos basados en redes neuronales con retropropagación (*BP-NN*) para la previsión del tipo de cambio.

Ya en el 2011, las estrategias de arbitraje más populares estaban enfocadas en el acarreo de costo y econométrico (Hsu *et al.*, 2011). Sin embargo, estos enfoques tuvieron dificultades para lidiar con los datos de operación con periodo corto y encontrar oportunidades de arbitraje. Esto a su vez condujo al fracaso de dichas estrategias. Luego, se mostró que existían oportunidades de arbitraje libre de riesgo en su totalidad en el mercado de divisas, pero la duración de segundos persistía (Ito *et al.*, 2012). Dichas oportunidades de arbitraje se daban cuando el mercado estaba activo y había mucha volatilidad. El spread negativo y las oportunidades de arbitraje a nivel macro eran nulas, pero se mostró que las técnicas algorítmicas podían pasar por alto esto. Simultáneamente, en este periodo se popularizaron las operaciones de alta frecuencias (Ito *et al.*, 2012; Popovic y Durovic, 2014). Subsecuente a esto, se buscaron anomalías al operar en periodos de semanas y días en el mercado de divisas, por lo cual se exploraron modalidades y buscó alguna relación entre las anomalías que sucedían en los periodos de días y minutos en el EURUSD (Popovic y Durovic, 2014). Gracias a esto se encontró que las mejores oportunidades de arbitraje se daban los viernes vendiendo dólares y comprando euros a las 00:00 GMT+2 del día, o al revés a las 03:00 GMT+2 en el mismo día,

esto se mostró por el aprovechamiento de que el mercado es más grande en volumen que en lo que oferta, por lo que debía ser altamente líquido y eficiente. Sin embargo, se descubrió que hay menos oportunidad de arbitraje en dichas monedas durante el miércoles y jueves.

A mediados del 2015, se exploró el rendimiento de los sistemas de *operación por pares* y los métodos con los que se seleccionan los pares, el cual forma parte del arbitraje al seleccionar pares de divisas (Huck y Afawubo, 2015). Normalmente se enfocaban en métodos de cálculo de distancia, pero este trabajo tomó enfoques estadísticos y econométricos, así como la cointegración y estacionariedad que hacen el sistema de comercio mucho más exigente desde un punto de vista computacional; debido a que, el método de distancia generaba retornos insignificantes en exceso. La selección de pares al seguir el criterio estacionario condujo a un rendimiento débil, revelando que la cointegración proporciona un retorno alto, estable y robusto.

Zhenyu Cuia y Stephen Taylor en el 2019 (Cui y Taylor, 2020), vincularon el problema de la detección de arbitraje de *k*-divisas con la identificación del ciclo de longitud máxima en un grafo construido a partir de la matriz de tipo de cambio de precio de venta más alta. Posteriormente se analizó un algoritmo que utilizó el producto *max-plus* (o suma-máxima) para encontrar la duración del ciclo más largo de una longitud determinada. Su tiempo de ejecución se compara con un enfoque de búsqueda de permutación análoga. Con esto, finalmente se demostró que el rendimiento en tiempo de ejecución por medio de este método es mejor para el arbitraje triangular, cuádruple y quíntuple, que con el método tradicional.

La diferencia de este trabajo respecto a lo que se ha hecho es que la mayoría se enfoca en la detección y predicción de arbitraje tradicional y/o triangular en pares convencionales (Cui y Taylor, 2020; Dow, 1998; Akram *et al.*, 2008; Aiba *et al.*, 2002; Wang *et al.*, 2008), mientras que la presente investigación se enfocará en la caracterización de arbitraje general. Una de las premisas es que existen oportunidades de arbitraje de tres puntos (Cui y Taylor, 2020; Wang *et al.*, 2008), adicionalmente se usarán múltiples indicadores técnicos como características en vez de usar una única herramienta. Aprovechando que el NZDUSD es uno de los pares menos vigilados en el mercado, se usará este instrumento con la hipótesis de que existen más oportunidades por no destacar frente a otros pares de divisas.

Capítulo 3. Metodología

Existen diversas investigaciones sobre creación de herramientas, análisis de técnicas y sistemas para realizar inversiones en el mercado de divisas. Sin embargo, no se ha establecido y probado una metodología por completo. Esto se debe a que el éxito de cada proyecto es gracias a la ganancia generada en el mercado de divisas, no por el éxito en el ámbito científico, propiciando que los creadores de cada uno de estos sistemas, técnicas y herramientas protejan su capital intelectual. En este capítulo se describe a detalle cada una de las actividades llevadas a cabo para cumplir con los objetivos de este proyecto.

3.1. Obtención y preparación de datos

Los precios históricos de cada par de divisas originales fueron tomados de *Dukascopy Bank SA* a nivel de minuto de acuerdo a lo establecido en la literatura (Cui y Taylor, 2020). Para llevar a cabo lo anterior, se ingresó a la página de *Dukascopy (https://www.dukascopy.com/)*, se abrió una cuenta gratuita en la página y se descargó su plataforma, *JForex*. En *JForex* se descargan los datos requeridos con la configuración que se desee, en este caso se buscó descargar la mayor cantidad de datos posibles. Se pretendió rebasar la cantidad de 10 años de datos en minuto, pero la información comenzaba a tener lagunas (incluso a nivel de meses en pérdida de información), que a periodo de minuto fue demasiada información perdida. Dicha información cumple con las características mostradas en la Tabla 1. Para más información sobre los acrónimos usados, véase Apéndice II en la Tabla 16.

Tabla 1. Información detallada de datos históricos descargado de *Dukascopy (https://www.dukascopy.com/swiss/english/home/)*.

Característica	Configuración
Plataforma	JForex
Nivel de periodo	Minuto (60 segundos)
Periodo de tiempo	01/01/2008 al 01/01/2018
Pares de divisas	AUDNZD, AUDUSD, EURNZD, EURUSD, GBPNZD, GBPUSD, NZDCAD, NZDCHF, NZFJPY, NZDUSD, USDCAD, USDCHF y USDJPY.
Cantidad de minutos	~4.1 millones de minutos por divisa

Al descargar los datos históricos, la información no se descarga para una divisa en particular, sino por pares. De acuerdo a los objetivos planteados en este trabajo, se requiere generar el par NZDUSD a través de las divisas: AUD, EUR, GBP, CAD, CHF y JPY. Cada periodo de tiempo

contiene la información mostrada en la Tabla 2.

Tabla 2. Detalles de los campos de información por periodo en base de datos histórica.

Característica	Descripción
Fecha	Día, mes y año al que pertenece el periodo
Hora	Hora, minuto y segundo al que pertenece el precio
Precio de apertura	Precio con el que inició el periodo
Precio mínimo	Precio mínimo que se alcanzó en el periodo
Precio máximo	Precio máximo que se alcanzó en el periodo
Precio de cierre	Precio con el que cerró el periodo
Volumen	Fuerza de movimiento del precio en ese periodo

3.1.1. Limpieza y rellenado de la base de datos cruda

Una vez obtenida la información cruda de cada uno de los pares de divisas, se revisaron los datos históricos para verificar que no existiesen lagunas (largos lapsos de tiempo sin información), precios en ceros, precios negativos, o periodos con información faltante o errónea; para esto se desarrolló un programa en *Python* que revisaba la información por medio de búsqueda lineal. Una de las divisas consideradas inicialmente fue el SGD (dólar de singapur), pero se encontró que todos sus cruces con otros divisas contenían lagunas a nivel de meses, incluso en fechas próximas a la que se desarrolló este proyecto. Esto condujo al descarte de dicha divisa. Al revisar el resto de la base de datos histórica, se probó la inexistencia de precios en cero, negativos o faltantes en los campos, por lo cual no fue necesario llevar a cabo limpieza o rellenado. Esto condujo a la siguiente etapa que fue la creación de la base de datos con series sintéticas.

3.1.2. Fórmula de obtención de pares sintéticos

Recuerde que para crear el NZDUSD a través de las divisas AUD, EUR, GBP, CAD, CHF y JPY, es necesario utilizar la Ecuación 13 (Capítulo 2). Observe, sin embargo, que algunos pares se encuentran invertidos. Por ejemplo, se requiere que el AUDNZD se encuentre como NZDAUD para usar tal cual la Ecuación 13. Es decir, es necesario que los pares de divisas estén acomodados de la siguiente manera: NZD*** y ***USD, donde *** representa cualquier divisa que no sean el NZD o USD. De otro modo, no se estaría generando el par sintético deseado, ya que el orden de las divisas en un par es importante: la primera divisa representa

la divisa que se está comprando y la segunda representa la que se está vendiendo. En otras palabras, comprar dólar neozelandés con dólar australiano en vez de dólar australiano con dólar neozelandés. Consecuentemente, de acuerdo a la fórmula, en vez de multiplicar por su precio, se dividió por él. De esta manera se logró procesar todos los pares en un sólo ciclo, obteniendo así, los pares sintéticos del NZDUSD a través del AUD, EUR, GBP, CAD, CHF y JPY. Si se graficaran los pares sintéticos y el original, se obtendría un comportamiento similar al que se muestra en la Figura 6.

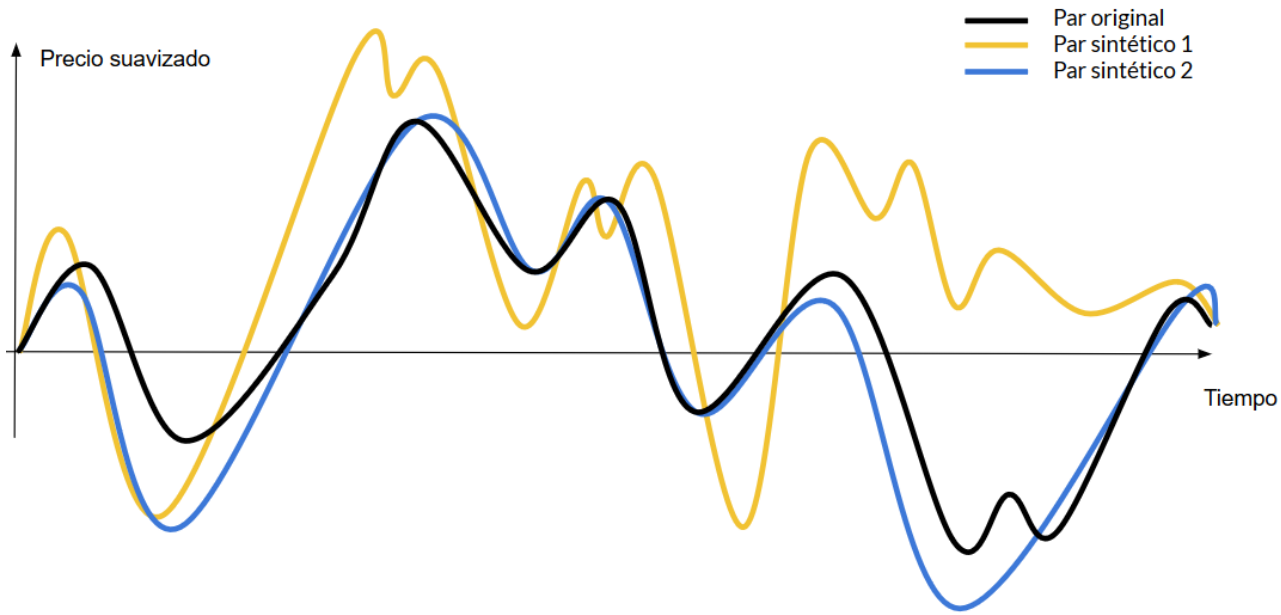

Figura 6. Planteamiento de base de datos histórica de series sintéticas con precios.

Para ver de manera más clara la diferencia entre precios de los pares sintéticos en relación al precio original se normalizaron los precios (S) respecto al par NZDUSD, por medio de la siguiente operación matemática:

$$S_{normalizado} = \frac{S_{\text{sintético}}}{S_{NZDUSD}}. \tag{15}$$

La ecuación anterior nos permite visualizar mejor la divergencia de cada una de las series sintéticas como se muestra en la Figura 7.

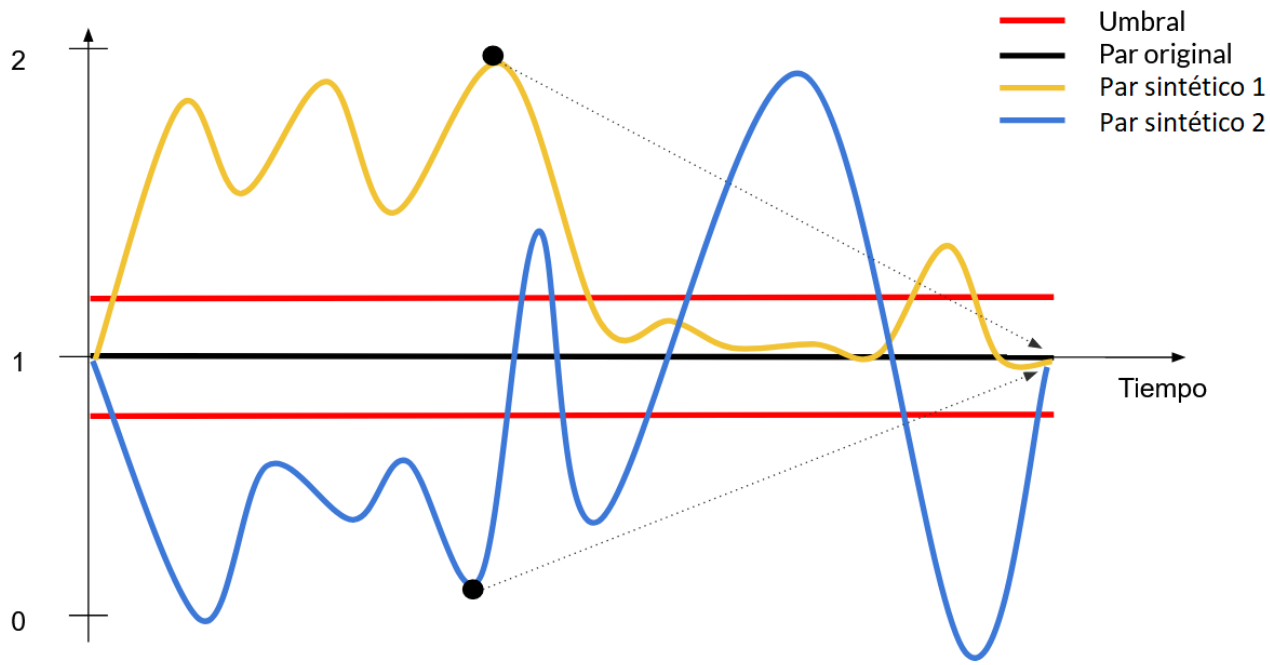

Figura 7. Planteamiento de base de datos histórica de series sintéticas con precios normalizados.

3.1.3. Umbral y etiquetado

Para etiquetar los puntos en la base de datos sintética, fue necesario definir un umbral de operación que tiene como objetivo distinguir un rango o umbral en el cual hay menor riesgo (si se está fuera del umbral) o mayor riesgo (si se está dentro del umbral) para operar con arbitraje. De esta manera, es posible observar gráficamente que series sintéticas están más próximas al precio original y cuales no. La idea principal detrás de esto es dar rigor a la definición de *máximos* y *mínimos relativos* de cada serie sintética. Ahora bien, fijemos una hora h en un día arbitrario. Sean $S^+_{h,1}, S^+_{h,2}, \ldots, S^+_{h,n}, n \epsilon \mathbb{Z}^+_{\geq 1}$ los precios de cierre de cada periodo mayores que el precio original de alguna divisa sintética arbitraria XYZ en la hora h y donde $\mathbb{Z}$ es mayor igual a 1, porque éste último representa el valor del precio original normalizado como se muestra en la Figura 7. Sean $S^-_{h,1}, S^-_{h,2}, \ldots, S^-_{h,m}, m \epsilon \mathbb{Z}^+_{<1}$ los precios de cierre de cada periodo estrictamente menores que el precio original de alguna divisa sintética arbitraria XYZ en la hora h. Definamos $XYZ^{sup}_h = \{S^+_{h,n} | n \epsilon [0, 59]\}$ y $XYZ^{inf}_h = \{S^-_{h,n} | n \epsilon [0, 59]\}$ como los conjuntos con los todos los precios superiores e inferiores. Consecuentemente la cardinalidad de dichos conjuntos se representa como $|XYZ^{sup}_h|$ y $|XYZ^{inf}_h|$, respectivamente. Los umbrales preliminares por divisa sintética se definen de la siguiente manera:

$$umbral_{h,sup}^{XYZ} = \frac{S_{h,1}^{+} + S_{h,2}^{+} + ... + S_{h,n}^{+}}{|XYZ_{h}^{sup}|} \tag{16}$$

y

$$umbral_{h,inf}^{XYZ} = \frac{S_{h,1}^{-} + S_{h,2}^{-} + ... + S_{h,n}^{-}}{|XYZ_{h}^{inf}|}. \tag{17}$$

Esto quiere decir que el umbral preliminar superior/inferior es el promedio de los precios superiores/inferiores con respecto al precio original por minuto de cada una de las series sintéticas, creando así una *serie sintética de precios promedios superiores/inferiores* a nivel de minuto, en su respectiva hora h. Una vez establecidos los umbrales superiores e inferiores por divisa y considerando $XYZ_1, XYZ_2, ..., XYZ_j$, $j \epsilon \mathbb{Z}^{+}$ como uno de las caminos sintéticos, los umbrales de operación generales se definen de la siguiente manera:

$$umbral_{h,sup} = \frac{umbral_{h,sup}^{XYZ_1} + umbral_{h,sup}^{XYZ_2} + ... + umbral_{h,sup}^{XYZ_j}}{j} \tag{18}$$

y

$$umbral_{h,inf} = \frac{umbral_{h,inf}^{XYZ_1} + umbral_{h,inf}^{XYZ_2} + ... + umbral_{h,inf}^{XYZ_j}}{j}. \tag{19}$$

En otras palabras, el umbral superior/inferior en un periodo de tiempo, es el promedio de los límites superiores/inferiores por divisa sintética a nivel de hora. Se manejó ese promedio de 60 minutos, ya que al promediar a nivel de minuto, el umbral se comporta como divisa sintética y no se aprecian los rangos de mayor o menor riesgo. Esto permitió etiquetar cada punto correspondiente a un precio en cada serie de la siguiente manera:

- Puntos dentro del umbral.
- Puntos fuera del umbral.
    - Debajo del umbral (excluyendo mínimos relativos).
    - Mínimos relativos.
    - Sobre el umbral (excluyendo máximos relativos).
    - Máximos relativos.

Para esta investigación los puntos que más nos interesan son los máximos relativos y los mínimos relativos, ya que es en estos periodos de tiempo donde potencialmente se puede tener un mayor margen de ganancia al operar. Además, es necesario asegurar que dichos puntos de precio (máximos y mínimos relativos) no correspondan a *picos* que duren segundos. Por lo cual, para definir los máximos/mínimos relativos se tomaron aquellos que estuvieran

fuera del umbral de operación y que en un radio de cinco minutos (cinco periodos antes y después del máximo o mínimo relativo en cuestión) sea el precio más alto/bajo respecto a su vecindad. Adicionalmente, estos puntos en la vecindad no tienen que cruzar el umbral como se muestra en la Figura 8.

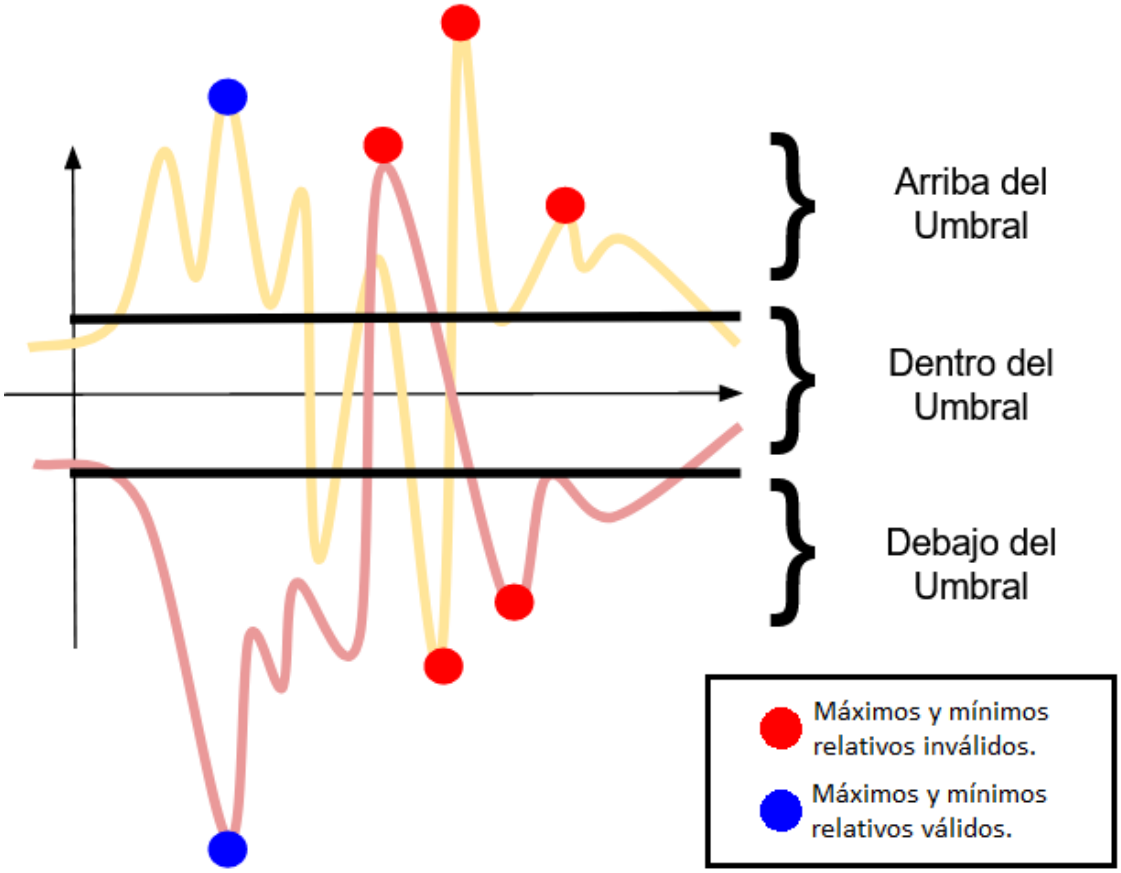

Figura 8. Descripción gráfica de etiquetado y ubicación de puntos. Los puntos azules son máximos y mínimos relativos válidos por ser el precio máximo o mínimo en un radio de 5 minutos, mientras que los rojos no cumplen la condición de los 5 minutos y que los puntos de la vecindad no crucen el umbral.

3.2. Experimentos

3.2.1. Escenarios del trabajo

En la presente investigación se consideraron tres escenarios con los cuales se desea trabajar, para poder atacar este problema de diversas formas:

- El primer escenario que llamaremos de manera coloquial *experimento binario*, tiene co-

mo objetivo identificar oportunidades de arbitraje en cada una de las series sintéticas sin considerar el tipo de operación a realizar (compra o venta). Para lograr esto, el tipo de etiquetado que se lleva a cabo es considerar únicamente los puntos máximos y mínimos relativos como positivos, mientras que los puntos dentro del umbral son considerados como negativos.

- El segundo escenario considera las oportunidades de arbitraje que se encuentren en el primer experimento, para poder clasificar las oportunidades de compra o venta. Para lograr esto, el etiquetado que se maneja en este experimento es también binario. De manera coloquial llamaremos a este segundo escenario como *experimento doble binario* o *experimento binario sobre binario*, dado que hipotéticamente al operar en la bolsa con algún indicador creado a partir del primer escenario, se vuelve a clasificar de manera binaria sobre el resultado anterior en caso de ser una oportunidad de operación.
- El tercer escenario lo llamaremos de manera coloquial *experimento multiclase* ya que se busca clasificar los puntos de una serie sintética considerando tres posibles grupos o eventos mutuamente excluyentes: mínimo relativo (oportunidad de compra), máximo relativo (oportunidad de venta) y los puntos dentro del umbral (o bien, que no es oportunidad para operar con esta técnica debido a la cercanía de los precios).

3.2.2. Etiquetado y balanceo de datos

Dado que la clasificación de los puntos depende del escenario de experimentación, en cada uno de ellos se descartaron los puntos que se ubicaban sobre y debajo del umbral que no pertenecían al conjunto de mínimos y máximos relativos, quedándonos únicamente con estos dos últimos y los puntos dentro del umbral.

Ahora bien, en el primer escenario, los casos negativos son los puntos dentro del umbral, mientras que los máximos y mínimos relativos son el caso positivo. En el segundo escenario se tienen los máximos relativos como casos positivos y los mínimos relativos, como casos negativos. En el escenario multiclase, al manejar tres etiquetas se denominó como etiqueta uno a los máximos relativos, como etiqueta dos a los mínimos relativos y como etiqueta tres a los puntos dentro del umbral.

Para balancear los datos se tomó una muestra aleatoria del conjunto más pequeño, en este caso de los mínimos relativos, y se extrajo esa misma porción del resto de los conjuntos. En total se tomó el 70 % del conjunto de mínimos relativos, el cual se usó para el entrenamiento de los algoritmos correspondientes y el 30 % restante se utilizó para las pruebas. Este proce-

dimiento de balanceo se repitió para cada uno de los tres escenarios considerando que varían las etiquetas, obteniendo así el total de puntos por conjunto mostrado en la Tabla 3. Como se puede observar, para el primer escenario se obtuvo un total de ciento ochenta mil puntos aproximadamente en cada conjunto para los entrenamientos, mientras que los conjuntos de prueba tenían un total de cincuenta y cuatro mil puntos aproximadamente cada uno. En el caso del segundo escenario la cantidad de etiquetas era la misma y el conjunto más pequeño siguió siendo el de mínimos relativos, pero se buscaba diferenciar únicamente entre compras y ventas, por lo que los conjuntos se redujeron a la mitad (noventa mil puntos cada uno). Para las pruebas se utilizaron un total de veintisiete mil puntos por conjunto en el segundo escenario. Para el tercer escenario, al requerir tres etiquetas, se obtuvieron noventa mil puntos aproximadamente por conjunto, ya que el conjunto más pequeño continuó siendo el de mínimos relativos. En las pruebas de este escenario, así como en el anterior, fue de 27 mil puntos aproximadamente.

Tabla 3. Cantidad de puntos por conjunto aproximada.

Escenario	Conjunto	Etapa	Cantidad de puntos
1er	0	Entrenamiento	180,000
1er	1	Entrenamiento	180,000
1er	0	Prueba	54,000
1er	1	Prueba	54,000
2do	0	Entrenamiento	90,000
2do	1	Entrenamiento	90,000
2do	0	Prueba	27,000
2do	1	Prueba	27,000
3er	0	Entrenamiento	90,000
3er	1	Entrenamiento	90,000
3er	2	Entrenamiento	90,000
3er	0	Prueba	27,000
3er	1	Prueba	27,000
3er	2	Prueba	27,000

3.2.3. Cálculo de características y selección

Por medio de la librería de python "Technical Analysis (TA)"(Lopez, 2017) se generaron las características de los conjuntos, los cuales corresponden a los valores arrojados por cada uno de los indicadores técnicos utilizados (vease la Sección 2.1.8) y sus variaciones. De un total de 34 indicadores técnicos, 23 de ellos pueden modificarse en sus parámetros, por lo que se realizaron cinco versiones alternas a cada uno. La versión original usaba los parámetros

recomendados, mientras que en las versiones alternas se modificaron el número de puntos anteriores para determinar la característica del precio deseado (Tabla 4), los distintos periodos fueron elegidos de tal manera que existiera una cantidad significativa de versiones y con una distancia significativa, cuidando que no sobrepase la media hora ya que de otro modo serían demasiado grandes los periodos. Además, pueden revisarse los precios en cualquier plataforma de FOREX a 5, 10, 15 y 30 minutos, permitiendo comparar la información obtenida con futuros trabajos. Por otro lado, se añadió el de 20 periodos ya que se buscaba tener una cantidad significativa de indicadores mas no demasiado grande. De esta manera, se obtuvo un total de 209 indicadores técnicos.

Tabla 4. Tabla de acrónimos usados para las versiones de indicadores técnicos y sus significados.

Acrónimo	Descripción
v0	Versión con parámetros recomendados
v1	Versión que revisa 5 periodos
v2	Versión que revisa 10 periodos
v3	Versión que revisa 15 periodos
v4	Versión que revisa 20 periodos
v5	Versión que revisa 30 periodos

Una vez generadas las 209 características para cada uno de los precios en la muestra, se procedió a seleccionar las características más relevantes por medio de seis algoritmos distintos implementados en python con las bibliotecas *sklearn* y *pandas* (para la manipulación de los grandes volumenes de información). Los algoritmos y parámetros usados para dicha actividad son los que se encuentran en la Tabla 5 (Agarwal, 2019). A los conjuntos de características generados se les puso el nombre del correspondiente algoritmo con el cual fueron producidos: Correlación, Chi2, RFE, Lasso, RF y LGBM, respectivamente. Este proceso se repitió para cada uno de los escenarios, por lo cual los conjuntos entre experimentos fueron distintos. Para distinguirlos se mantuvieron en diferentes carpetas relacionadas con su respectivo escenario.

Tabla 5. Parámetros usados en cada algoritmo seleccionador.

Algoritmo	Parámetro	Configuración
Correlación	Función	*f_regressor*
X^2	Función	*chi2*
RFE	No. Pliegues	10
	Métrica	f1
Regresión logística	Solucionador	*lbfgs*
	Penalización	*Ninguna*
Bosque aleatorio	No. estimadores	10
LightGBM	No. estimadores	500
	Nivel de estimadores	0.05
	No. hojas	32
	ratio de columnas	0.02
	Regularización en el peso L1	3
	Regularización en el peso L2	1
	Reducción mínima para nueva partición	0.01

3.2.4. Métricas de evaluación

Para evaluar el rendimiento de los algoritmos se usaron las siguientes métricas:

- Exactitud, que responde a la pregunta: ¿cuál es la proporción de predicciones correctas?
- Precisión, que responde a la pregunta: ¿qué proporción de positivos reales se han predicho correctamente?
- Sensibilidad, que responde a la pregunta: ¿qué proporción de predicciones positivas es correcta?
- Métrica f1, que se utiliza para combinar las medidas de precisión y sensibilidad en un sólo valor. Esto resulta práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.
- Pérdida de Hamming, responde a la pregunta: ¿qué proporción de predicciones fue incorrecta?

Estas métricas se describen a detalle en la Sección . Ahora bien, el resultado que se reportará y discutirá a lo largo de este trabajo es la media armónica (para más detalle, véase la Sección). Esta es una manera de mezclar las calificaciones de las métricas ya que la dispersión de éstas toman juego en su cálculo con la finalidad de obtener una mejor *calificación genérica*, o bien, tomar una mejor decisión considerando todos los resultados. No se consideró

Hamming loss en el cálculo, ya que esta métrica se considera que está implementada en la exactitud. Consecuentemente, los resultados específicos para cada métrica se presentan en los Apéndices.

3.2.5. Validación cruzada

Posteriormente se procedió a realizar entrenamientos por medio de validación cruzada para analizar cuál de los conjuntos de características (correlación, X^2, RFE, regresión logística, bosque aleatorio y lightGBM) era más eficiente de acuerdo a distintos algoritmos con la media armónica de las distintas métricas. Los parámetros usados en el algoritmo de validación fueron 10 pliegues con los clasificadores de SVM lineal, bosque aleatorio, MLP, árbol de decisión y *Naive Bayes*, los cuales son recomendaciones de distintos artículos (Yiu, 2020; Browne, 2000; Raul Garreta, 2013; Edwards, 2018; Abraham *et al.*, 2014; Pal, 2005; Ali *et al.*, 2012; Suykens y Vandewalle, 1999; Zhao *et al.*, 2012). Particularmente, el clasificador de Naive Bayes es utilizado como referencia ya que no se esperan buenos resultados de su desempeño, mientras que al clasificador SVM se le asignó un kernel lineal por cuestiones de tiempo de ejecución, porque con otros *kernels o núcleos* en múltiples conjuntos excedía el tiempo determinado para llevar estas actividades (menos de un mes) al considerar el número de elementos que se manejaban por conjunto y el número de características que contiene cada punto.

Estas a actividades se llevaron a cabo por medio del lenguaje de programación *python* con las bibliotecas de *sklearn* (librería con clasificadores y métricas), *pandas* (para manipulación de grandes volumenes de información), *numpy* (para llevar a cabo operaciones matemáticas de manera más sencilla) y *time* (para tomar el tiempo con el cual se llevaron a cabo las clasificaciones).

Tabla 6. Parámetros de clasificadores usados en validación cruzada en el primer escenario. La *configuración 1* representa un tipo de configuración que se usó en el clasificador, la segunda configuración es una versión alterna utilizada sólo para algunos algoritmos.

Algoritmo	Configuración 1	Configuración 2
SVM	Kernel lineal	—
Bosque aleatorio	100 árboles	200 árboles
MLP	50 capas ocultas	100 capas ocultas
Árbol de decisión	Mejor partición	Partición aleatoria
Naive Bayes	sin parámetros	—

Para el *experimento doble binario* y el *experimento multiclase* se usó el 70 % de los datos

para la validación cruzada. Dados los resultados arrojados durante el primer escenario, se tomó la decisión de descartar los clasificadores SVM y MLP. La discusión respecto a esta decisión se lleva a cabo en el Capítulo 6. Además, se implementó la métrica de *Hamming loss* porque indica el porcentaje de error. Los clasificadores utilizados se muestran en la Tabla 7.

Tabla 7. Parámetros de clasificadores usados en la validación cruzada del segundo y tercer escenario. La *configuración 1* representa un tipo de configuración que se usó en el clasificador, la segunda configuración es una versión alterna que se usó en caso de estar rellena de información la casilla.

Algoritmo	Configuración 1	Configuración 2
Bosque aleatorio	100 árboles	200 árboles
Árbol de decisión	Mejor partición	Partición aleatoria
Naive Bayes	sin parámetros	—

En el segundo escenario fue necesario generar un segundo conjunto independiente para las pruebas ya que los resultados obtenidos fueron inusualmente buenos. Para obtener dicho nuevo conjunto independiente, se generó de la misma manera que el primero, pero si el punto aleatorio tomado se encontraba en el primer conjunto independiente, se tomaba otro aleatorio. Así sucesivamente hasta obtener un conjunto totalmente distinto.

Particularmente para el tercer escenario, se consideraron originalmente dos tipos de promedios, *macro* y *micro*. Sin embargo, se descartó la idea de usar el *promedio micro*, ya que los datos estaban balanceados y *las clases* contaban con el mismo peso.

3.3. Prueba con conjuntos independientes de datos

Posteriormente a los entrenamientos, se probaron las mejores combinaciones de clasificadores-conjunto de características obtenidos durante la validación cruzada con el 30 % de los datos restantes.

3.4. Equipo computacional

Las especificaciones de la computadora usada para los experimentos son los siguientes:

- Computadora: DELL G7 7588.
- Procesador: Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz 2.21 GHz.

- Memoria RAM: 16 GB.
- Sistema operativo: Windows 10 Home (ver. 20H2).

Capítulo 4. Resultados y discusiones: Construcción de series sintéticas y conjunto de datos

En este capítulo se presentan los resultados de la construcción de la base de datos de series sintéticas y los conjuntos obtenidos a partir de ella. Además se presentan las interpretaciones e implicaciones de dichos resultados, como afectan y como se refleja al operar en la vida diaria en el mercado de divisas.

4.1. Base de datos histórica de series sintéticas

Dado que no existe una base de datos de series sintéticas en la literatura con la cual comparar, este fue uno de los objetivos principales del presente proyecto, sobre todo, porque en el *FOREX* usualmente los inversionistas evitan revelar las estrategias más productivas y sus hallazgos. En la Figura 9 se muestra una fracción de la base de datos de series sintética en la cual se puede observar la divergencia de precios entre pares sintéticos con respecto al par original (NZDUSD, representado en negro). En dicha figura, al generar el NZDUSD por medio del dólar australiano, la ineficiencia del mercado se hace presente de manera prominente respecto al resto de las divisas (EUR, GBP, CAD, CHF y JPY) que parecen no distar mucho del par original.

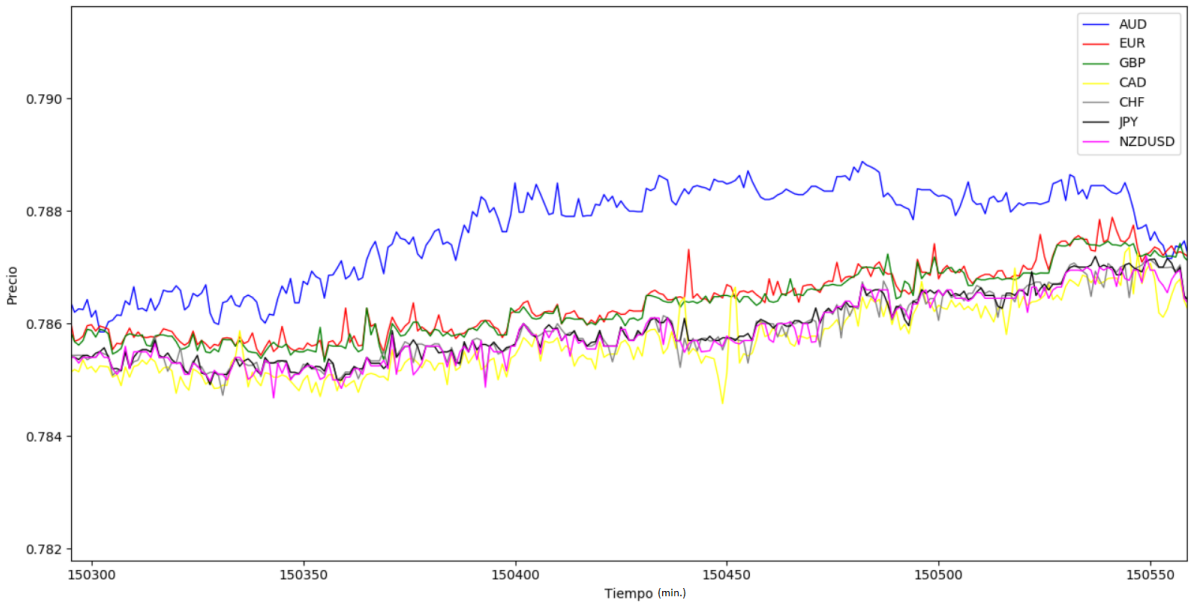

Figura 9. Comparación entre pares sintéticos en un periodo de tiempo arbitrario. El tiempo está a nivel de minuto.

Observe que dicho gráfico, en ese periodo arbitrario, revela la existencia de oportunidades

de arbitraje triangular con el par sintético del dólar australiano. También se puede estimar que las ventanas de oportunidad duran aproximadamente desde cinco minutos hasta ocho horas. Sin embargo, con el resto de pares sintéticos no se tiene claro el número de oportunidades ni la duración.

En caso de llevarse a cabo operaciones de arbitraje doble triangular se necesitarían al menos dos series sintéticas divergentes respecto al precio original. Sin embargo, en este punto no es posible saber que otra serie sintética pueda llevar a cabo el arbitraje doble triangular. Lo ideal es que a simple vista se logre observar dos series sintéticas diverger como en el caso del dólar australiano.

4.2. Normalización de base de datos histórica de series sintéticas

Una vez normalizada la base de datos de series sintéticas (Figura 10), es posible apreciar de una mejor manera que el par sintético del AUD dista del precio original considerablemente durante dos días para ese periodo arbitrario en la base de datos: Existen diferencias de precios tanto por encima como por debajo del precio original, las cuales se pueden denominar como diferencias positivas y diferencias negativas. La amplitud es de aproximadamente 8 pips en los puntos más altos y bajos, mientras que se conserva una duración de entre cinco minutos a quinientos en el mejor de los casos. Además, se puede observar que el JPY y CHF divergen del precio original (línea negra), pero ya no en gran medida.

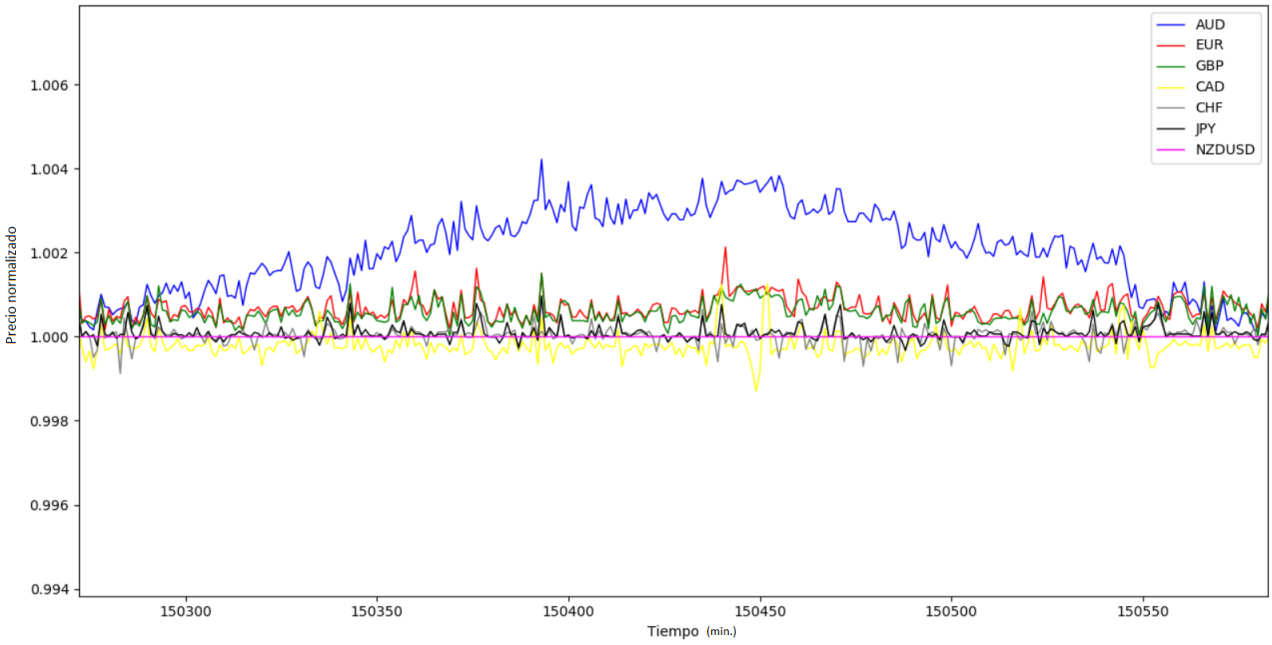

Figura 10. Gráfico normalizado de pares sintéticos en un periodo de tiempo arbitrario.

Sin embargo, esto nos da la idea que entre el minuto 1000 y 1500 pudo haber existido una oportunidad de arbitraje doble triangular con las series sintéticas del AUD y CAD. Para resolver este problema que se tiene de identificación de oportunidades de arbitraje triangular (sencillo o doble) se emplea el umbral de operación, del cual se presenta resultados en la siguiente sección.

4.3. Umbral de operaciones

Las líneas negras en la Figura 11 representan los umbrales de operación positivo y negativo (el cual fue generado con las ecuaciones 18 y 19). Recordemos que estos sirven como un margen de seguridad de operación, lo que significa que existe menor riesgo de perder alguna operación con arbitraje triangular fuera de los umbrales. Las series sintéticas ubicadas en la parte superior presentan oportunidades de venta, mientras que las series debajo del umbral son oportunidades de compra. Ahora bien, las series sintéticas ubicadas dentro del umbral son de alto riesgo, ya que su precio es cercano al original (el precio del NZDUSD).

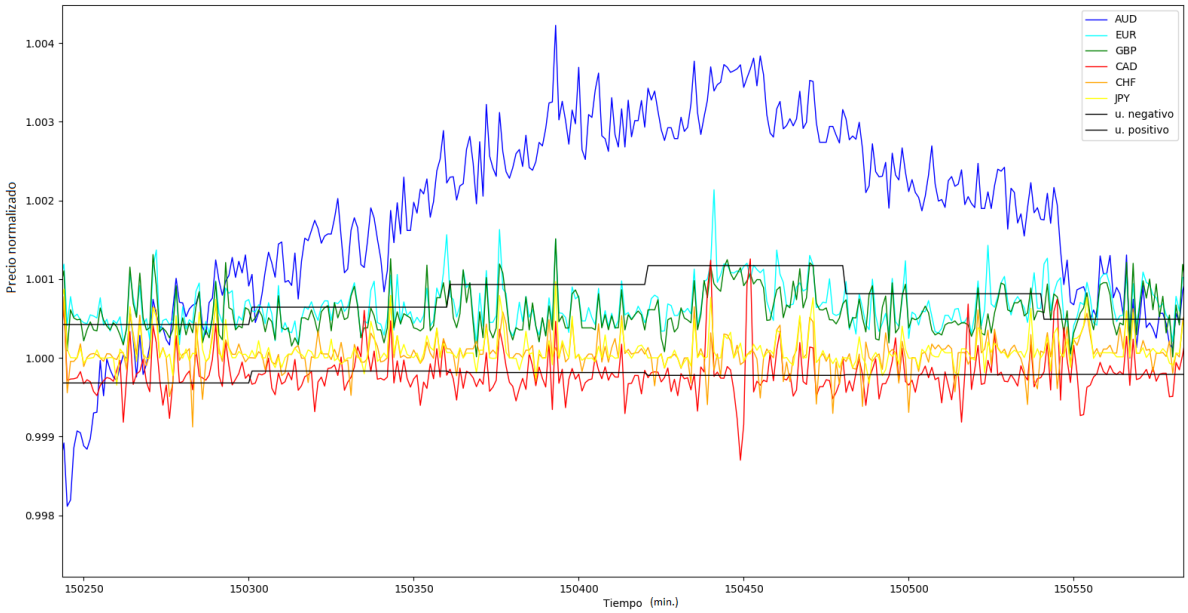

Figura 11. Gráfico normalizado de pares sintéticos con umbrales superiores e inferiores en dos horas arbitrarias de la base de datos histórica. Periodo de tiempo dentro del intervalo mostrado en los gráficos 9 y 10. Cerca del minuto 150450, existe una oportunidad de arbitraje triangular con cobertura vendiendo la serie sintética del AUD y comprando la serie sintética de CAD gracias al mínimo relativo fuerte que presenta el CAD cerca del dicho minuto.

Por otro lado, si se examina los pares sintéticos inferiores, la serie con mayor oportunidad de *compra* es el dólar canadiense. En este caso la mejor oportunidad de arbitraje triangular aparece cerca del minuto 150450 ya que el dólar canadiense tiene un *pico* fuerte mientras

que el dólar australiano se encuentra en los puntos más altos de precio alcanzado en esos dos días. Esto quiere decir que el mejor punto para abrir ambas operaciones de arbitraje triangular sucede en el minuto 150450 aproximadamente mientras que el cierre de ambas operaciones podría ser, idealmente, posterior al minuto 150550, cuando ambas divisas están prácticamente dentro del umbral. Nótese que la serie sintética del CAD estuvo oscilando en el umbral inferior mientras que la serie sintética del AUD se mantuvo fuera del umbral superior durante un largo periodo de tiempo.

4.4. Etiquetado

La Tabla 8 muestra que de un total de 4,131,361 puntos por serie sintética a lo largo de los diez años a nivel de minuto, se tienen aproximadamente 33500 máximos y mínimos relativos, o bien, cerca del 0.8 % de puntos por serie sintética son máximos y mínimos relativos. Esto podría interpretarse como aproximadamente cuatro oportunidades de arbitraje triangular por día, las cuales son suficientes considerando que es una estrategia neutra, o sea con bajo nivel de riesgo. Además, dado que dichas oportunidades funcionan también para operar con *arbitraje triangular*, entonces el número de oportunidades se incrementa potencialmente al doble.

Tabla 8. Despliegue de ubicación de puntos por serie sintética.

Divisa	Tipo de Cambio	Dentro del Umbral	Debajo del Umbral	Arriba del Umbral	Mínimos Relativos	Máximos Relativos	Total
AUD	Bid	3297722	125874	671407	6707	29651	4131361
AUD	Ask	2222712	1013013	836278	34030	25328	4131361
EUR	Bid	3380278	3978	731661	163	15281	4131361
EUR	Ask	2258490	1050605	766826	34919	20521	4131361
GBP	Bid	3445921	1231	675262	16	8931	4131361
GBP	Ask	2258702	1074078	744218	35519	18844	4131361
CAD	Bid	3801149	322214	5490	2462	46	4131361
CAD	Ask	2256046	770701	1049640	20372	34602	4131361
CHF	Bid	3915627	204684	8825	2163	62	4131361
CHF	Ask	2246821	766576	1064370	19583	34011	4131361
JPY	Bid	3997113	120815	9075	4278	80	4131361
JPY	Ask	2259039	782899	1035996	20533	32894	4131361
Promedio		2944968,333	519722,3333	633254	15062,08333	18354,25	

Tabla 9. Despliegue porcentual de ubicación de puntos por serie sintética. Los valores en negro indican los valores más altos, en rojo indican los más bajos.

Divisa	Tipo de Cambio	Dentro del Umbral	Debajo del Umbral	Arriba del Umbral	Mínimos Relativos	Máximos Relativos	Total
AUD	Bid	79,82%	3,05%	16,25%	0,16%	0,72%	100,00%
AUD	Ask	53,80%	24,52%	20,24%	**0,82%**	0,61%	100,00%
EUR	Bid	81,82%	0,10%	17,71%	0,00%	0,37%	100,00%
EUR	Ask	54,67%	25,43%	18,56%	**0,85%**	0,50%	100,00%
GBP	Bid	83,41%	0,03%	16,34%	0,00%	0,22%	100,00%
GBP	Ask	54,67%	26,00%	18,01%	**0,86%**	0,46%	100,00%
CAD	Bid	92,01%	7,80%	0,13%	0,06%	0,00%	100,00%
CAD	Ask	54,61%	18,65%	25,41%	0,49%	**0,84%**	100,00%
CHF	Bid	94,78%	4,95%	0,21%	0,05%	0,00%	100,00%
CHF	Ask	54,38%	18,56%	25,76%	0,47%	**0,82%**	100,00%
JPY	Bid	96,75%	2,92%	0,22%	0,10%	0,00%	100,00%
JPY	Ask	54,68%	18,95%	25,08%	0,50%	**0,80%**	100,00%

En la tabla se mantiene los precios de compra y venta (*bid* y *ask*) por separado. Esto tiene una sencilla explicación aplicada a la vida real: cuando uno busca comprar o vender alguna divisa, el precio depende del tipo de operación, así como en las casas de cambio. No es lo mismo cambiar pesos mexicanos a dolares americanos que de dolares americanos a pesos mexicanos. Este tren de pensamiento lleva a que la mejor estrategia es vender caro y comprar barato, lo que en términos de este proyecto significa, vender lo más alto fuera del umbral superior y comprar lo más bajo fuera del umbral inferior respectivamente. Dicho esto, para detectar las oportunidades de compra con el precio más bajo, en la Tabla 9, se utiliza la información de las divisas con tipo de cambio *bid*. Al ordenar los pares sintéticos con más oportunidades de compra se tiene: AUD con 0.16 %, JPY con 0.1 %, CAD con 0.06 % y CHF con 0.05 %. Nótese que el EUR y GBP, no presentan oportunidades de compra, mientras que el CAD y CHF presentan porcentajes de 0.06 % y 0.05 %, lo que significa que existe al menos una oportunidad de arbitraje triangular cada dos días en cada una de estas divisas sintéticas. Sin embargo, en la divisa sintética de AUD existe al menos una o dos oportunidades por día de arbitraje triangular. Por otro lado, para oportunidades de venta con el precio más alto, se buscan los porcentajes más altos en las divisas sintéticas con tipo de cambio *ask*, los cuales son: CAD, CHF y JPY con 0.84 %, 0.82 % y 0.8 %, respectivamente. Esto quiere decir que existen más posibilidades de operar a la venta que a la compra, ya que aquí se cuenta con un aproximado de 9 oportunidades al día de operar a la venta con arbitraje triangular en dichas divisas sintéticas. Las series sintéticas de AUD, EUR y GBP, presentaron menos oportunidades de operación a la venta con un total aproximado de 5 a 7 oportunidades por día. En resumen,

existen más oportunidades de venta que de compra con la estrategia de arbitraje triangular. Además, se puede observar como la mayor cantidad de precios o puntos en la series de tiempo se encuentran dentro del umbral.

Un detalle importante es que las oportunidades detectadas son de arbitraje triangular. Para un arbitraje doble triangular se necesita que coincida una oportunidad de compra y una oportunidad de venta. Considerando el hecho de que son escasas las de venta, las posibilidades de arbitraje doble triangular son aún más escasas, al menos en esos 10 años de información.

4.5. Conclusiones parciales

- Se encontraron ventanas de oportunidad a la compra (0.5 %) y a la venta (0.8 %), dichas oportunidades son escasas en relación al total de puntos de la base de datos, pero las ventanas de oportunidad si son lo suficientemente grandes como para analizar, tomar decisiones y operar (5 minutos), contradiciendo algunos artículos de la literatura que afirman que dichas oportunidades duran sólo segundos (Akram *et al.*, 2008; Ito *et al.*, 2012).
- Las oportunidades de arbitraje triangular son mayores al usar un par poco convencional (NZDUSD), respecto al par más estudiado en la literatura, EURUSD, el cual tiene en promedio 0.3 % de oportunidades (Popovic y Durovic, 2014), mientras que el NZDUSD puede llegar a tener hasta 0.8 % de oportunidades de compra-venta.
- Una ventaja de este trabajo es que se estudió las oportunidades de arbitraje triangular en el mercado en periodos de minutos, mientras que otros artículos lo estudian a nivel de hora y día (Cui y Taylor, 2020; Wang *et al.*, 2008).

Capítulo 5. Resultados y discusión del primer escenario: oportunidad de operación

En este capítulo se presentan los resultados de la selección de características, entrenamientos y pruebas del primer escenario. Asimismo se interpretan y discuten dichos resultados, qué tipo de características fueron tomadas, los mejores conjuntos, clasificadores y combinaciones.

5.1. Conjunto de características

El detalle de los conjuntos generados puede consultarse en los Apéndices de este trabajo (Tabla 17), además de cada característica contenida. En la Fig. 12 se observa que los tipos de indicadores dominantes son los de volatilidad y tendencia, mientras que los del tipo volumen y *otros* se registran menos. Esto da a entender que el volumen no tiene gran influencia en el mercado de divisas, al menos con respecto a operaciones de arbitraje. Por otro lado, la volatilidad y tendencia parecen ser un factor importante para determinar la existencia de oportunidades de arbitraje. Los conjuntos de X^2 y RFE contienen únicamente indicadores de volatilidad, sin embargo, bosque aleatorio y lightGBM tomaron de cada tipo. En cuanto al número de características por conjunto, el único patrón encontrado es que los métodos de incrustado toman la cantidad más grande de características, seguido por los de envoltura. Sin embargo, los conjuntos más pequeños son generados por los métodos de filtro. En cuanto a los indicadores osciladores, tienen presencia en algunos conjuntos, más no son tan prominentes como los de volatilidad y tendencia. El conjunto con el mayor porcentaje de características de este tipo es el de lightGBM, que en compensación tomó pocos de volatilidad.

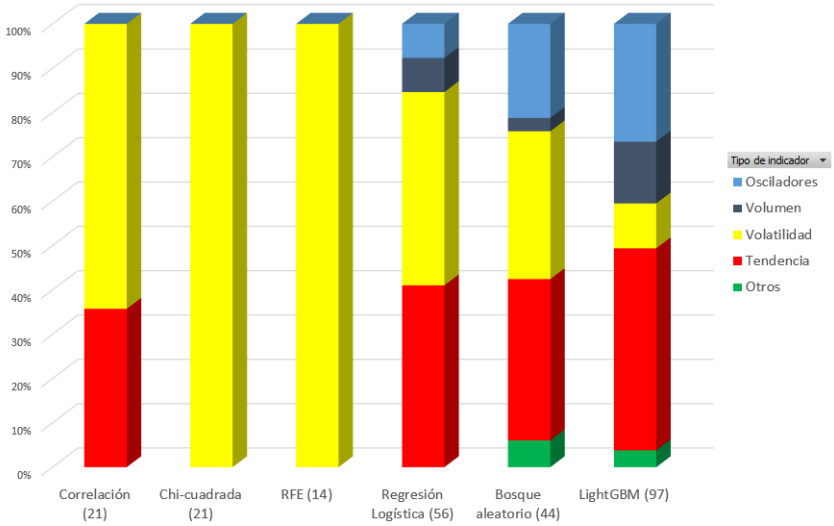

Figura 12. Conjuntos de características generados para el primer escenario y el porcentaje de tipo de indicadores técnicos en cada uno de ellos. El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Entre las características dominantes por tipo de indicador se encuentran:

- Tendencia: *ADX* y *Parabolic SAR*.
- Oscilador: *Stochastic Oscillator*.
- Volatilidad: *Bollinger Bands* y *Donchian Channel*.
- Volumen: *ADI, Easy of Movement* y *Volume-price Trend*.
- *Otros*: *Daily Return* y *Daily Log Return*.

Estas características se presentan en los conjuntos de manera recurrente en sus diferentes versiones, sobre todo los de tendencia y volatilidad. Por otro lado, las características que no aparecen en los conjuntos son las siguientes:

- Oscilador: *Ultimate Oscillator (UO)* y *Rate of Change (ROC)*
- Volumen: *Volume Weighted Average Price (VWAP)*

Cabe mencionar que existen múltiples indicadores que se usan en un sólo conjunto, en una versión, perteneciendo a las categorías de Osciladores, Volumen y *Otros*. Estos indicadores en su mayoría se encuentran en el conjunto *lightGBM*, el cual obtuvo el mayor número de características. Los indicadores de *KST Oscillator, Ichimoku* y *CCI* que son pertenecientes a la categoría de tendencia, también se utilizan una sola vez.

5.2. Análisis con media armónica: Entrenamiento

En esta sección se muestran los resultados de los entrenamientos de cada clasificador (SVM con kernel lineal, bosque aleatorio, MLP, árbol de decisión y *Naive Bayes*) con cada uno de los conjuntos generados de características. Para determinar las mejores combinaciones clasificador-conjunto se discutirán los resultados partiendo de cada uno de los clasificadores, para posteriormente comparar únicamente entre las mejores combinaciones respecto a sus resultados en las pruebas.

Para el caso del clasificador SVM con kernel lineal, en la Figura 13 se puede observar que las mejores medias armónicas fueron arrojados por los conjuntos generados por RFE, regresión logística y X^2, en dicho orden. Además, cada media es menor a 0.77, lo que indica que de manera general, usar el clasificador de SVM tiene como máxima calificación 77/100 y como mínimo 52/100. Sin embargo, la peor calificación general fue dada por el conjunto de lightGBM. Dicho esto, el mejor resultado se obtiene con el conjunto RFE con una calificación de 0.7677. Ahora bien, al considerar la desviación estándar (o STD), se observa que los conjuntos de RFE y X^2 son superiores, notablemente RFE, mientras que los conjuntos con más alta STD son los de regresión logística y bosque aleatorio. Por lo cual, considerando esto y la media armónica (para más detalle, véase la Sección), podría inferirse que la mejor combinación es SVM-RFE. Existe una mejora en la detección de las oportunidades de arbitraje al reducir el número de características dado que la segunda media armónica más baja es el conjunto TODAS. Por otro lado, lightGBM tiene menor calificación y un mayor STD, lo cual puede deberse a que en dicho conjunto se hayan descartado algunas características que sean de gran importancia para el objetivo de este escenario. Cabe destacarse que para este clasificador, el orden de mejores conjuntos está dado por el tipo de selección (exceptuando el de TODAS): primero el único conjunto dado por método de envoltura, seguido por aquellos dados por los métodos embebidos, y al final aquellos obtenidos por un método de filtro.

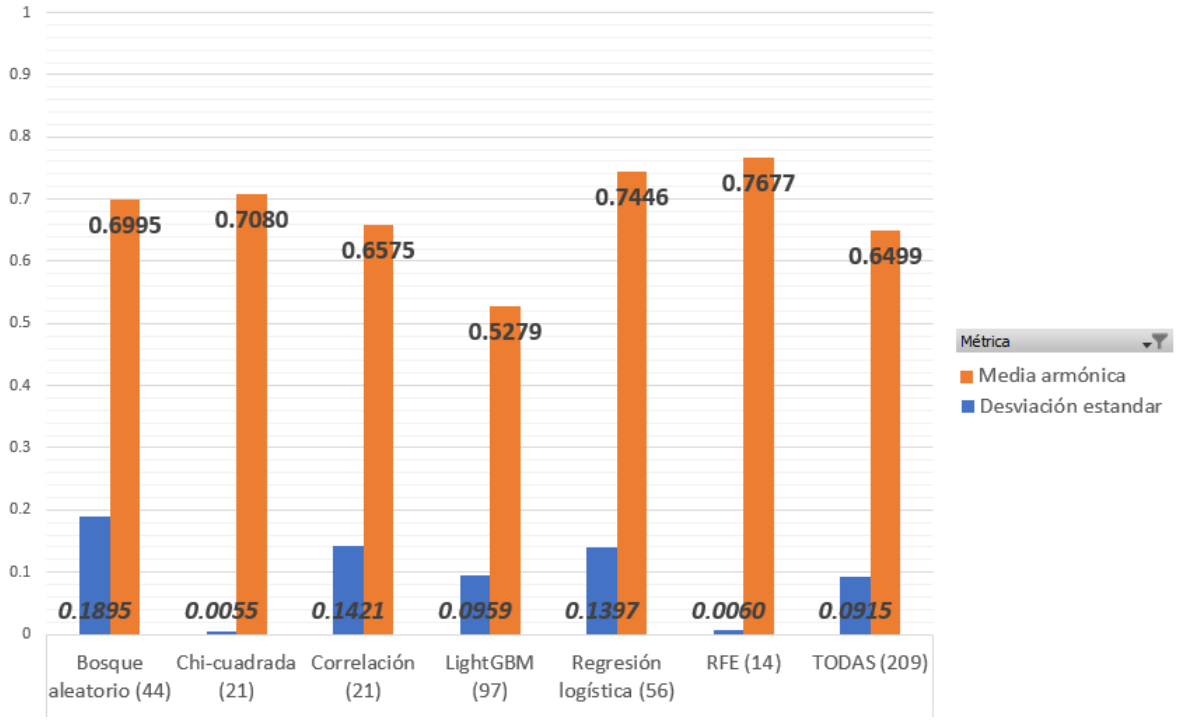

Figura 13. Resultados de clasificación en el primer escenario con el clasificador SVM (con kernel lineal). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

En la Figura 14 se muestran los resultados de bosque aleatorio con la configuración de 100 árboles (BA 100). Las peores calificaciones son arrojadas por la combinación de dicho clasificador con los conjuntos de X^2 y correlación. Probablemente dichos conjuntos de características obtuvieron menores calificaciones por su baja cantidad y calidad de características, ya que RFE también tiene una cardinalidad baja (14). Por otro lado, las mejores calificaciones las obtuvo el conjunto bosque aleatorio con un total de 0.9297, seguido por regresión logística con una calificación de 0.9254. Al considerar el STD, sobresalen posteriormente los conjuntos con calificaciones más baja, sin embargo, los que tiene STD más alto son lightGBM, TODAS y RFE, en ese orden. Consecuentemente se consideraría que bosque aleatorio no es el conjunto más apropiado, pero la diferencia de STD entre cada uno de estos conjuntos es menor a 0.01, o bien, de 0.003 entre bosque aleatorio y correlación (el conjunto con mejor STD). En términos generales se podría decir que lo anterior no afecta al desempeño de la combinación BA100-BA. En este caso el conjunto TODAS se encuentra entre los tres mejores conjuntos para este clasificador (probablemente por el funcionamiento del clasificador, ya que necesita de más información). Sin embargo, no está en primer lugar, dicho puesto se lo lleva el conjunto de bosque aleatorio, el cual tiene un total de 44/209 características. El número de características es considerablemente bajo, lo que implica que si existen características menos indispensables, pero el conjunto de bosque aleatorio logró discriminarlas con efectividad para lograr una calificación de 0.9297 de media armónica. Se puede observar que el orden de los mejores conjuntos

por media armónica está dado por el tipo de seleccionador (exceptuando el conjunto TODAS): primero aquellos dados por métodos embebidos, luego por método de envoltura y al final de filtro.

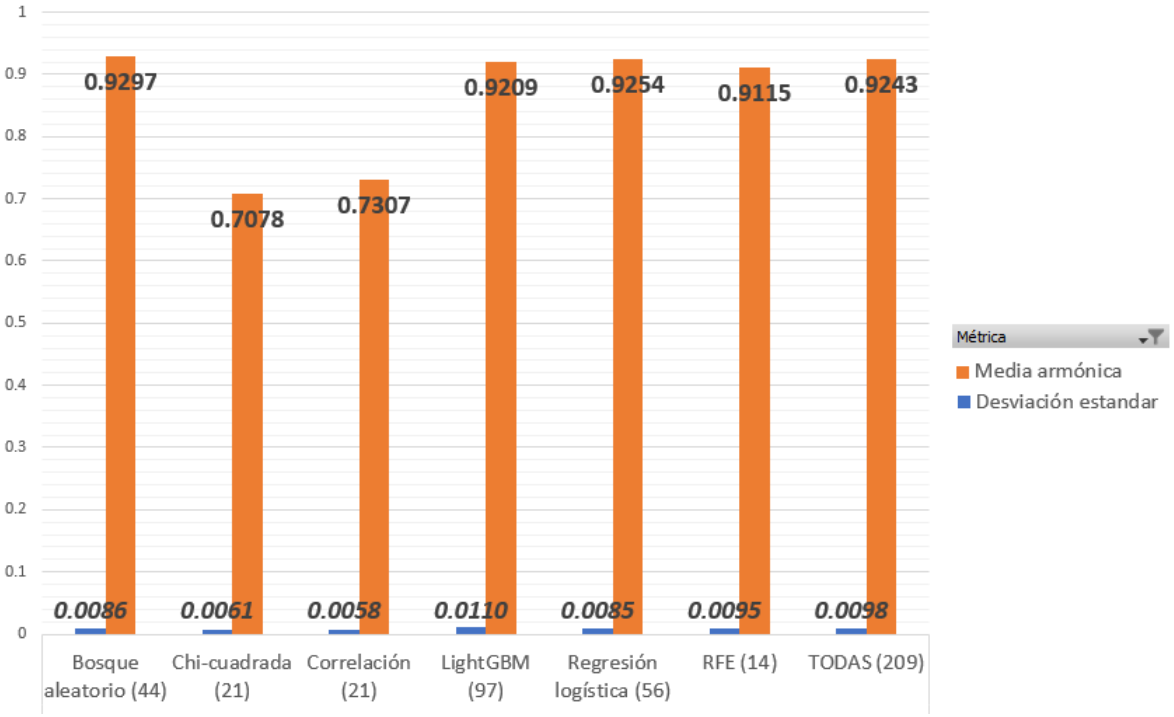

Figura 14. Resultados de clasificación en el primer escenario con el clasificador bosque aleatorio (con 100 árboles en los nodos). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

En la Figura 15 se muestran los resultados de bosque aleatorio con la configuración de 200 árboles (BA200). Nuevamente, las peores calificaciones se obtienen con los conjuntos de X^2 y correlación, y las mejores calificaciones con el conjunto bosque aleatorio con 0.9302 de calificación en la media armónica. Similarmente a la figura anterior (Fig. 14), los conjuntos con menores calificaciones se distinguen por tener una cantidad menor de características. Además, el segundo mejor conjunto fue dado por regresión logística con una calificación de 0.9260. En cuanto al STD, sobresalen los conjuntos con calificaciones más baja, mientras que los STD más alto representan a los conjuntos lightGBM, TODAS y RFE. Nuevamente, la diferencia de STD entre cada uno de estos conjuntos es menor a 0.01, o bien, de 0.004 entre bosque aleatorio y X^2 (el conjunto con mejor STD), lo cual no afecta al desempeño de la combinación BA200-BA. El conjunto TODAS está entre los mejores tres conjuntos, dada la información requerida por el clasificador para la construcción de los árboles. Ahora bien, se puede observar que bosque aleatorio y regresión logística son los conjuntos con mejor media armónica pero con mayor cantidad de características (desconsiderando TODAS) con un total de 44 y 56, respectivamente. Esto quiere decir que dichos conjuntos contienen características esenciales, a diferencia de lightGBM y RFE que contienen 97 y 14, respectivamente, que caen en el caso extremo de con-

tener demasiadas o pocas características. Una vez más se tiene que los métodos de selección de conjuntos parecen tener un papel en las calificaciones de media armónica (exceptuando el de TODAS): primero los conjuntos dados por métodos embebidos, seguido por método de envoltura y al final de filtro.

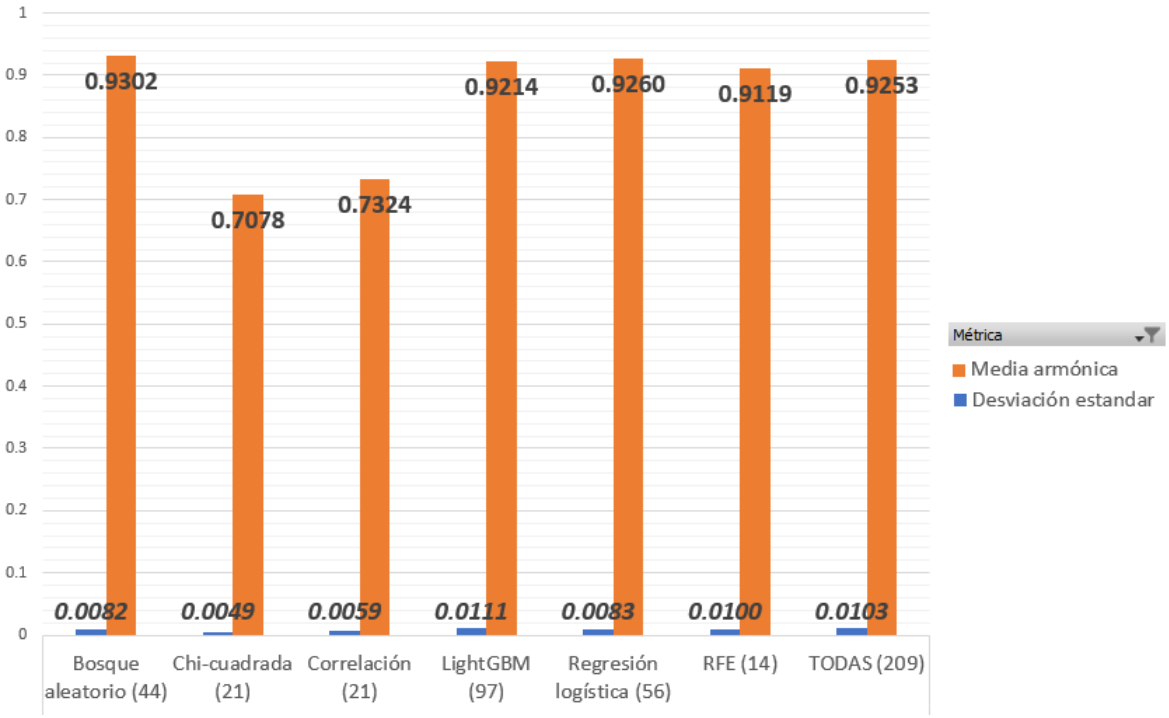

Figura 15. Resultados de clasificación en el primer escenario con el clasificador bosque aleatorio (con 200 árboles en los nodos). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Al comparar ambos clasificadores de bosque aleatorio, las mejores combinaciones de ambos clasificadores suceden con el conjunto de bosque aleatorio (*BA100-BA* y *BA200-BA*), mientras que las peores son con X^2 (*BA100-*X^2 y *BA200-*X^2) y correlación (*BA100-correlación* y *BA200-correlación*). Sin embargo, *BA200-BA* es mejor que *BA100-BA* por 0.01 de media armónica y 0.0004 de STD. De manera general, el clasificador de BA200 sobresale con menor desviación estándar y mejores calificaciones respecto a BA100, lo cual se debe a las distintas configuraciones que tiene el algoritmo. Por otro lado, los conjuntos dados por métodos embebidos obtuvieron la mejor calificación, mientras que los peores fueron los arrojados por métodos de filtrado.

En el caso del clasificador MLP con 50 capas ocultas (Fig. 16), se tiene como mejor conjunto RFE con 0.8545 de calificación, seguido por bosque aleatorio con 0.8057. En contraste, la calificación más baja fue obtenida por el conjunto de lightGBM, seguido por el conjunto con todas las características. Se puede observar que los conjuntos con calificaciones más bajas son aquellos que contienen un gran número de características. Sin embargo, dichos conjuntos tienen la menor desviación estándar, con un valor menor a 0.007. Por otro lado, RFE (que fue

el de mejor calificación con este clasificador) tiene un STD de 0.0591, pero no llega al 0.1, por lo cual no afecta en gran medida al desempeño del mejor conjunto y es algo a tener en consideración. En el caso de MLP, no le favorece manejar grandes conjuntos de características como se puede observar en el caso de TODAS y lightGBM, que constan de 209/209 y 97/209 características, respectivamente. De hecho los conjuntos con cantidades selectas son los que se ven beneficiados como es el caso de RFE, bosque aleatorio y correlación. En este caso, RFE (que contiene 14/209 características) tiene una alta dispersión de calificaciones por pliegue respecto al resto, la cual es considerablemente alta (aproximadamente ±6 % de calificación). Por lo cual, un conjunto más apropiado es el de bosque aleatorio con una calificación de 0.8057 y un STD de ±2 %. Cabe mencionarse que en este clasificador, no hubo un patrón particular de mejores medias armónicas respecto al método de selección usado en los conjuntos.

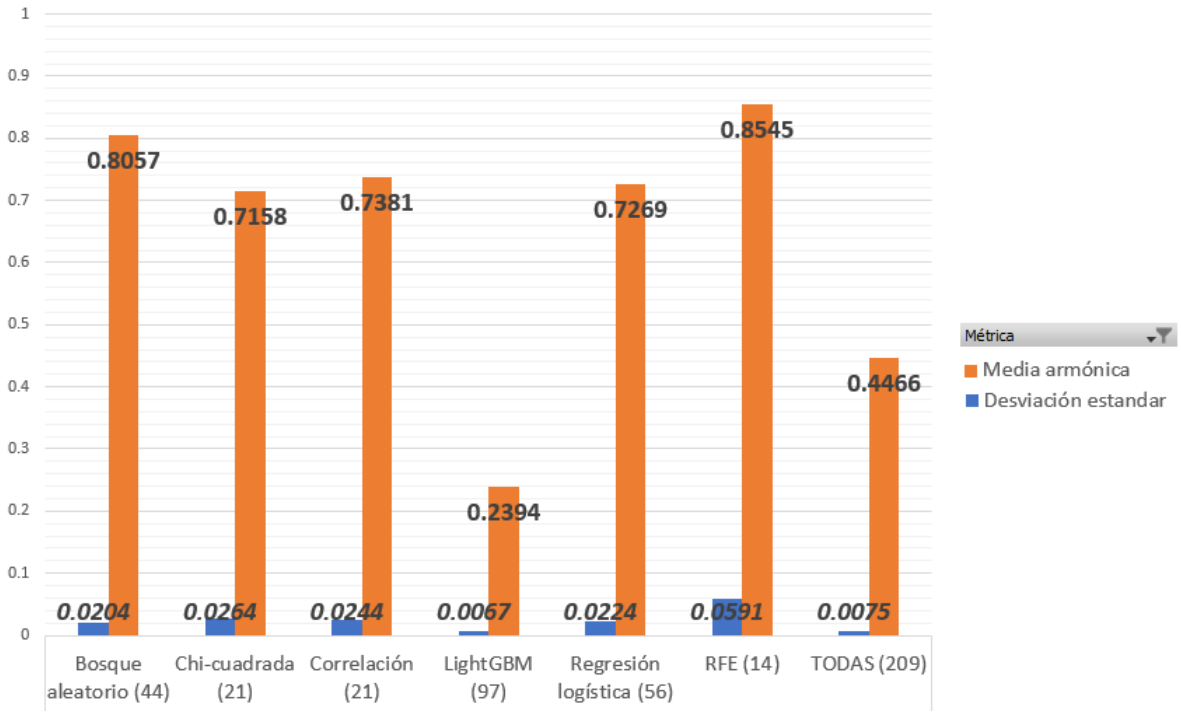

Figura 16. Resultados de clasificación en el primer escenario con el clasificador MLP (con 50 pliegues ocultos). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Por otro lado, en el caso de MLP con 100 capas ocultas (Fig. 17), el mejor conjunto es RFE con 0.8622 de calificación, el segundo mejor es bosque aleatorio con 0.07957 y el peor conjunto es el de todas las características, seguido por lightGBM. Similarmente a la figura anterior, los conjuntos con menor calificación son aquellos que tienen una cantidad grande de características. Ahora bien, en cuanto a los valores de STD, los valores más bajos los obtuvo lightGBM, TODAS y correlación. Nótese que RFE tiene un valor de 0.0223, lo cual respecto al más bajo es una diferencia de 0.011 y respecto al más alto es de 0.01. Por lo anterior, de manera general, se pretende decir que la combinación de MLP100-RFE no se ve afectado en gran medida para

ser la mejor combinación. El fenómeno observado en el escenario anterior entre cantidad de características y el funcionamiento del clasificador MLP se repite; los conjuntos más pequeños se ven beneficiados y los más grandes tienen una caída de valor en media armónica considerable. Sin embargo, las calificaciones subieron ligeramente al tener más capas ocultas (de 50 a 100 capas ocultas en los parámetros) respecto su contraparte. Desde otro punto de vista, utilizar el conjunto TODAS no es conveniente. Además, como en el caso de MLP50, no se obtuvo un patrón entre medias armónicas y método de selección de los conjuntos.

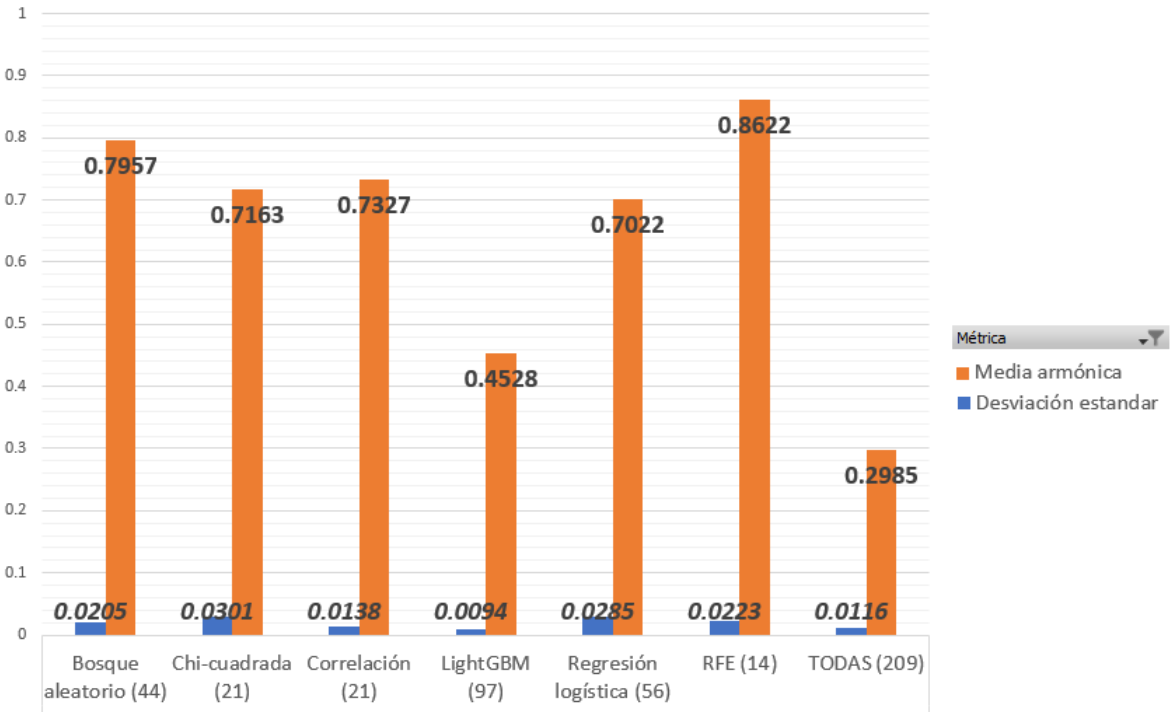

Figura 17. Resultados de clasificación en el primer escenario con el clasificador MLP (con 100 pliegues ocultos). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Al comparar ambos clasificadores MLP se puede observar que los mejores conjuntos son *MLP50-BA* y *MLP100-RFE*. Dichos conjuntos tienen calificación de 0.8057 y 0.8622, respectivamente, además de una diferencia de 0.06 de calificación posicionando la combinación de *MLP100-RFE* sobre la de *MLP50-BA*. Cabe destacar que BA cuenta con 44 características, mientras que RFE tan sólo 14. Además, difieren en el conjunto con calificación más baja, ya que uno es con *lightGBM* (MLP50-lightGBM) y el otro con el de todas las características (*MLP50-TODAS*), aunque si se observa el panorama general, estos conjuntos tienen las calificaciones más bajas en ambos casos. De manera general, se puede inferir que el clasificador de MLP trabaja mejor con una cantidad selecta de características, entre mayor sea la información, menor es su eficiencia, así como sucedió con los conjuntos de TODAS y lightGB.

Para el clasificador de árbol de decisión también se tienen dos versiones, los resultados del primero se pueden observar en la Figura 18, que corresponden al clasificador configurado para

realizar la mejor partición posible en los nodos. Los conjuntos con mejor calificación son bosque aleatorio, TODAS y regresión logística, en dicho orden, con bosque aleatorio en la cabecera con un total de 0.8895. En contraste, los peores conjuntos son aquellos que tienen un menor número de características, en particular los que fueron generados por los algoritmos seleccionadores por filtro, X^2 y correlación. Notablemente, los conjuntos con menor media armónica resultaron ser los que arrojaron una menor dispersión de datos. En cuanto a la desviación estándar, todos los conjuntos tienen un valor menor a 0.0085. Bosque aleatorio está en los valores medios de STD con un total 0.0072, mientras que los conjuntos con STD alto tienen valores arriba de 0.0075, la cual no es una gran diferencia. Ahora bien, la cantidad de elementos en los conjuntos afecta al funcionamiento del clasificador, por eso se puede notar una caída en los conjuntos generados por los seleccionadores de filtro (como se mencionó previamente), seguido por RFE (el único conjunto generado por un algoritmo de selección por envoltura) que no tiene una calificación mala porque la mayoría de las características que tiene están contenidos en los conjuntos con media armónica superior (los conjuntos dados por algoritmos de selección por método embebido). Sin embargo, el conjunto TODAS tiene una calificación alta (respecto a los conjuntos de regresión logística y lightGBM), más no supera a bosque aleatorio. Lo anterior indica que bosque aleatorio, aparte de tener una cantidad suficiente de información, las características que posee son de calidad y/o relevantes para el objetivo del presente escenario.

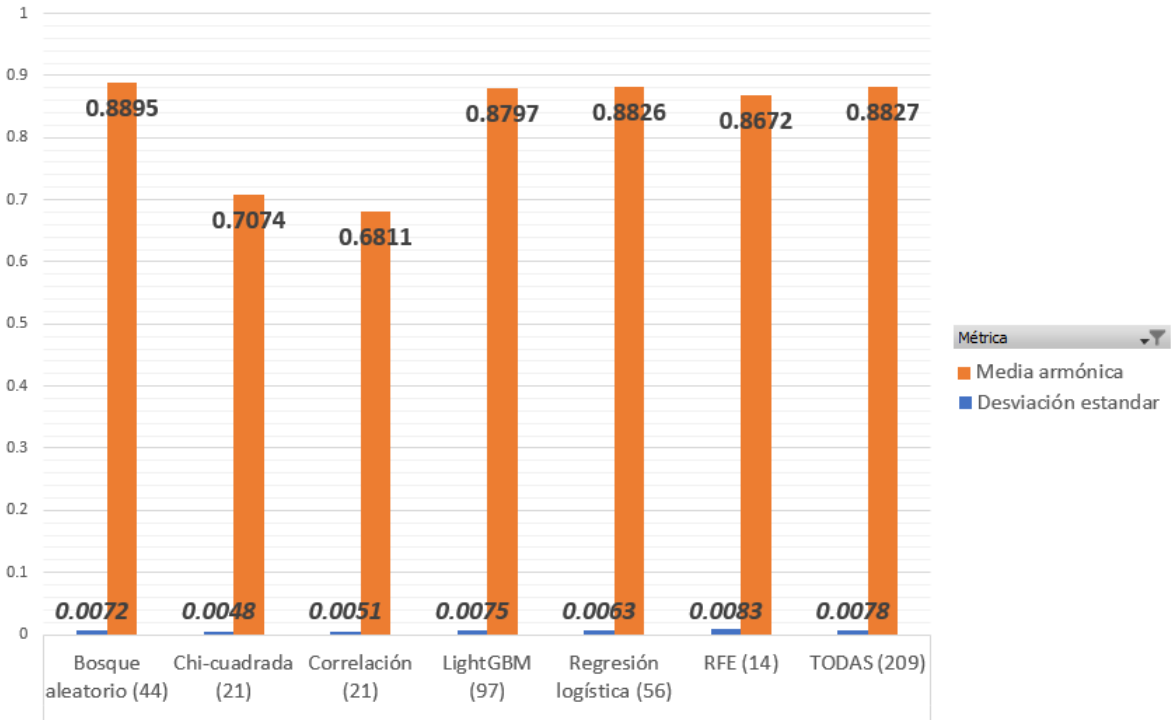

Figura 18. Resultados de clasificación en el primer escenario con el clasificador AD (con la mejor partición). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

En el clasificador AD con partición aleatoria en los nodos, se obtuvieron resultados similares, sin embargo, los resultados de los mejores conjuntos cambiaron ligeramente (Fig. 19). En primer lugar se tiene a bosque aleatorio, seguido por regresión logística y en tercero a TODAS. Los conjuntos con bajas calificaciones son los mismos que en la configuración anterior, X^2 y correlación, donde correlación tiene la menor calificación con un total de 0.6809. La desviación estándar para los conjuntos es menor a 0.01 (o bien, ±1 %). Notablemente, el más alto (RFE) obtuvo un valor de 0.0099, mientras que los valores más bajos fueron dados por X^2, bosque aleatorio y regresión logística. Lo interesante de esto es que la mejor combinación por media armónica, se encuentra entre las combinaciones con menor STD, lo cual favorece al conjunto *AD(mejor)-BA*. Ahora bien, se puede observar que los conjuntos dados por los algoritmos de selección por filtro tienen las calificaciones más bajas, seguido por el conjunto dado por el algoritmo de selección por envoltura (RFE), mientras que las calificaciones más altas fueron dadas por los conjuntos obtenidos por los algoritmos de selección por métodos embebidos (exceptuando lightGBM que es inferior a TODAS).

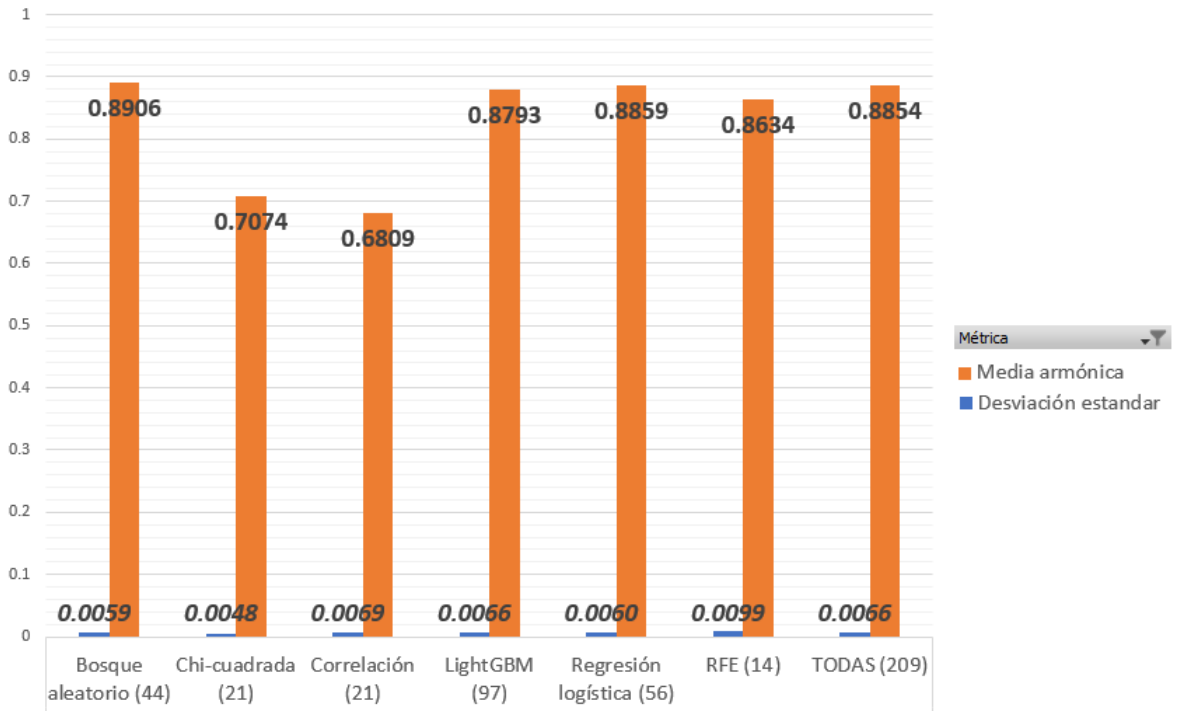

Figura 19. Resultados de clasificación en el primer escenario con el clasificador AD (con partición aleatoria). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Al comparar las mejores calificaciones arrojadas por el mismo clasificador pero con distintos parámetros, se tiene que por una diferencia de 0.01 en la media armónica, *AD(aleatorio)-BA* fue mejor que *AD(mejor)-BA*. De hecho, los resultados en general son parecidos, pero destacan los del clasificador con partición aleatoria en los primeros cuatro conjuntos con mejor calificación. Esto indica que en cuanto a media armónica, el clasificador con partición aleatoria es superior

en este escenario. En cuanto a desviación estándar, ambos clasificadores obtuvieron buenos resultados, con valores menores a 0.01. Sin embargo, AD(*aleatorio*) destaca por tener más conjuntos con valores más bajos. De manera general, se puede decir que el clasificador de árbol de decisión funciona bien con los conjuntos dados por algoritmos de selección por método embebido.

Finalmente, se presentan los resultados obtenidos por el clasificador *Naive Bayes* (Fig. 20). Se puede observar que los mejores conjuntos son arrojados por correlación y X^2 (contrario a los demás clasificadores), con un total de 0.7269 y 0.7249 respectivamente. Sin embargo, los conjuntos con bajas calificaciones (igual o menores a 0.0473) son lightGBM, regresión logística y el de todas las características. Bosque aleatorio y RFE obtuvieron calificaciones medianas con este clasificador, aunque comparado con otros clasificadores, son bajas. Particularmente con este clasificador se puede hacer notar que trabaja mejor con conjuntos generados por seleccionadores de filtro, mientras que el conjunto de bosque aleatorio está en tercera posición como mejor conjunto, sin embargo, su media armónica es de 0.4830. Ahora bien, al considerar la desviación estándar, el valor más alto lo tiene bosque aleatorio con 0.0575, mientras que los valores más bajos lo tienen X^2 con un total de 0.0064 y correlación con 0.0072. Esto favorece la combinación NB-correlación ya que fue la combinación con mejor media armónica.

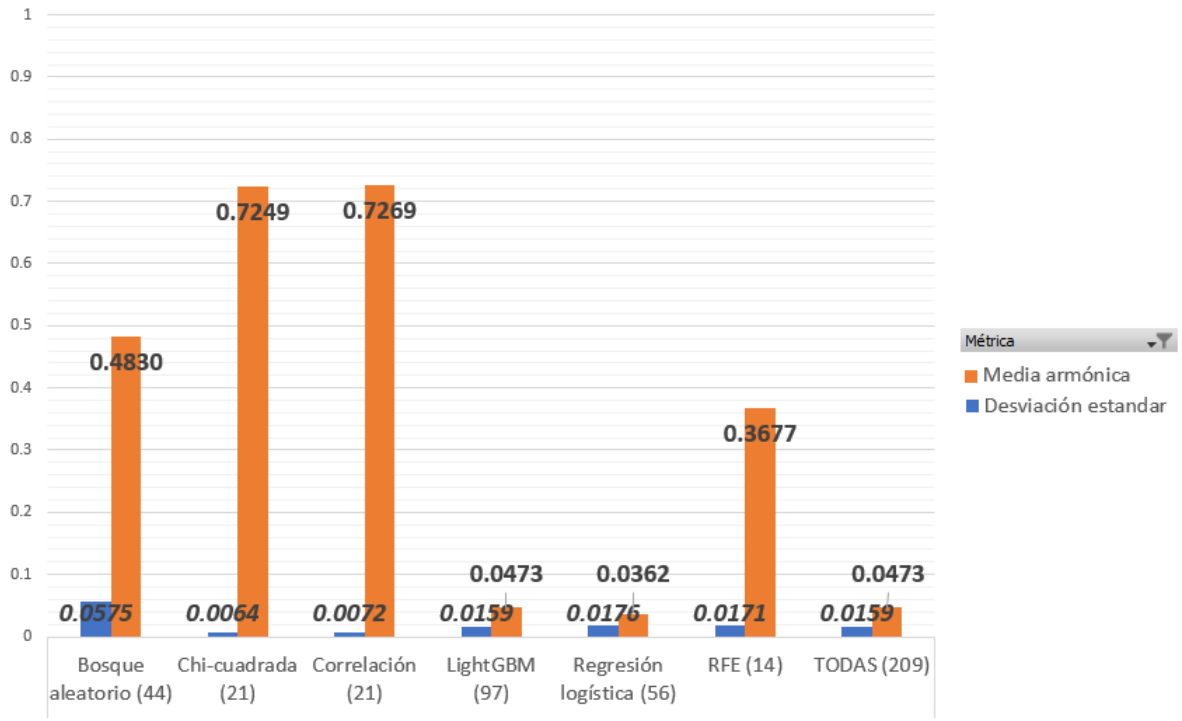

Figura 20. Resultados de clasificación en el primer escenario con el clasificador *Naive Bayes*. El número entre paréntesis corresponde al número de características contenido en cada conjunto.

El análisis anterior nos lleva a concluir que las mejores combinaciones clasificador-conjunto

en este escenario durante los entrenamientos fueron de: *SVM-RFE,BA100-BA, BA200-BA, MLP50-BA, MLP100-RFE, AD(mejor)-BA, AD(aleatorio)-BA* y *NB-Correlación*. Además, los mejores conjuntos fueron RFE, BA y regresión logística, ya que estos presentan mejores calificaciones sin importar el clasificador de manera general. La única excepción fue *NB*, ya que sobresalen los conjuntos generados por métodos de filtro. Por otro lado, los peores conjuntos se puede considerar que son los de correlación y X^2. Esto podría deberse a la poca cantidad de características contenidas, ya que al comparar los indicadores en los conjuntos con los que arrojaron una mejor calificación, a los conjuntos de correlación y X^2 les hacían falta indicadores que se repetían en otros conjuntos, o bien, características dominantes. Para determinar un mejor y peor clasificador, es necesario hacer otras consideraciones, pero esto se discutirá en la sección siguiente.

5.2.1. Análisis con media armónica: Pruebas

En esta sección se muestran los resultados obtenidos con el conjunto independiente de datos para cada una de las combinaciones clasificador-conjunto de características seleccionado en la etapa anterior. Para determinar la mejor de dichas combinaciones se comparan primero los entrenamientos con las pruebas de cada combinación, para después incluir el tiempo de prueba como factor para seleccionar la mejor combinación durante las pruebas.

En la Tabla 10 se tienen las mejores combinaciones clasificador-conjunto propuestos en la sección anterior. En ella se puede apreciar una comparación de los resultados obtenidos durante las etapas de entrenamiento y prueba. En particular, los resultados durante la prueba son similar o un poco más bajos que los resultados durante los entrenamientos. Existen casos en los que las calificaciones obtenidas durante la prueba superan a las de entrenamiento, sin embargo, es por menos de 0.5 %, lo cual no supone un problema. Dicho esto, se puede observar que las combinaciones con calificaciones más altas son las de BA100-BA y BA200-BA, seguido por AD(*mejor*)-BA y AD(*aleatorio*)-BA. Además, las calificaciones más bajas son dados por SVM y NB. Por otro lado, las combinaciones con desviación estándar más alto son los propuestos junto con el clasificador de MLP. Nótese que las calificaciones obtenidas en las combinaciones con dicho clasificador son más bajas las de AD. Las combinaciones con dispersión de datos más baja son AD(*aleatorio*)-BA, SVM-RFE, AD(*mejor*)-BA y NB-correlación, en ese orden.

Finalmente se tiene el tiempo de clasificación el cual cobra suma relevancia para el problema propuesto en este proyecto, ya que por como se definieron los máximos y mínimos relativos, no debe de tomar más de 5 minutos en llevar a cabo la tarea, por esto, las combi-

Tabla 10. Tabla comparativa entre las mejores combinaciones durante las etapas de entrenamiento (μ_{HE}) y prueba (H_P) en el primer escenario. Nótese que σ es la desviación estándar obtenida en la etapa de entrenamiento, mientras que T_P se refiere al tiempo de clasificación en pruebas.

Combinación	μ_{HE}	σ	H_P	T_P (min.)
SVM-RFE	76.77 %	±0.6 %	76.89 %	1.7268
BA100-BA	92.97 %	±0.86 %	92.42 %	9.1021
BA200-BA	93.02 %	±0.82 %	92.48 %	18.1804
MLP50-BA	80.57 %	±2.04 %	78.24 %	49.0992
MLP100-RFE	86.22 %	±2.23 %	86.84 %	49.0802
AD(*mejor*)-BA	88.95 %	±0.72 %	88.42 %	0.9930
AD(*aleatorio*)-BA	89.06 %	±0.59 %	88.47 %	0.0979
NB-correlación	72.69 %	±0.72 %	72.70 %	0.0169

naciones relacionadas con los clasificadores de MLP y BA son automáticamente descartados, mientras que las mejores combinaciones en cuestión de tiempo son NB, AD(*aleatorio*)-BA y AD(*mejor*)-BA. Tomando en cuenta todas la media armónica, desviación estándar y tiempo, las mejores combinaciones son AD(*mejor*)-BA y AD(*aleatorio*)-BA, siendo el segundo la mejor opción.

En la Fig. 21 se puede apreciar únicamente los resultados obtenidos durante las pruebas. En ella se hace notar que los clasificadores de BA y MLP tardan más de lo esperado, mientras que en calificaciones no destacan SVM y MLP. En general, el desempeño de MLP fue bastante mediocre para el tiempo que tardó y SVM fue rápido, mas no lo suficientemente bueno, por esto, dichos clasificadores se omitirán en futuros escenarios.

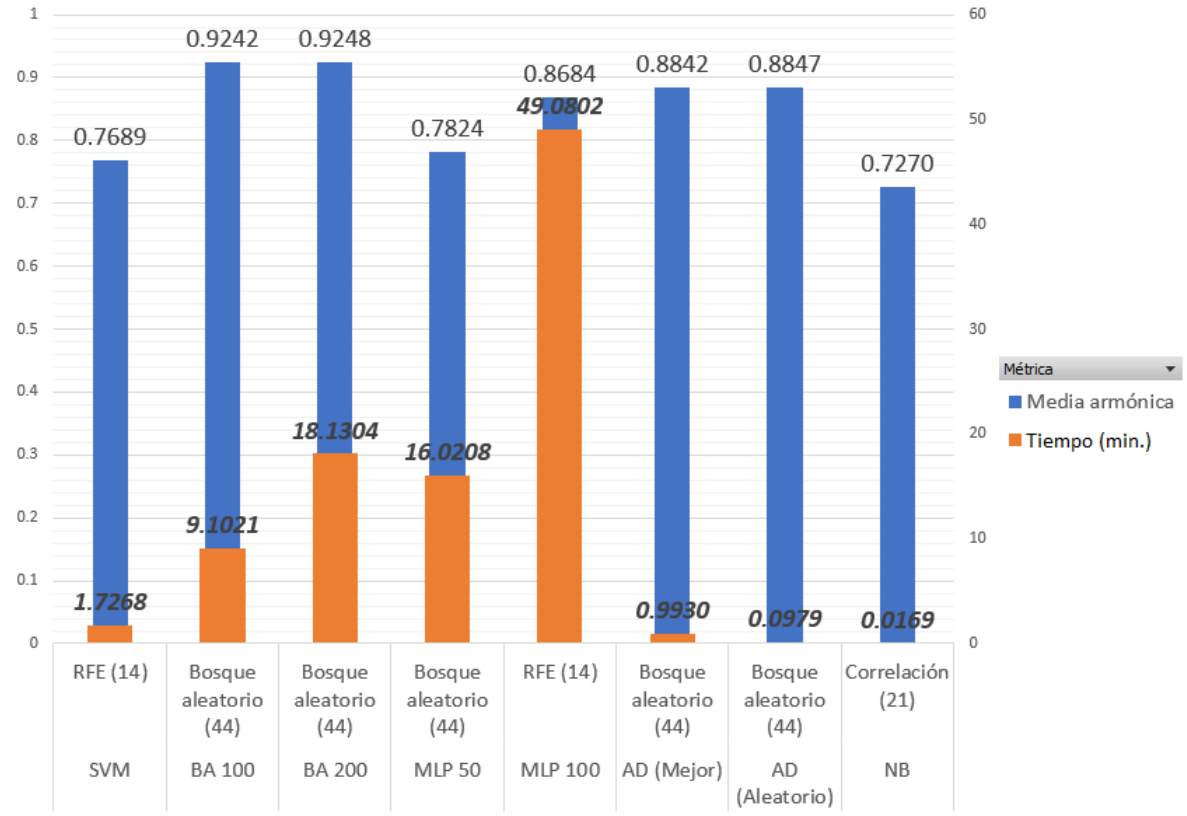

Figura 21. Resultados de las mejores combinaciones clasificador-conjunto durante la prueba en el primer escenario. El número entre paréntesis corresponde al número de características contenido en cada conjunto. El primer renglón de algoritmos se refiere al conjunto, mientras que el segundo renglón de algoritmos se refiere al clasificador involucrado.

Por otro lado, se demuestra que arrojar buenas calificaciones, como en el caso de las combinaciones pertenecientes a los clasificadores de bosque aleatorio, si no se realiza la clasificaciones en el tiempo requerido para el problema, no es una buena combinación. Esto denota que, en cuestión de tiempo, los clasificadores de SVM, AD y NB son excelentes, mientras que por calificación, los mejores son bosque aleatorio y árbol de decisión.

Otro aspecto a mencionar es que el conjunto de bosque aleatorio está presente en casi todas las combinaciones, posicionándolo como el mejor conjunto para este escenario. De manera general, se puede observar que AD es de los mejores clasificadores para este problema en cualquiera de sus dos variantes, mientras que por velocidad y calificación, la mejor combinación es AD(*aleatorio*)-BA.

5.3. Conclusiones parciales

En esta sección se presentan las conclusiones adquiridas para el primer escenario, en el cual el problema es la identificación de oportunidades de arbitraje.

- Los seleccionadores de filtro arrojan los conjuntos menos óptimos para el problema planteado en este proyecto, seguido por los conjuntos dados por métodos de envoltura, mientras que los mejores son generados por métodos embebidos como lo es el conjunto de bosque aleatorio.
- Los conjuntos de RFE y bosque aleatorio tienen el desempeño más destacable en combinación con casi todos los clasificadores propuestos, mientras que el rendimiento general más bajo lo tiene los conjuntos de X^2 y correlación.
- Durante las pruebas se mostró que los clasificadores más lentos son MLP y bosque aleatorio, por otro lado, los más veloces son SVM, NB y árbol de decisión. De estos, MLP, SVM y NB no presentaron buenas calificaciones de media armónica en ninguna de las dos etapas respecto al resto de los clasificadores.
- En los conjuntos tienden a dominar los indicadores de tipo volatilidad y tendencia, a su vez, los menos frecuentes son los de tipo volumen y *otros*. Esto implica que las oportunidades de arbitraje son identificadas de una mejor manera con la volatilidad y tendencia, por otro lado, el volumen tiene poco efecto a esta herramienta y/o estrategia.
- Con los resultados presentes y el análisis dado, la mejor combinación para el primer escenario es AD(*aleatorio*)-BA.

Capítulo 6. Resultados y discusión del segundo escenario: tipo de operación en oportunidades de arbitraje

En este capítulo se presentan los resultados de la selección de características, entrenamientos y pruebas del segundo escenario, el cual pretende determinar si una oportunidad de operación de arbitraje en el mercado de divisas es compra o venta. Asimismo se interpretan y discuten dichos resultados, qué tipo de características fueron tomadas, los mejores conjuntos, clasificadores y combinaciones.

6.1. Conjunto de características

El detalle de los conjuntos generados puede consultarse en los Apéndices de este trabajo (Tabla 32), además de cada característica contenida. En la Figura 22, se puede observar que el tipo de característica dominante es volatilidad, seguido por los indicadores de tendencia. En este escenario tienen menos presencia los indicadores de volumen, osciladores y *otros*. Dicho escenario busca determinar con gran eficiencia el tipo de operación que se pueda llevar a cabo sobre las oportunidades encontradas, lo que implica que para determinar si es una compra o una venta en una operación de arbitraje, la tendencia y la volatilidad del mercado tendrán un gran peso en la decisión. En general, el número de características obtenidos por conjunto fue menor respecto al escenario anterior, por lo que se puede inferir que no se requiere de tanta información para valorar el tipo de operación. Por otro lado, la proporción de indicadores tomados por conjunto se respeta. Además, el mayor número lo tuvieron los conjuntos generados por métodos embebidos, seguido de los métodos de envoltura y al final los de filtro.

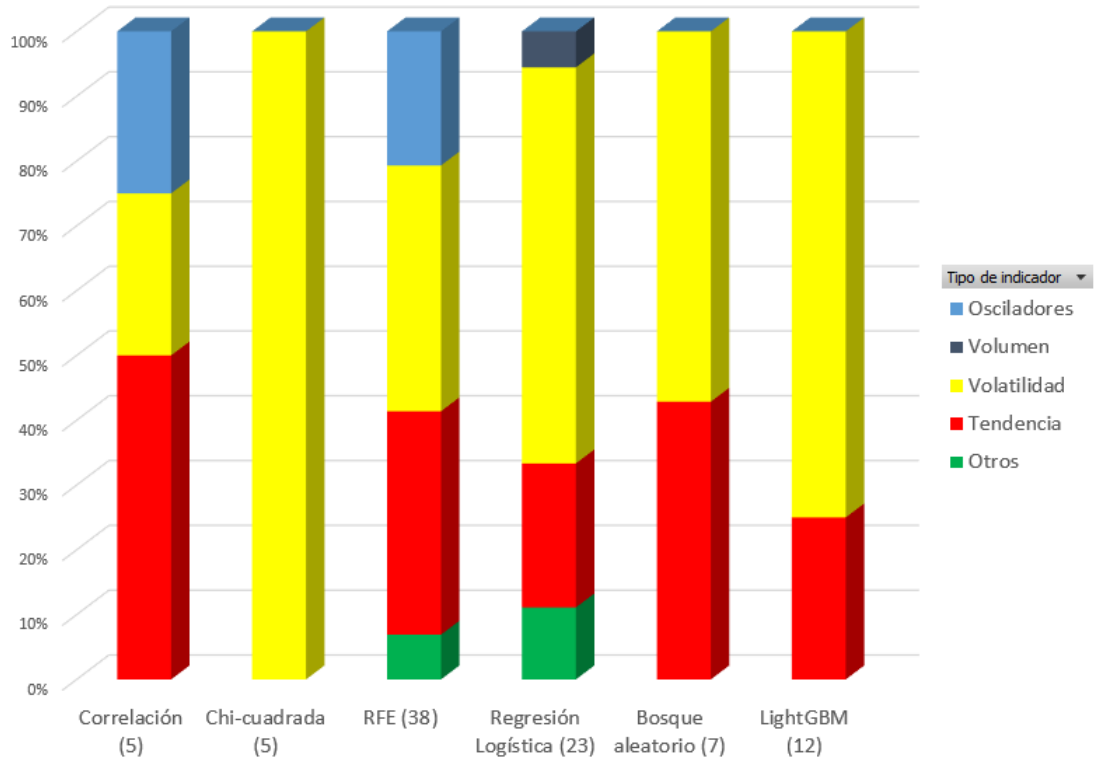

Figura 22. Conjuntos de características generados para el segundo escenario y el porcentaje de tipo de indicadores técnicos en cada uno de ellos. El número entre paréntesis corresponde al número de características contenido en cada conjunto.

En particular, los indicadores dominantes en los conjuntos son los siguientes:

- Tendencia: *MACD*
- Volatilidad: *Bollinger Bands* y *Donchian Channel*

Por el contrario, los indicadores sin utilizar fueron:

- Tendencia: *ADX, Vortex Indicator, DPO, KST Oscillator* e *Ichimoku*.
- Osciladores: *Money Flow Index, True Strenght Index, Ultimate Oscillator, Stochastic Oscillator, Williams %R, Awesome Oscillator, KAMA* y *ROC*.
- Volatilidad: *Average True Range.*
- Volumen: *ADI, OBV, CMF, Force Index, EoM, NVI* y *VWAP.*
- Otros: *Cumulative Return*.

Observe que la mayor parte de los indicadores no usados pertenecen a las categorías de osciladores y volumen. Los indicadores de dichas categorías parecen ser más presentes en los conjuntos de manera porcentual, pero es por la cardinalidad de elementos por conjunto.

6.2. Análisis con media armónica: Entrenamiento

En esta sección se discute la eficiencia de los clasificadores de bosque aleatorio, árbol de decisión y *Naive Bayes* (NB) en el segundo escenario. Para determinar las mejores combinaciones clasificador-conjunto de características se toman las calificaciones más altas (media armónica) por clasificador, para posteriormente comparar entre el mejor de cada uno.

Para el clasificador Bosque Aleatorio (BA) con la configuración de 100 árboles (Fig. 23) las calificaciones más bajas fueron dadas por los conjuntos generados por los algoritmos de selección por filtrado (X^2 y correlación). Por otro lado, las mejores calificaciones fueron dadas por el resto de los conjuntos, con una calificación perfecta de 1/1. Sin embargo, la diferencia entre las calificaciones más bajas y más altas es menor a 1 %. Esto da pie a usar cualquier conjunto ante estos resultados. Además, la desviación estándar es nula en los conjuntos con calificación de 1, solamente los conjuntos de X^2 y correlación tienen un STD diferente de 0, la cual es menor a 2.5 %. En este caso, parece no haber distinción entre usar todas o un subconjunto de características, por lo cual, lo ideal es optar por el conjunto con menor número de características y con calificación de 1, o sea, bosque aleatorio, el cual tiene un total de siete características.

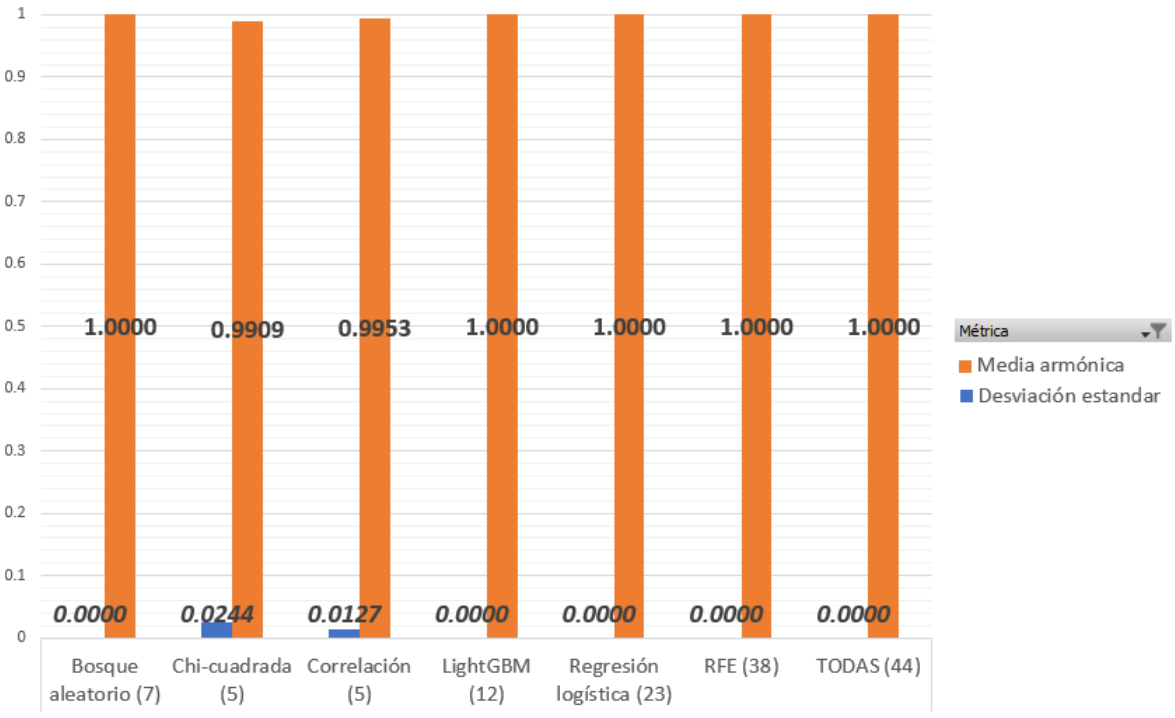

Figura 23. Resultados de clasificación en el segundo escenario con el clasificador bosque aleatorio (con 100 árboles en los nodos). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Este mismo fenómeno de buenas calificaciones se repite para el clasificador de BA con 200

árboles (Fig. 24), con una diferencia prácticamente nula respecto a su versión con 100 árboles, ya que el conjunto de correlación cambia en 0.0001 de valor de media armónica y difiere en 0.001 de STD. De manera similar, podemos concluir que lo mejor es optar por el conjunto con menor número de características, bosque aleatorio, con un total de siete.

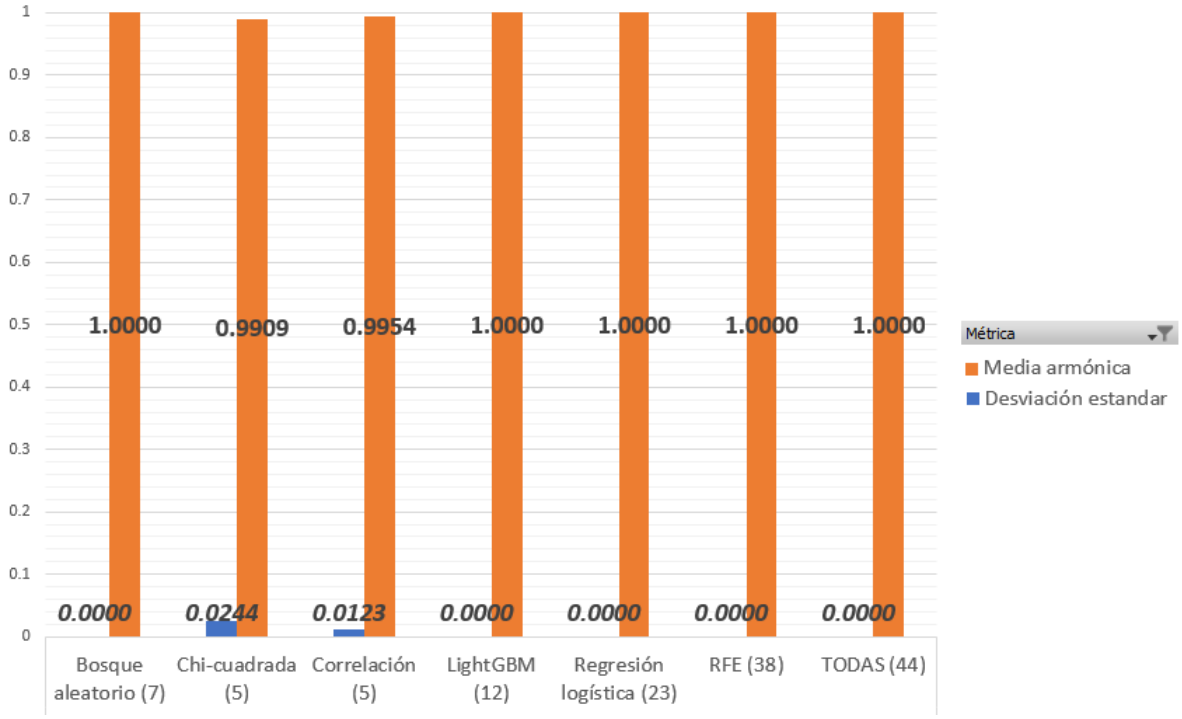

Figura 24. Resultados de clasificación en el segundo escenario con el clasificador bosque aleatorio (con 200 árboles en los nodos). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Así mismo, con el clasificador de árbol de decisión con la mejor partición en los nodos (Figura 25), se puede observar que los conjuntos generados por algoritmos de filtro son los que difieren de las medias armónicas y desviaciones estándar excelentes. Por otro lado, parece no haber problema alguno con usar el total o parcial número de características, alentando a nuevamente usar el conjunto de bosque aleatorio que contiene tan sólo siete características de un total de 44.

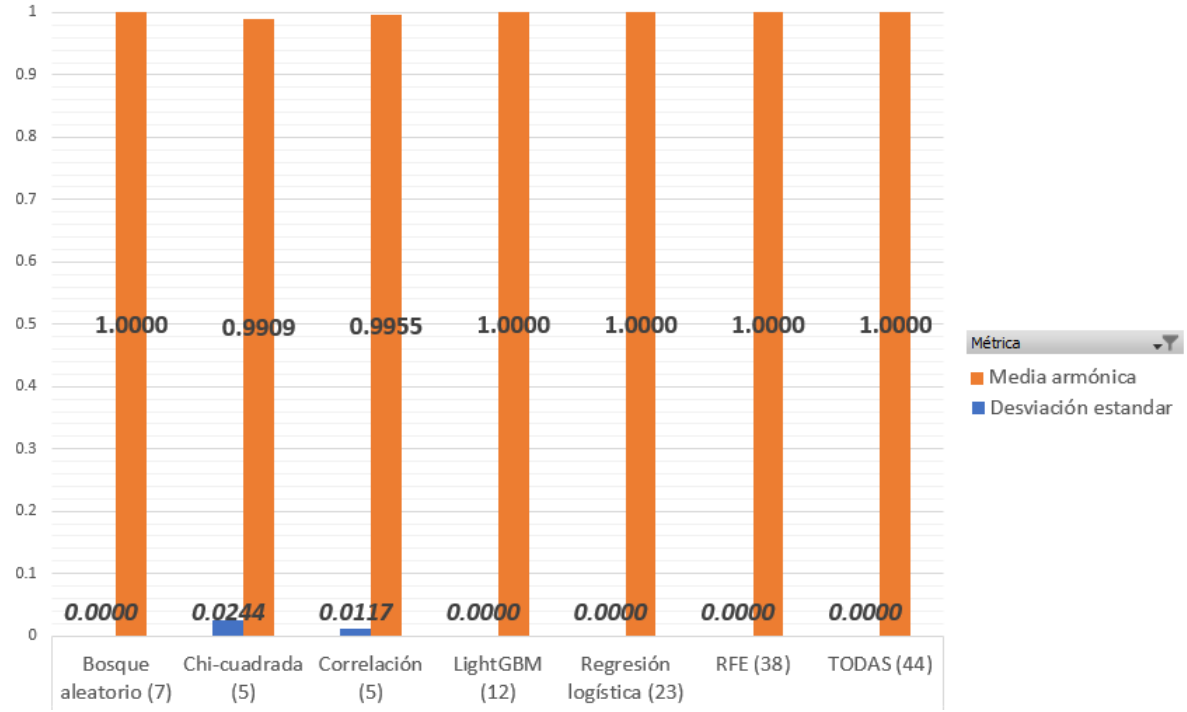

Figura 25. Resultados de clasificación en el segundo escenario con el clasificador AD (con la mejor partición). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Con la variante de partición aleatoria en los nodos (Figura 26), se obtiene la misma conclusión que con la variante de mejor partición en los nodos, ya que los resultados son prácticamente iguales, la única distinción aparente es la desviación estándar del conjunto de correlación que incrementa por 0.0026 y la media armónica que incrementa por 0.0002. De estos resultados se puede inferir que el conjunto indicado para usar es el de bosque aleatorio por su buen desempeño, nula dispersión entre los múltiples pliegues de validación cruzada y el número de características.

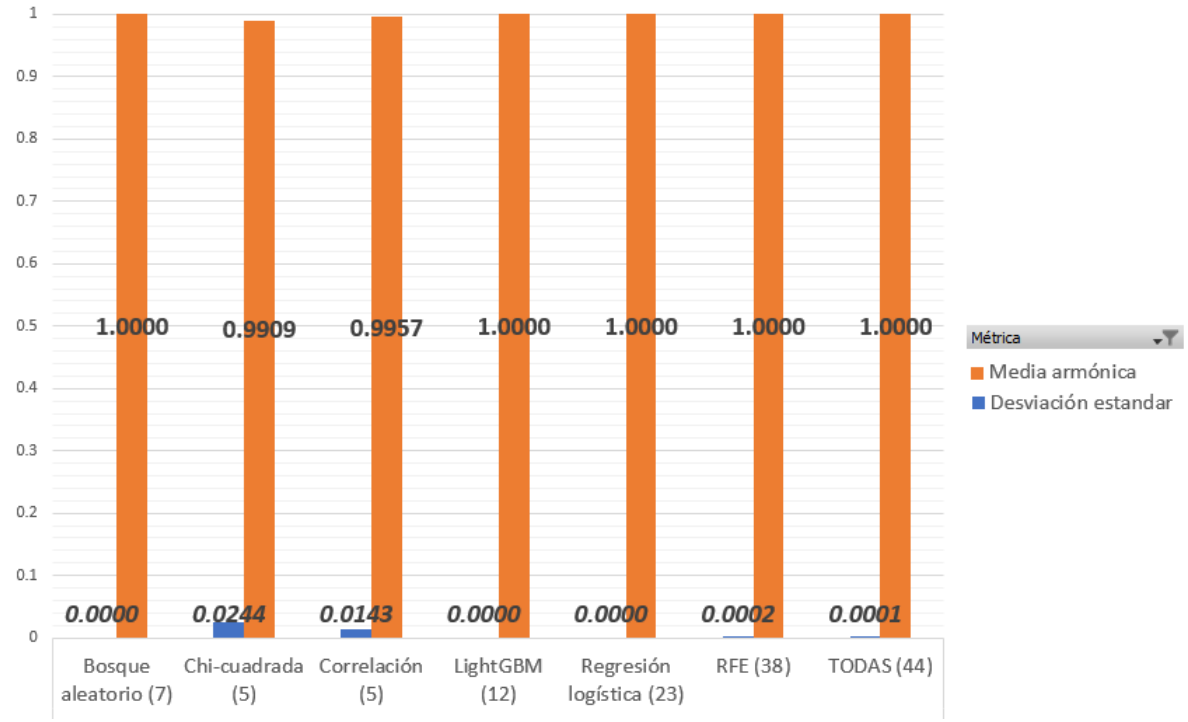

Figura 26. Resultados de clasificación en el segundo escenario con el clasificador AD (con partición aleatoria). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Ahora bien, para el clasificador de Naive Bayes (Figura 27), las mejores calificaciones son dadas por los conjuntos de bosque aleatorio, correlación y lightGBM, mientras que las más bajas corresponden a X^2, RFE, regresión logística y TODAS. El dato que hace la distinción entre los mejores conjuntos es la desviación estándar, ya que bosque aleatorio arrojó un valor de 0, mientras que el resto es superior a 0. Por otro lado, este clasificador da la oportunidad en segundo lugar al conjunto de correlación con un STD prácticamente nulo (0.001) y una media armónica de 1/1. Además, dicho conjunto contiene tan sólo cinco características, que si bien es menor que la cardinalidad del conjunto de bosque aleatoria, no vale la pena incrementar en 0.001 la desviación estándar por calcular dos características menos, cuando dichas características no toman más de 0.1 segundos en ser calculadas.

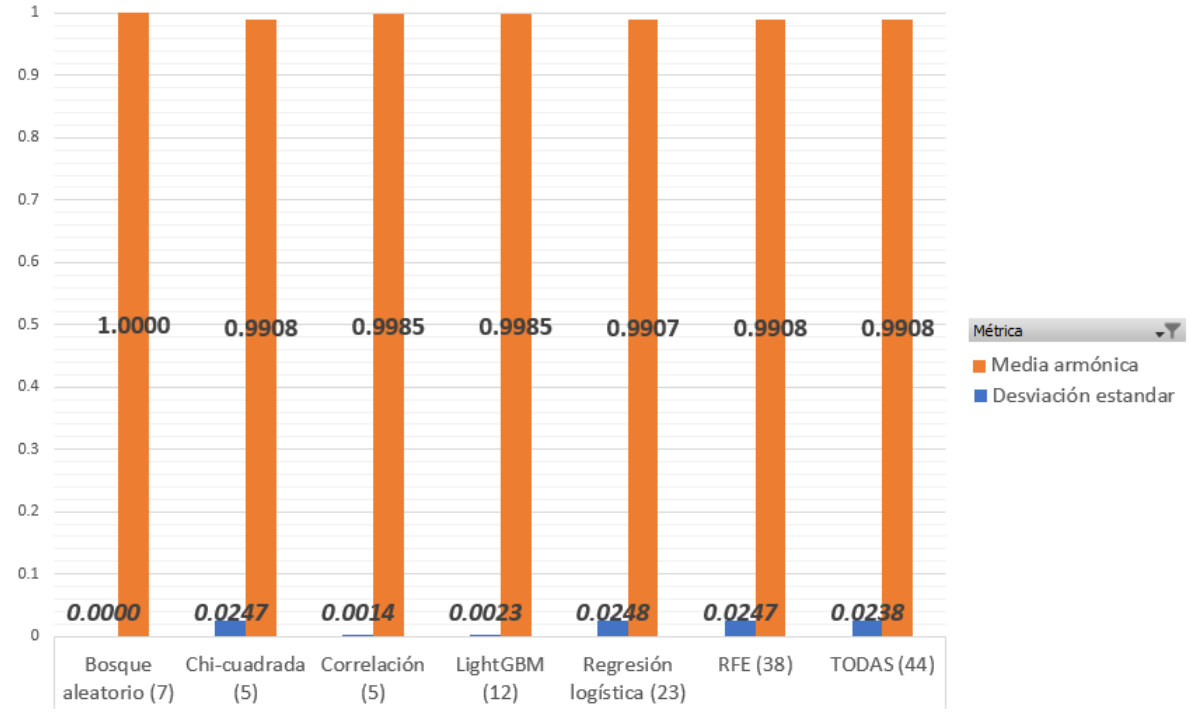

Figura 27. Resultados de clasificación en el segundo escenario con el clasificador *Naive Bayes*. El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Los gráficos anteriores indican que en este caso es mejor guiarse por el número de características ya que los valores de media armónica son prácticamente los mismos con un valor de 1, mientras que la desviación estándar es casi o igual a 0. Los conjuntos con baja calificación son los de correlación y X^2, mientras que el mejor conjunto en todos los clasificadores fue el de bosque aleatorio. Por lo tanto, para determinar con mayor precisión y rigor la mejor combinación, se debe analizar la posible interacción de este conjunto con cada uno de los clasificadores durante las pruebas en la siguiente sección.

6.2.1. Análisis con media armónica: Pruebas

En esta sección se muestran los mejores resultados de cada uno de los clasificadores (entrenados en la etapa anterior) con su mejor conjunto, o bien, la mejor combinación. La mejor combinación durante el segundo escenario es determinada primero al comparar los entrenamientos con las pruebas de cada una de las mejores combinaciones, después se toman en consideración el tiempo como factor para determinar la mejor combinación durante las pruebas.

En la Tabla 11 se observa que las calificaciones obtenidas tanto en las pruebas como en

los entrenamientos fueron prácticamente perfectas, la desviación estándar es de 0, pero las combinaciones difieren en la cantidad de tiempo con la cual llevaron a cabo la tarea pertinente al escenario. En este caso, todos los tiempos son menores a medio segundo. Sin embargo, sobre todos ellos destaca la combinación AD(*aleatorio*)-BA, la cual pudo determinar si una operación es compra o venta con una efectividad del 100 % en tan solo 0.0035 minutos, o bien, menos de medio segundo.

Tabla 11. Tabla comparativa entre las mejores combinaciones durante las etapas de entrenamiento (μ_{HE}) y primera prueba (H_{P1}) en el primer escenario. Nótese que σ es la desviación estándar obtenida en la etapa de entrenamiento, mientras que T_{P1} se refiere al tiempo de clasificación en la primera prueba.

Combinación	μ_{HE}	σ	H_{P1}	T_{P1} (min.)
BA100-BA	100 %	±0 %	100 %	0.254741
BA200-BA	100 %	±0 %	100 %	0.508826
AD(*mejor*)-BA	100 %	±0 %	100 %	0.011234
AD(*aleatorio*)-BA	100 %	±0 %	100 %	0.003561
NB-BA	100 %	±0 %	100 %	0.004047

En la Fig. 28 se puede apreciar los resultados de dicha prueba sin la presencia de los resultados del entrenamiento. Dicha figura acentúa el descarte de los conjuntos relacionados con el clasificador de BA que cuentan con los tiempos más altos (BA100-BA y BA200-BA), que comparados con el resto no son tardados por ser todos menores a medio minuto. Los resultados anteriores parecen ser demasiado buenos, por lo cual se realizó una segunda prueba con un conjunto independiente de datos distinto al primero (ver Sección 3.2.5).

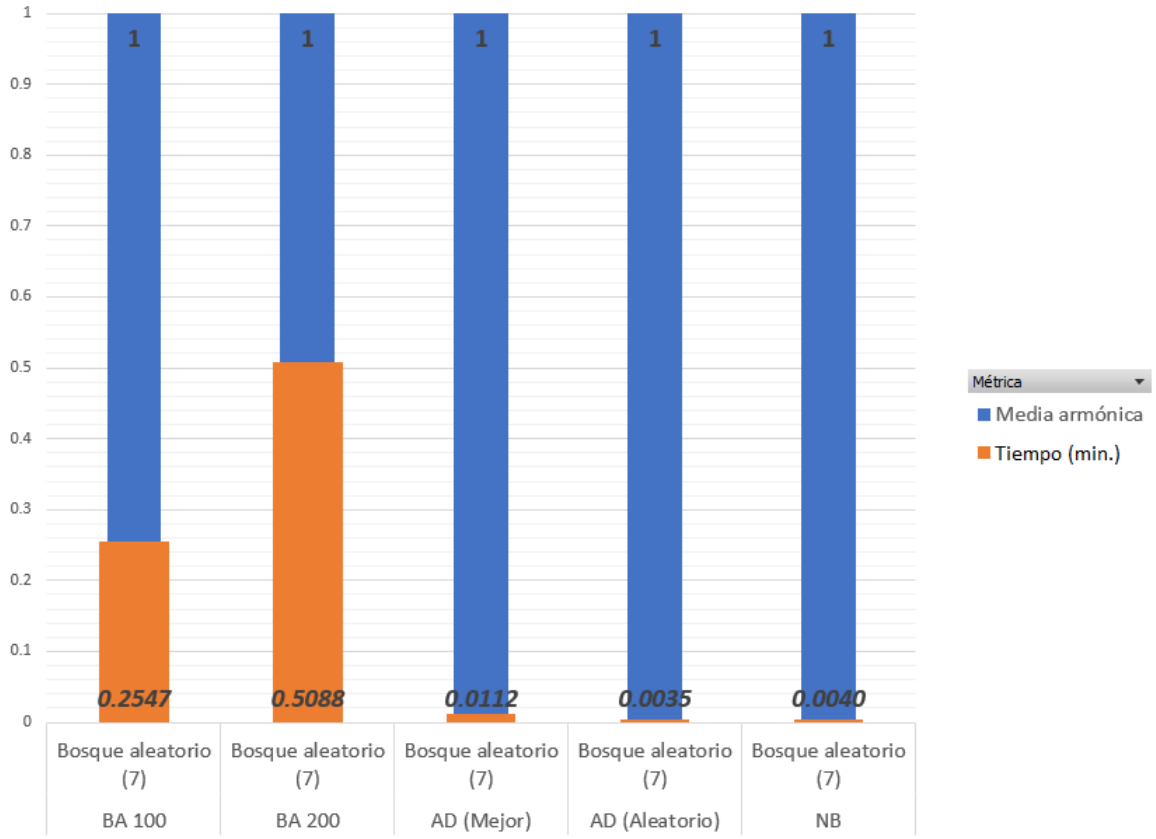

Figura 28. Resultados de las mejores combinaciones clasificador-conjunto durante la etapa de pruebas en el segundo escenario. Primera prueba. El número entre paréntesis corresponde al número de características contenido en cada conjunto. El primer renglón de algoritmos se refiere al conjunto, mientras que el segundo renglón de algoritmos se refiere al clasificador involucrado.

Los resultados obtenidos en la tabla anterior se repiten en la segunda prueba (Tabla 12): Las medias armónicas durante los entrenamientos y pruebas fue del 100 % y la desviación estándar para cada combinación tiene un valor de 0 %. Nuevamente la diferencia está en los tiempos de clasificación, en la cual se hacen presentes como mejores las combinaciones de AD(*aleatorio*)-BA y NB-BA.

Tabla 12. Tabla comparativa entre las mejores combinaciones durante las etapas de entrenamiento (μ_{HE}) y segunda prueba (H_{P2}) en el segundo escenario. Nótese que σ es la desviación estándar obtenida en la etapa de entrenamiento, mientras que T_{P2} se refiere al tiempo de clasificación en la segunda prueba.

Combinación	μ_{HE}	σ	H_{P2}	T_{P2} (min.)
BA100-BA	100 %	±0 %	100 %	0.254758
BA200-BA	100 %	±0 %	100 %	0.508867
AD(*mejor*)-BA	100 %	±0 %	100 %	0.011241
AD(*aleatorio*)-BA	100 %	±0 %	100 %	0.003536
NB-BA	100 %	±0 %	100 %	0.004066

Estas observaciones se confirman con la Figura 29, la cual plasma dicha información y

presenta resultados similares a los de la figura anterior. Sin embargo, al comparar los tiempos en la Tabla 13, se puede observar una mínima diferencia.

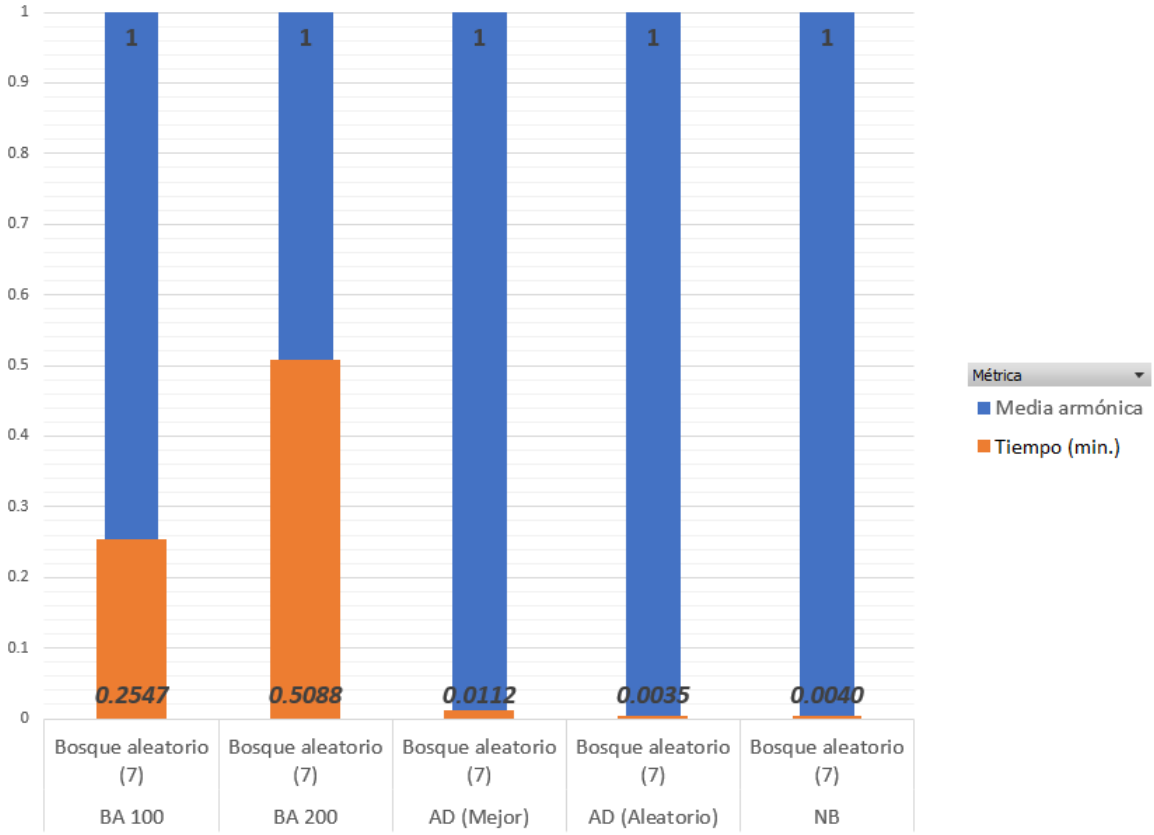

Figura 29. Resultados de las mejores combinaciones clasificador-conjunto durante la etapa de pruebas en el segundo escenario. Segunda prueba. El número entre paréntesis corresponde al número de características contenido en cada conjunto. El primer renglón de algoritmos se refiere al conjunto, mientras que el segundo renglón de algoritmos se refiere al clasificador involucrado.

Lo anterior demuestra que no es una tarea complicada determinar el tipo de operación que se realizará al usar la herramienta de arbitraje. De hecho, esto es posible al usar indicadores de volatilidad y tendencia, tal cual es el caso del conjunto de bosque aleatorio, al incluir algún indicador como *Bollinger Bands* para determinar si se comprará o venderá.

Tabla 13. Tabla comparativa de tiempos entre pruebas del segundo escenario. Los subíndices *P1* y *P2* corresponden a *prueba 1* y *prueba 2*, respectivamente.

Combinación	T_{P1} (min.)	T_{P2} (min.)	Diferencia
BA100-BA	0.254741	0.254758	0.000017
BA200-BA	0.508826	0.508867	0.000041
AD(*mejor*)-BA	0.011234	0.011241	0.000007
AD(*aleatorio*)-BA	0.003561	0.003536	0.000025
NB-BA	0.004047	0.004066	0.000019

En este escenario, debido al desempeño similar de las combinaciones es un tanto complicado determinar cuales son los peores y mejores conjuntos y clasificadores. Sin embargo, se aprovechó el factor tiempo para tomar una decisión, aunque al comparar entre ambas pruebas, la diferencia era mínima. Dicho esto, el mejor conjunto en general es bosque aleatorio ya que incluso el clasificador de NB que no es tan bueno, trabajó bien. En cuanto a peores y mejores clasificadores, ninguno fue realmente malo, incluso *NB* arrojó buenas combinaciones, pero el clasificador de árbol de decisión con partición aleatoria es superior para este escenario. En contexto, el problema parece demasiado sencillo, pero estos clasificadores entrenados sólo tienen sentido después de obtener resultados con el primer escenario.

6.3. Conclusiones parciales

En esta sección se presenta de manera puntual las conclusiones parciales que se adquirieron al plantearse el problema de identificación de tipo de oportunidad de operación con arbitraje. Nótese que estos resultados son validos únicamente tras obtener dichas oportunidades en el primer escenario.

- Determinar el tipo de operación (compra o venta) con el arbitraje triangular como estrategia es una tarea sencilla de acuerdo a los resultados presentados en cada una de las etapas (1 entrenamiento y 2 pruebas).
- Persiste la poca eficacia de los conjuntos dados por los métodos de filtro respecto al resto de los conjuntos, aunque su desempeño general no fue malo. Por otro lado, los conjuntos de *RFE, bosque aleatorio, lightGBM, regresión logística* y *TODAS* son los que mejor funcionan con los clasificadores propuestos, sin embargo, destaca el conjunto de bosque aleatorio por la baja cantidad de elementos en él.
- En cuanto a clasificadores, el más lento en ambas pruebas fue el de bosque aleatorio, mientras que el más rápido fue NB, pero destaca árbol de decisión en desempeño general.
- Los conjuntos generados tienden a contener indicadores de tipo volatilidad y tendencia, mientras que los indicadores menos frecuentes pertenecen a la categoría de osciladores, volumen y *otros*. Esto denota que para determinar el tipo de operación con la estrategia de arbitraje triangular, lo ideal es trabajar con la volatilidad y tendencia.
- Por último, dados los resultados y el análisis presentado en este capítulo, es posible concluir que la mejor combinación para el segundo escenario es AD(*aleatorio*)-BA.

Capítulo 7. Resultados y discusión del tercer escenario: oportunidades de operación con etiquetado multiclase

En este capítulo se presentan los resultados de la selección de características, entrenamientos y pruebas del tercer escenario. Asimismo se interpretan y discuten dichos resultados, qué tipo de características fueron tomadas, los mejores conjuntos, clasificadores y combinaciones.

7.1. Conjunto de características

Los conjuntos generados pueden consultarse en los Apéndices de este trabajo (Tabla 54), además de cada característica contenida y sus detalles. Como se muestra en la Fig. 30, en los conjuntos obtenidos para el tercer escenario se tiene una mayor presencia de indicadores osciladores, comparado con los escenarios anteriores. Sin embargo, los tipos dominantes continúan siendo los de volatilidad y tendencia. Esto quiere decir que para determinar una oportunidad y el tipo de operación que se llevará a cabo, toma mucha relevancia el momento, ímpetu o impulso que tiene el precio. Por otro lado, el volumen nuevamente tiene poca presencia. Esto puede que se deba a la misma volatilidad del mercado, pero no existen evidencias de esta afirmación. El número de características utilizadas incrementó considerablemente respecto a los escenarios anteriores, sin embargo, la proporción de elementos por tipo de seleccionador se respetó. Implicando que para llevar esta doble tarea se necesita más información.

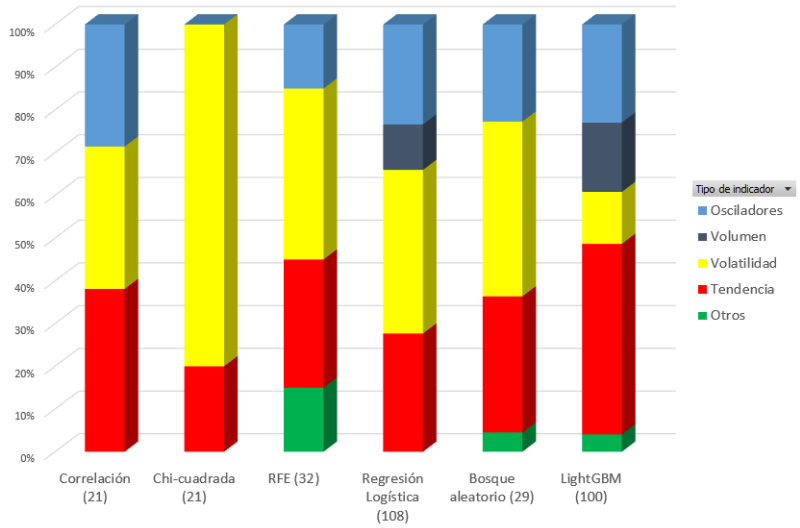

Figura 30. Conjuntos de características generados para el tercer escenario y el porcentaje de tipo de indicadores técnicos en cada uno de ellos. El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Las características que aparecen con más frecuencia son las que se presentan a continuación:

- Tendencia: *Parabolic SAR*.
- Osciladores: *RSI*.
- Volatilidad: *Bollinger Bands y Donchian Channel*.
- Volumen: *ADI* y *Ease of Movement*.
- Otros: *Daily Log Return*

Por el contrario, las características menos frecuentes o sin presencia en los conjuntos del tercer escenario son:

- Tendencia: *ADX*
- Osciladores: *Ultimate Oscillator*
- Volumen: *VWAP*

Gracias al gran número de características en regresión logística y lightGBM, hay indicadores que son usados una única vez en dichos conjuntos, como *Chaikin Money Flow, Negative Volume Index*, entre otros pertenecientes a la categoría de volumen, osciladores y *otros*.

7.2. Análisis con media armónica: Entrenamiento

Similarmente a los dos capítulos anteriores, se procederá a determinar las mejores combinaciones clasificador-conjunto de características, discutiendo los resultados partiendo de cada uno de los clasificadores con la finalidad de tomar la mejor combinación. Para la etapa de pruebas, se comparará únicamente los resultados entre las mejores combinaciones encontradas en esta sección. Cabe mencionar que en este escenario se cuenta con los mismos clasificadores que en el escenario anterior: bosque aleatorio, árbol de decisión y *Naive Bayes*.

El clasificador de bosque aleatorio con 100 árboles (Figura 31) obtuvo buenas calificaciones con los conjuntos de regresión logística, *TODAS* y bosque aleatorio. Particularmente, la mejor calificación fue dada por el primero con un valor de 0.9494, mientras que los conjuntos con calificaciones más bajas fueron X^2 y correlación con valor debajo de 0.82. En cuanto a la desviación estándar no hay una diferencia significativa. Ya que todos los valores oscilan entre 0.0034 a 0.0054. Sin embargo, el valor más bajo se lo lleva X^2, y el más alto fue dado

por el conjunto de bosque aleatorio. Por otro lado, el conjunto de regresión logística contiene 108/209 características, posicionándolo como el subconjunto con la cardinalidad más grande, lo que implica que es mejor utilizar dicho subconjunto en lugar del conjunto TODAS, dada la calificación de media armónica y su STD (que respecto a TODAS tiene una diferencia de 0.001). Así, la mejor combinación es *BA100 - regresión logística*. Cabe destacar además que, con respecto al primer escenario, los mejores conjuntos siguen siendo generados por los métodos de envoltura y embebidos.

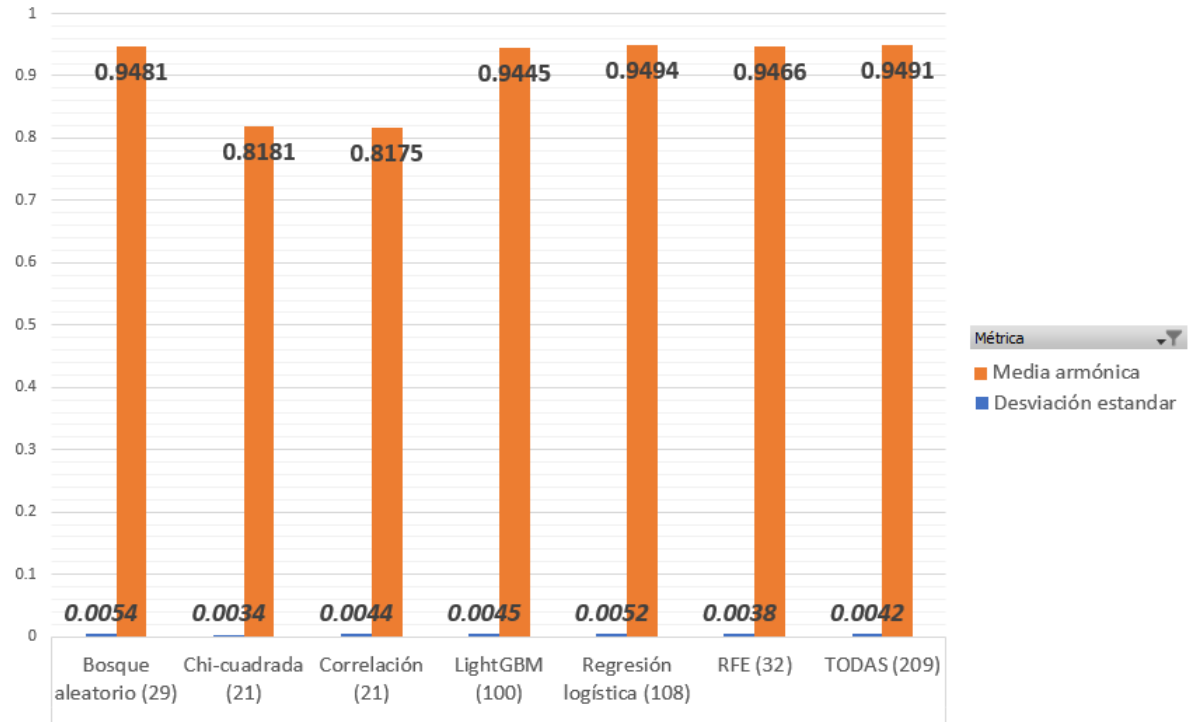

Figura 31. Resultados de clasificación en el tercer escenario con el clasificador bosque aleatorio (con 100 árboles en los nodos). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Similarmente, BA con 200 árboles obtuvo el resultado más alto con regresión logística con una puntuación de 0.9495, seguido por los conjuntos TODAS y regresión logística. Los conjuntos con menor calificación fueron aquellos generados por los métodos de filtrado, con un valor de 0.81. Observe que, el conjunto de regresión logística tiene el peor valor de STD respecto al resto de los conjuntos, sin embargo, dicho valor (0.005) no es lo suficientemente alto. En general, todos los conjuntos tienen un STD bajo, lo cual no afecta al desempeño del conjunto. En cuanto al número de características, la diferencia entre usar todas ellas (209) y usar el conjunto de regresión logística (108) es grande en cantidad, pero en calificación no es tanta (0.0002 de diferencia de media armónica), mientras que en STD difieren por 0.0009. Así, la mejor combinación es dada por *BA200 - regresión logística*. Nuevamente, los peores conjuntos son aquellos generados por los algoritmos de selección con método de filtrado.

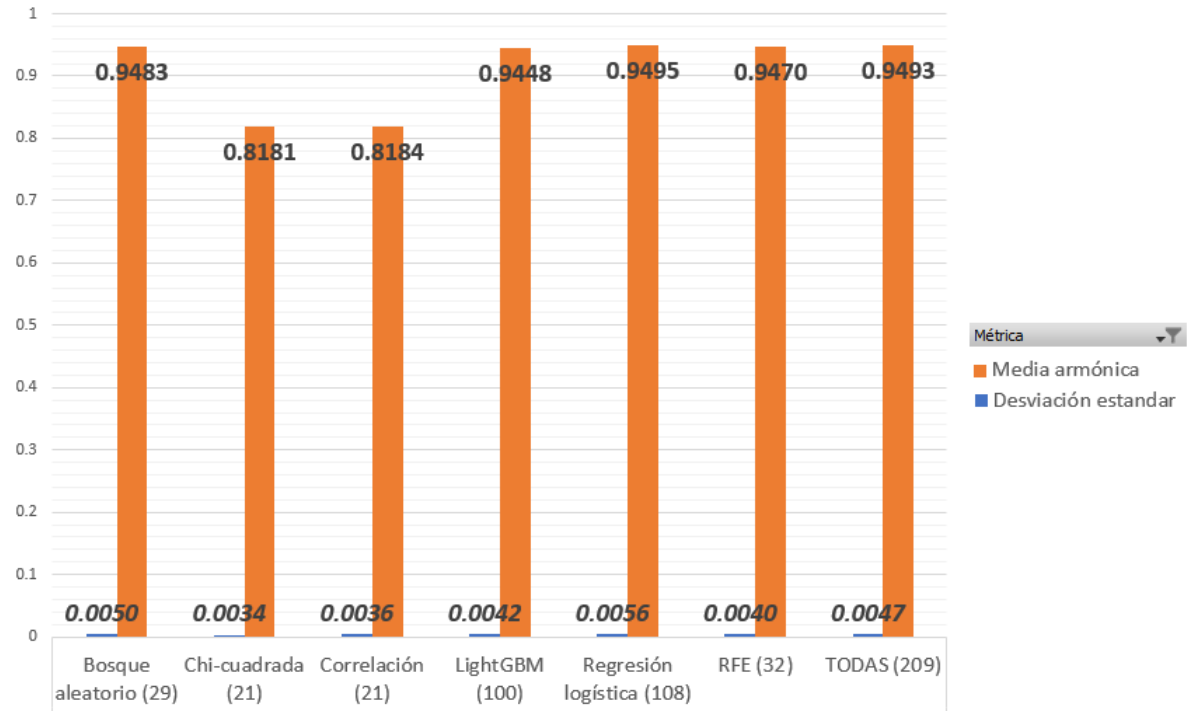

Figura 32. Resultados de clasificación en el tercer escenario con el clasificador bosque aleatorio (con 200 árboles en los nodos). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Al comparar ambos clasificadores de bosque aleatorio, se puede observar que la mejor calificación es de *BA200-regresión logística*, con una diferencia de 0.0001 sobre *BA100-regresión logística*, lo cual no es significativa. En general, mantienen calificaciones y desviaciones estándar similares. Por otro lado, el clasificador BA trabaja bien con conjuntos de cardinalidad alta, mientras que aquellos de cardinalidad baja adquieren un desempeño relativamente pobre respecto al resto de los conjuntos.

Por otro lado, el clasificador de árbol de decisión con la mejor partición (Fig. 33) obtuvo un buen desempeño con el conjunto de regresión logística (con un total de 0.9163), mientras que el desempeño más bajo fue dado por los conjuntos generados por medio de algoritmos de selección por filtro. Esto da la idea de que un conjunto con suficiente información puede tener un buen aprovechamiento ya que incluso el conjunto TODAS obtuvo un desempeño alto. Sin embargo, con los conjuntos de X^2 y correlación (que tienen menos características), el clasificador no tiene un desempeño bueno o media armónica superior a 0.9. En cuanto a la desviación estándar todos los valores son menores a 0.0046, exceptuando el conjunto de correlación que tiene un valor de 0.01, pero en general STD no afecta en gran medida al desempeño de los conjuntos gracias a que es bajo en promedio. Al comparar el conjunto TODAS con regresión logística, se puede apreciar una ligera mejoría por parte de este último. En cuanto a la desviación estándar, le gana también en estabilidad y la reducción de términos es significativa (poco menos de la mitad). Así, el mejor conjunto para este clasificador es el de regresión logística.

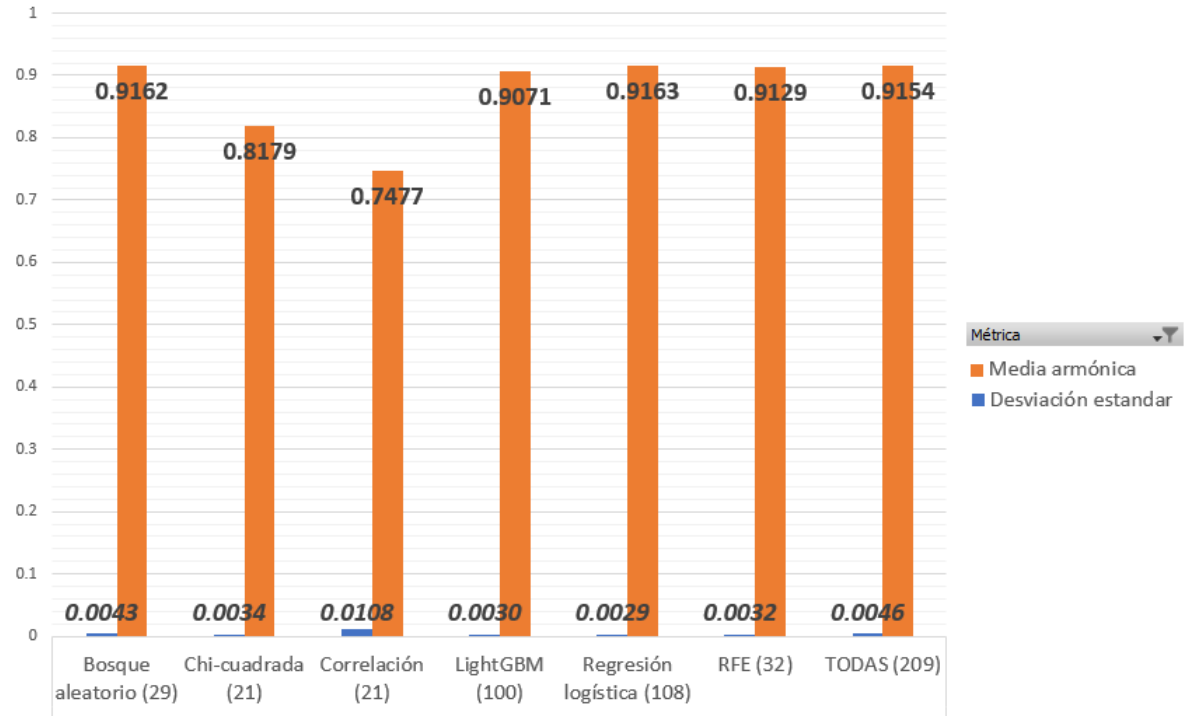

Figura 33. Resultados de clasificación en el tercer escenario con el clasificador AD (con la mejor partición). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Con respecto a su contra parte con partición aleatoria (Figura 34), se puede observar que regresión logística obtuvo la calificación más alta con 0.9186 en media armónica, seguido por TODAS y bosque aleatorio. Las calificaciones más bajas fueron de X^2 y correlación con 0.81. En si, el comportamiento fue idéntico al de la figura anterior, pero con una mejoría en los conjuntos buenos y empeoramiento en los conjuntos de desempeño bajo. Ahora bien, el STD incrementó para casi todos los conjuntos, excepto por el generado por bosque aleatorio y correlación. Sin embargo, no fue de manera exuberante, sino mínima (0.001 aproximadamente), lo cual se interpreta como un ±0.1 %. En general, la mejor combinación continúa siendo con el conjunto de regresión logística: la reducción de características respecto al conjunto general incrementa su desempeño en 0.002, lo cual no es demasiado, pero se obtienen con casi la mitad de las características, ahorrando una considerable cantidad de tiempo.

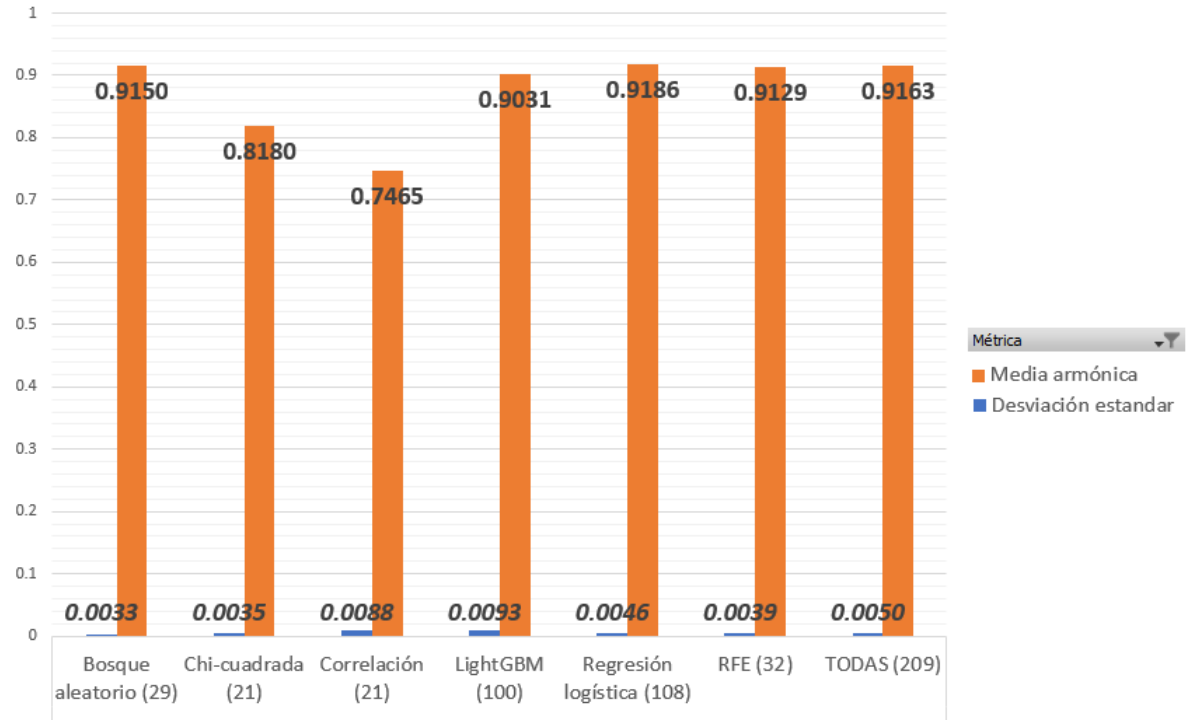

Figura 34. Resultados de clasificación en el tercer escenario con el clasificador AD (con partición aleatoria). El número entre paréntesis corresponde al número de características contenido en cada conjunto.

Al comparar los clasificadores de árbol de decisión, se puede observar que ambos trabajan bien con el conjunto de regresión logística. En general, las mejores calificaciones son dadas por los conjuntos generados por los métodos de envoltura y embebidos, mientras que las más bajas son dadas por los conjuntos generados por los métodos de filtro. Ahora bien, se puede notar un incremento en el STD de X^2, lightGBM, regresión logística, RFE y TODAS, en el clasificador con partición aleatoria. No obstante, decrementa con bosque aleatorio y correlación. Fuera de eso, al comparar los mejores conjuntos en ambos clasificadores, se observa una ligera mejoría por parte del clasificador con partición aleatoria en la media armónica por un total de 0.0023, por lo demás, se mantienen cambios mínimos en general.

En la Figura 35 se aprecia que el clasificador de *Naive Bayes* tiene una calificación de 0.8511 con el conjunto de bosque aleatorio (calificación más alta de media armónica), seguido por X^2 y correlación. Después de esos conjuntos, el desempeño general cae en un 50 %, esto debido a la cantidad y calidad de características que manejan los conjuntos. Aunado a este hecho se encuentra el funcionamiento particular de este clasificador, el cual compara la probabilidad de pertenencia a una clasificación dada la incidencia de una característica, sin importar el resto de ellas. En cuanto a la desviación estándar se puede apreciar que los conjuntos de bajo desempeño tienen un STD bajo (exceptuando el conjunto de regresión logística, lo cual podría deberse al tipo de características contenidas o al método con el que fue generado, ya que es el único obtenido por método de envoltura). Por otro lado, los conjuntos con

desempeño supuestamente alto tienen un STD elevado, con un valor aproximado de ±1 %. En particular, se puede apreciar como el conjunto TODAS fue el que tuvo peor desempeño junto con lightGBM, por lo que sin duda reducir grandemente la cantidad de características favoreció al desempeño del clasificador. De esta manera, es posible concluir que, para este clasificador, la mejor combinación es con bosque aleatorio.

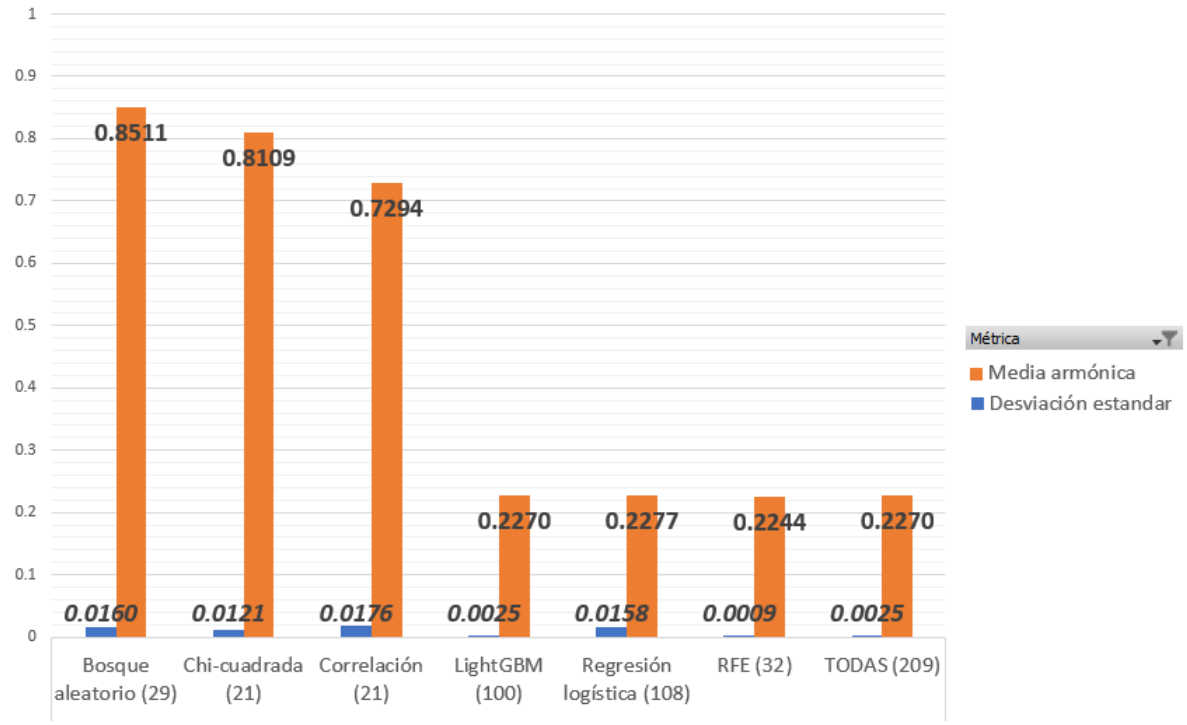

Figura 35. Resultados de clasificación en el tercer escenario con el clasificador *Naive Bayes*. El número entre paréntesis corresponde al número de características contenido en cada conjunto.

De acuerdo al análisis anterior, se tiene que las mejores combinaciones de conjuntos por cada clasificador son las siguientes: *AD(aleatorio)-regresión logística, AD(mejor)-regresión logística, BA100-regresión logística, BA200-regresión logística* y *N.Bayes-B.A*. Dicho esto, el mejor conjunto observado es el de regresión logística por trabajar mejor con casi todos los clasificadores (con la excepción de *N. Bayes*), mientras que los conjuntos menos óptimos para este escenario son X^2 y correlación por las bajas calificaciones en cada uno de los gráficos mostrados con anterioridad más la dispersión de los datos que se aprecia por la desviación estándar. Ahora bien, para determinar un mejor y peor clasificador, es necesario revisar las pruebas y su tiempo de ejecución durante dicha etapa.

7.2.1. Análisis con media armónica: Pruebas

En esta sección se consideran únicamente los mejores resultados de cada uno de los clasificadores de la etapa anterior (bosque aleatorio, árbol de decisión y NB) con su mejor conjunto. Para determinar la mejor combinación durante el tercer escenario se comparan los resultados arrojados en los entrenamientos con los de pruebas de cada una de las mejores combinaciones. Además, se toma en consideración el tiempo como factor para determinar la mejor combinación durante las pruebas.

En la Tabla 14 se puede observar que los valores de media armónica obtenidos durante los entrenamientos en todos los casos es superior a los valores de las pruebas. Además, las combinaciones tienen un excelente desempeño (en su mayoría una media armónica superior a 91 %). De hecho, la mejor combinación a partir de la media armónica es BA200-*regresión logística* con un total de 94.95 % en la primera fase y con 94.62 durante la segunda. En cuanto al tiempo, las combinaciones asociadas al clasificador BA no son óptimos por la definición dada de máximos y mínimos relativos en este trabajo. Así, se tiene que las únicas combinaciones factibles son AD(*mejor*)-*regresión logística*, AD(*aleatorio*)-*regresión logística* y NB-BA. Dicho esto, el mejor tiempo lo tiene NB-BA, pero al tomar en cuenta la media armónica, sin duda es mejor AD(*aleatorio*)-*regresión logística*. De hecho, esta última combinación tiene el valor más bajo de STD, por lo cual parece ser el clasificador-conjunto indicado para el objetivo de este escenario.

Tabla 14. Tabla comparativa entre las mejores combinaciones durante las etapas de entrenamiento (μ_{HE}) y prueba (H_P) en el tercer escenario en términos de media armónica. Nótese que σ es la desviación estándar obtenida en la etapa de entrenamiento, mientras que T_P se refiere al tiempo de clasificación en pruebas.

Combinación	μ_{HE}	σ	H_P	T_P (min.)
BA100-*regresión logística*	94.94 %	±0.52 %	94.59 %	16.63
BA200-*regresión logística*	94.95 %	±0.56 %	94.62 %	33.25
AD(*mejor*)-*regresión logística*	91.63 %	±0.29 %	91.17 %	3.33
AD(*aleatorio*)-*regresión logística*	91.86 %	±0.0046 %	91.25 %	0.23
NB-BA	85.11 %	±1.6 %	85.08 %	0.02

En la Figura 36 se observa de manera más clara que el clasificador de BA tiene una alta calificación, pero un tiempo de ejecución que no favorece al problema. Por otro lado, el clasificador NB adquirió la calificación más baja de todas, dejando al clasificador AD como el más indicado. Al comparar los tiempos de las combinaciones de este último clasificador, es fácil determinar cual es preferible por la diferencia de tiempo de ejecución (3 segundos aproximadamente de diferencia).

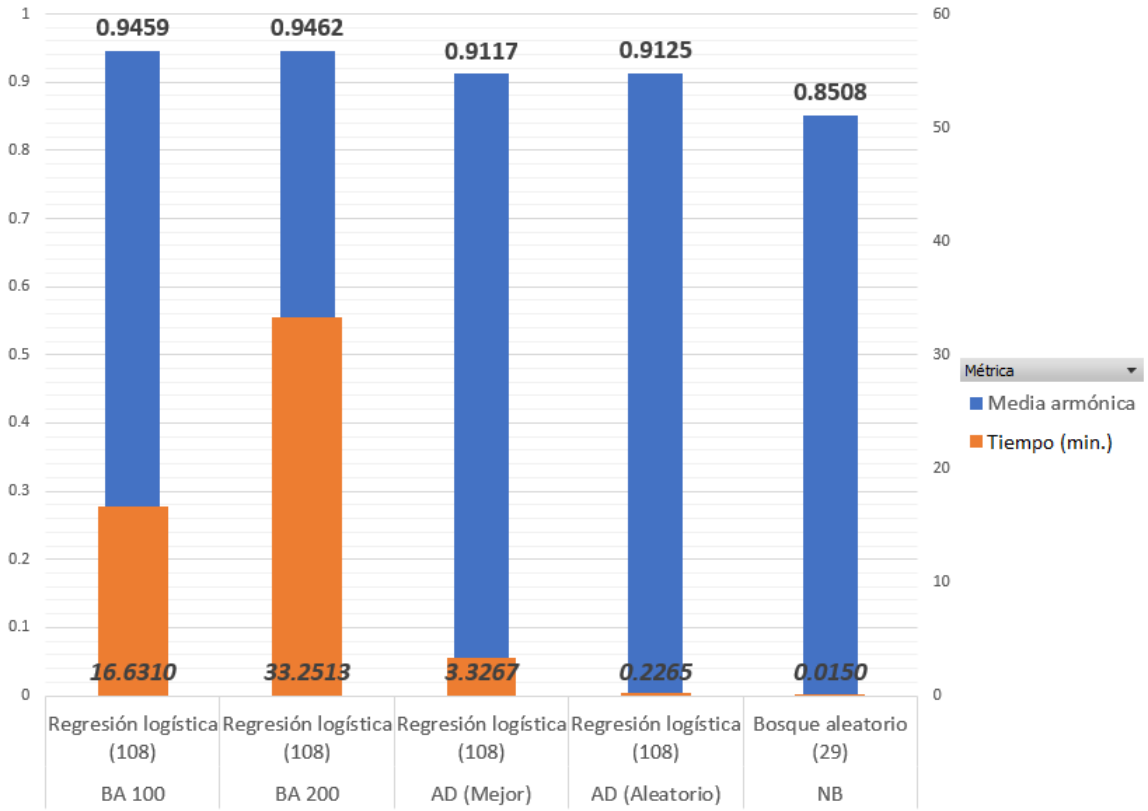

Figura 36. Resultados de las mejores combinaciones clasificador-conjunto durante la etapa de prueba en el tercer escenario. El número entre paréntesis corresponde al número de características contenido en cada conjunto. El primer renglón de algoritmos se refiere al conjunto, mientras que el segundo renglón de algoritmos se refiere al clasificador involucrado.

Así, se concluye que la mejor combinación es la de AD(*aleatorio*)-*regresión logística*. De hecho, solamente se acentuó la ineficiencia en tiempo del clasificador de bosque aleatorio, ya que en calificación arrojó los mejores resultados. Por otro lado, el clasificador de árbol de decisión obtuvo resultados buenos en un tiempo excelente. En cuanto al clasificador de NB, fue un buen punto de referencia. Si consideramos un mejor conjunto para este escenario, dicho conjunto es regresión logística por su prevalencia, a diferencia del primer y segundo escenario donde el conjunto de bosque aleatorio fue el mejor. En general, este escenario, donde se combinan las tareas de los dos escenarios anteriores (detección de oportunidades y determinación de tipo de operación en oportunidades de arbitraje) obtuvo excelentes resultados.

7.3. Conclusiones parciales

En esta sección se presentan las conclusiones parciales de manera puntual del problema planteado en el tercer escenario que consta de identificación de oportunidades de operación de arbitraje y el tipo de operación a realizar en dichas oportunidades. Nótese que este problema

es una combinación de los dos escenarios anteriores.

- Los seleccionadores de filtro arrojan los conjuntos menos convenientes para el problema planteado en este proyecto (exceptuando el caso en el que se usa el clasificador de NB). Además, los conjuntos generados por métodos de envoltura y embebidos son los que presentan un desempeño sobresaliente general en cuanto a valores de media armónica se refiere.
- El clasificador más lento es bosque aleatorio, mientras que los más veloces son los de NB y árbol de decisión. Ahora bien, entre todos ellos destaca AD con partición aleatoria por su velocidad y calificaciones, ya que NB obtuvo valores bajos de media armónica.
- Las mejores calificaciones son dadas por las combinaciones del clasificador de bosque aleatorio. Sin embargo, este clasificador es demasiado lento para el problema propuesto (como se mencionó en el punto anterior), por lo cual, no es posible operar en el mercado de manera efectiva con la definición de máximos y mínimos relativos dado en este trabajo.
- Los conjuntos, como en los escenarios anteriores, adquirieron más indicadores de volatilidad y tendencia, mientras que los pertenecientes a *volumen* y *otros* son pocos. Esto confirma la eficacia de esta clase de indicadores técnicos para identificar operaciones y el tipo de operación de arbitraje triangular.
- Durante el tercer escenario, el conjunto de *regresión logística* es el que tiene un mejor desempeño con cualquier tipo clasificador propuesto. Así mismo, la mejor combinación es AD(*aleatorio*)-*regresión logística*, seguido por AD(*mejor*)-*regresión logística*, por el buen desempeño general que tuvieron.

Capítulo 8. Epílogo

8.1. Resumen

El arbitraje financiero es una herramienta en el mercado de divisas que a través del tiempo ha perdido el interés a los ojos de inversionistas y a su vez etiquetada como poco eficiente por la existencia de oportunidades demasiado escasas. Sin embargo, se encontraron ventanas de oportunidad en un par de divisas poco convencional (NZDUSD), respecto a otros pares más estudiados como lo es EURUSD, por decir un ejemplo. Dichas ventanas de oportunidad a la compra constan del 0.5 % y a la venta del 0.8 %. Aunque estos números reflejan bajas oportunidades a simple vista, en relación a la base de datos pueden ser suficientes por día. Además, estas oportunidades cuentan con un margen de tiempo lo suficientemente grande para analizar, tomar decisiones y operar (aproximadamente 5 minutos). Ahora bien, un factor importante de esta investigación es que las oportunidades de arbitraje triangular fueron estudiadas en el mercado en periodos de minutos, cuando lo usual es estudiarlas a nivel de hora y día.

Al visualizar como un ’punto’ cada periodo y sus valores en relación al precio en la base de datos, las oportunidades de arbitraje triangular se etiquetaron respecto al resto de ’puntos’ del mercado y un umbral que serviría como margen de disminución de pérdidas al operar en el mercado. Se consideraron tres escenarios, por lo que en cada uno de ellos se manejó un etiquetado distinto.

En el primer escenario se buscó identificar las oportunidades respecto al resto de los puntos de la base de datos sin importar si era una compra o una venta, por lo cual se usaron dos etiquetas (oportunidad o no oportunidad). En este escenario se obtuvieron seis conjuntos de características obtenidos tras usar algoritmos seleccionadores, dichos conjuntos fueron nombrados de acuerdo al algoritmo utilizado: correlación, chi-cuadrada, RFE, regresión logística, bosque aleatorio y lightGBM. Posteriormente se validó la eficacia de dichos conjuntos por medio de la técnica de validación cruzada en combinación de distintos algoritmos de clasificación: Máquina de soporte vectorial o SVM con kernel lineal, bosque aleatorio o BA (generando 100 y 200 árboles), perceptrón multicapa o MLP (con 50 y 100 capas ocultas), árbol de decisión o AD (con partición aleatoria y la mejor) y *Naive Bayes*.

Los resultados obtenidos del primer escenario mostraron que los seleccionadores de filtro arrojan los conjuntos menos convenientes, ya que las calificaciones de dichos conjuntos con los clasificadores fueron bajas. Sin embargo, los conjuntos generados por métodos embebidos

fueron realmente buenos. En particular, los conjuntos de RFE y BA tienen un desempeño destacable con cada clasificador. Durante las pruebas los clasificadores más lentos fueron MLP y BA, mientras que los más veloces fueron SVM, NB y AD. Se observó también que MLP, SVM y NB no tienen un desempeño bueno respecto al resto. Dado que los resultados de los clasificadores MLP y SVM durante el primer escenario mostraron un desempeño pobre en relación con la media armónica, desviación estándar y tiempo, éstos fueron descartados para escenarios posteriores. Al analizar los conjuntos, dominan los indicadores de tipo volatilidad y tendencia. Por otro lado, los menos frecuentes son de tipo volumen y *otros*. Esto da a entender que para la identificación de oportunidades de arbitraje en el mercado, los indicadores de volumen y *otros* tienen poco efecto o influencia, seguido por los osciladores, contrario a los de volatilidad y tendencia, que juegan un papel importante para dicha estrategia. Finalmente, los resultados de entrenamiento y pruebas indicaron que la mejor combinación para identificar oportunidades de arbitraje triangular en el mercado de divisas fue el clasificador AD con partición aleatoria con el conjunto de características de BA generado para el primer escenario.

El segundo escenario funcionó como una continuación del primero, tras identificar dichas oportunidades de inversión. Se buscó definir si las oportunidades correspondían a una compra o una venta, por lo que se usaron dos etiquetas (compra o venta). En dicho escenario se volvieron a generar los conjuntos de características, pero a partir del mejor conjunto del escenario anterior (BA). Se encontró que determinar el tipo de operación es una tarea sencilla, al obtener resultados excelentes validados incluso mediante pruebas con dos conjuntos independientes. Los conjuntos generados por métodos de filtro fueron poco eficientes respecto a los demás pese a que su desempeño general no fue malo. Los conjuntos restantes (*RFE, bosque aleatorio, lightGBM, regresión logística* y *TODAS*) tuvieron un desempeño excelente con cada clasificador propuesto en tiempo y media armónica, por lo que se definió al mejor por la cardinalidad de los conjuntos. Lo anterior estableció al conjunto BA como el más apropiado para la tarea planteada en este escenario por su desempeño y baja cantidad de elementos. En cuanto a clasificadores, el más lento en ambas pruebas fue BA, mientras que el más rápido fue NB, pero AD lo supera por realizar la tarea en un tiempo considerablemente bajo y tener un desempeño general mejor. Ahora bien, al analizar los conjuntos generados durante este escenario, resulta claro que para determinar el tipo de operación, los indicadores de volatilidad y tendencia aquieren un rol más importante. Esta última afirmación es un tanto obvia ya que los indicadores de tendencia por su funcionamiento y naturaleza revelan el tipo de operación. Por otro lado, los de categoría oscilatorios, volumen y *otros* no muestran mucha presencia en los conjuntos, quizá sí de manera porcentual, pero al observar su cardinalidad se revela que estos contienen a lo mucho 1 elemento en general. Por último, al analizar los resultados del entrenamiento y las pruebas se concluye que la mejor combinación es dada por el clasificador AD (con partición aleatorio) y el conjunto de características BA.

El tercer escenario es una combinación de los dos escenarios previos, en el que se buscó identificar la oportunidad e inmediatamente definir si era compra o venta, por ello se manejaron tres etiquetas en este escenario: no oportunidad, oportunidad de compra y oportunidad de venta. Los conjuntos generados por seleccionadores de filtro no tuvieron un desempeño bueno para el problema planteado en este escenario. Mientras que los conjuntos generados por métodos de envoltura y embebidos fueron sobresalientes en cuanto a media armónica. Por otro lado, en cuanto a velocidad para clasificar, el mejor fue NB, seguido por AD y al último BA. En cuanto a desempeño de clasificación, el mejor fue BA, luego AD, y por último NB. Esto nos deja con AD con partición aleatoria como el mejor clasificador, ya que su tiempo de clasificación a pesar de no ser el del puesto número uno, fue excelente y sus calificaciones son mejores que las de NB. El clasificador de BA tiene buenas calificaciones, pero para el problema planteado resulta muy lento por la manera en que se definieron los máximos y mínimos relativos. Recuerde que se tiene una restricción de cinco minutos, lo cual no lo hace buen candidato. Este escenario confirmó que los indicadores de volatilidad y tendencia son esenciales para la detección de oportunidades en arbitraje triangular, de acuerdo a la cantidad de estos indicadores contenida en los conjuntos generados. Ahora bien, los indicadores osciladores no tienen una presencia fuerte, pero son más numerosos que los de volumen y *otros*. Finalmente, el conjunto de *regresión logística* destacó sobre el de *BA* a diferencia de los escenario anteriores, que junto con el clasificador AD con partición aleatoria los posiciona como la mejor combinación. Si se quisiera incrementar el desempeño sacrificando un poco de tiempo (estando aún dentro de los cinco minutos), se podría optar por la combinación de AD con la mejor partición y el conjunto de características de regresión logística.

8.2. Conclusión principal

Identificar el tipo de operación a realizar con arbitraje triangular no es una tarea complicada, el detalle está en predecir dichas oportunidades. Sin embargo, existen suficientes oportunidades durante el día que, de acuerdo a la teoría y una buena identificación de las mismas, el riesgo que se corre al operar con arbitraje es bajo. Esto podría estar influenciado por usar un par de divisas poco convencional y haber trabajado con periodos de un minuto en la base de datos. Los tres escenarios anteriores presentaron resultados positivos, sin embargo, entre los primeros dos y el tercer escenario, lo más efectivo es clasificar con múltiple etiqueta. Con ello se incrementa la cardinalidad del conjunto de características, pero está dentro del margen de tiempo deseado y cuenta con un mejor desempeño. En cuanto al tipo de característica, en todos los escenarios sobresalen los indicadores de volatilidad y tendencia. Los indicadores oscilatorios no fueron descartados del todo, seleccionándose algunos. Sin embargo, los que

mostraron baja o nula presencia son los de volumen y *otros*. Ahora bien, en cuanto a los seleccionadores de características, los conjuntos generados con métodos de filtro no destacaron, sobresalieron los que fueron obtenidos por métodos embebidos y por envoltura, en particular, RFE, BA y regresión logística. Por otro lado, al considerar únicamente la media armónica, el mejor clasificador es BA, pero es sumamente lento para el problema planteado. En velocidad, se tiene como mejor clasificador a NB, pero su resultados fueron pésimos. Esto llevó a destacar como mejor clasificador al AD, en particular si tiene como parámetro partición aleatoria en los nodos. Finalmente, se concluye que la mejor combinación es *AD (aleatorio)-regresión logística* implementando un múltiple etiquetado.

8.3. Trabajo futuro

Los aspectos a considerar para continuar este trabajo es un cuarto escenario donde se consideren cinco etiquetas: puntos dentro del umbral, sobre el umbral (excluyendo máximos relativos), debajo del umbral (excluyendo mínimos relativos), máximos relativos y mínimos relativos. Lo anterior debido a que los puntos debajo y sobre el umbral que excluyen máximos y mínimos relativos pueden generar ruido para la identificación de las mejores oportunidades de operación arbitraje triangular. Ahora bien, los indicadores usados como características en realidad son pocos en relación a los indicadores existentes ya que continuamente se están desarrollando nuevos, por lo cual se podrían considerar otros menos populares o convencionales. Incluso, se debe considerar el hecho de que se usaron cantidades dispares de tipo de características. Otro aspecto sumamente importante, es que estos resultados son efectivos únicamente en el par NZDUSD, por lo que se podría generalizar el método y ampliar la investigación con otros pares de divisas, idealmente, pares no convencionales o menos estudiados. Finalmente, el avance continuo de la ciencia invita a probar otros algoritmos de selección, clasificación e incluso otro método para validar el desempeño de las combinaciones.

Literatura citada

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., y Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, **8**: 14.

Agarwal, R. (2019). The 5 feature selection algorithms every data scientist should know. Recuperado el 20-02-2020 de: https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2.

Aguilar-Rivera, R., Valenzuela-Rendón, M., y Rodríguez-Ortiz, J. (2015). Genetic algorithms and darwinian approaches in financial applications: A survey. *Expert Systems with Applications*, **42**(21): 7684–7697.

Aiba, Y. y Hatano, N. (2006). A microscopic model of triangular arbitrage. *Physica A-statistical Mechanics and Its Applications*, **371**(2): 572–584.

Aiba, Y., Hatano, N., Takayasu, H., Marumo, K., y Shimizu, T. (2002). Triangular arbitrage as an interaction among foreign exchange rates. *Physica A-statistical Mechanics and Its Applications*, **310**(3): 467–479.

Akram, Q. F., Rime, D., y Sarno, L. (2008). Arbitrage in the foreign exchange market: Turning on the microscope. *Journal of International Economics*, **76**(2): 237–253.

Alexander, C. (2008). *Market risk analysis, pricing, hedging and trading financial instruments*, Vol. 3. John Wiley & Sons.

Ali, J., Khan, R., Ahmad, N., y Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, **9**(5): 272.

Baillie, R. T. y Bollerslev, T. (1991). Intra-day and inter-market volatility in foreign exchange rates. *The Review of Economic Studies*, **58**(3): 565–585.

Biesiada, J. y Duch, W. (2007). Feature selection for high-dimensional data—a pearson redundancy based filter. En: *Computer recognition systems 2*. Springer, pp. 242–249.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, **44**(1): 108–132.

Cai, M.-c. y Deng, X. (2008). Complexity of exchange markets. En: *Handbook on Information Technology in Finance*. Springer, pp. 689–705.

Cui, Z. y Taylor, S. (2020). Arbitrage detection using max plus product iteration on foreign exchange rate graphs. *Finance Research Letters*, **35**: 101279.

Dash, M. y NS, A. K. (2013). Exchange rate dynamics and forex hedging strategies. *Investment Management and Financial Innovations*, **10**(8): 8–16.

Dávila, A. V. y Herrera, G. F. (2015). Estrategia para invertir en el mercado de divisas (forex) basada en redes neuronales. *Revista Politécnica*, **35**(2): 124.

Davis, J. y Goadrich, M. (2006). The relationship between precision-recall and roc curves. En: *Proceedings of the 23rd international conference on Machine learning*. pp. 233–240.

de Vazelhes, W., Carey, C., Tang, Y., Vauquier, N., y Bellet, A. (2019). metric-learn: Metric learning algorithms in python. *arXiv preprint arXiv:1908.04710*.

Dembczyński, K., Waegeman, W., Cheng, W., y Hüllermeier, E. (2010). Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. En: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 280–295.

Dow, J. (1998). Arbitrage, hedging, and financial innovation. *Review of Financial Studies*, **11**(4): 739–755.

Edwards, G. (2018). Machine learning | an introduction. Recuperado el 28-02-2020 de: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0].

Goodhart, C., Love, R., Payne, R., y Rime, D. (2002). Analysis of spreads in the dollar/euro and deutschemark/dollar foreign exchange markets. *Economic Policy*, **17**(35): 535–552.

Grimaldi, M., Kokaram, A., *et al.* (2006). Discrete wavelet packet transform and ensembles of lazy and eager learners for music genre classification. *Multimedia Systems*, **11**(5): 422–437.

Hackeling, G. (2014). *Mastering Machine Learning with scikit-learn*, Vol. 1. Packt Publishing Ltd, primera edición. 35 Livery Street, Birmingham.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Tesis de doctorado, Department of Computer Science, Waikato University, New Zealand.

Hsu, Y.-C., Chen, A.-P., y Chang, J.-H. (2011). An inter-market arbitrage trading system based on extended classifier systems. *Expert Systems With Applications*, **38**(4): 3784–3792.

Huck, N. y Afawubo, K. (2015). Pairs trading and selection methods: is cointegration superior? *Applied Economics*, **47**(6): 599–613.

Ito, T., Yamada, K., Takayasu, M., y Takayasu, H. (2012). Free lunch! arbitrage opportunities in the foreign exchange markets. Reporte técnico, National Bureau of Economic Research.

Jeni, L. A., Cohn, J. F., y De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. En: *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, pp. 245–251.

Kaufman, P. (2011). *Alpha Trading: Profitable Strategies That Remove Directional Risk*, Vol. 1. John Wiley Sons Inc, primera edición. 222 Rosewood Drive, Danvers.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., y Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. En: *Advances in neural information processing systems*. pp. 3146–3154.

Kočenda, E. y Moravcová, M. (2019). Exchange rate comovements, hedging and volatility spillovers on new eu forex markets. *Journal of International Financial Markets, Institutions and Money*, **58**: 42–64.

Kramer, O. (2016). Scikit-learn. En: *Machine learning for evolution strategies*. Springer, pp. 45–53.

Krishnan, R. y Menon, S. S. (2009). Impact of currency pairs, time frames and technical indicators on trading profit in forex spot market. *International journal of Business insights & Transformation*, **2**(2).

Lai, T. L. y Xing, H. (2008). *Statistical Models and Methods for Financial Markets*, Vol. 1. Springer, New York, NY.

Lopez, D. (2017). Technical analysis (ta) library. Documentación de librería de python. [Recuperado el 25-03-2020 de: https://github.com/bukosabino/ta.

Mahmoodzadeh, S. y Gençay, R. (2014). Tick size change in the wholesale foreign exchange market. Reporte técnico, Working paper, Simon Fraser University, BC.

Mavrides, M. (1992). *Triangular Arbitrage in the Foreign Exchange Market: Inefficiencies, Technology, and Investment Opportunities*, Vol. 1. Praeger, primera edición. River Forest, Illinois.

Miner, R. C. (2008). *High probability trading strategies: Entry to exit tactics for the forex, futures, and stock markets*, Vol. 328. John Wiley & Sons.

Moosa, I. A. (2003). *International financial operations*, Vol. 1. Springer, New York NY, primera edición. San Diego, CA.

Mukherjee, S. y Sharma, N. (2012). Intrusion detection using naive bayes classifier with feature reduction. *Procedia Technology*, **4**: 119–128.

Nuti, G., Mirghaemi, M., Treleaven, P., y Yingsaeree, C. (2011). Algorithmic trading. *Computer*, **44**(11): 61–69.

Olivera, J. H. (1991). Equilibrio social, equilibrio de mercado e inflación estructural. *Desarrollo Económico*, pp. 487–493.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, **26**(1): 217–222.

Popovic, S. y Durovic, A. (2014). Intraweek and intraday trade anomalies: evidence from forex market. *Applied Economics*, **46**(32): 3968–3979.

Raul Garreta, G. M. (2013). *Learning scikit-learn: machine learning in python*, Vol. 1. Packt Publishing Ltd, segunda edición. 35 Livery Street, Birmingham.

Rish, I. (2001). An empirical study of the naive bayes classifier. En: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3, pp. 41–46.

Shieh, M.-D. y Yang, C.-C. (2008). Multiclass svm-rfe for product form feature selection. *Expert Systems with Applications*, **35**(1-2): 531–541.

Suykens, J. A. y Vandewalle, J. (1999). Training multilayer perceptron classifiers based on a modified support vector method. *IEEE transactions on Neural Networks*, **10**(4): 907–911.

Tsang, E., Markose, S., y Er, H. (2005). Chance discovery in stock index option and futures arbitrage. *New Mathematics and Natural Computation*, **1**(3): 435–447.

Uribe, G. y Álvarez, J. (1978). *Media geométrica y media armónica*, Vol. 6 de *1*. Servicio Nacional de Aprendizaje (SENA).

Vaidya, J. y Clifton, C. (2004). Privacy preserving naive bayes classifier for vertically partitioned data. En: *Proceedings of the 2004 SIAM international conference on data mining*. SIAM, pp. 522–526.

Vajda, V. (2014). Could a trader using only “old” technical indicator be successful at the forex market? *Procedia Economics and Finance*, **15**: 318–325.

Verbeke, M., Van Asch, V., Daelemans, W., y De Raedt, L. (2014). Lazy and eager relational learning using graph-kernels. En: *International Conference on Statistical Language and Speech Processing*. Springer, pp. 171–184.

Vidyamurthy, G. (2004). *Pairs Trading: Quantitative Methods and Analysis*, Vol. 1 de *1*. John Wiley Sons, Inc., primera edición. 111, River Street, Hoboken, NJ.

Wang, F., Li, Y., Liang, L., y Li, K. (2008). Triangular arbitrage in foreign exchange rate forecasting markets. En: *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*. pp. 2365–2371.

Wen, H., Ciamarra, M. P., y Cheong, S. A. (2018). How one might miss early warning signals of critical transitions in time series data: A systematic study of two major currency pairs. *PloS one*, **13**(3): e0191439.

Wissner-Gross, A. D. y Freer, C. E. (2010). Relativistic statistical arbitrage. *Physical Review E*, **82**(5): 56104.

Yazdi, S. H. M. y Lashkari, Z. H. (2013). Technical analysis of forex by macd indicator. *International Journal of Humanities and Management Sciences (IJHMS)*, **1**(2): 159–165.

Yiu, T. (2020). Understanding cross validation. Recuperado el 28-02-2020 de: https://towardsdatascience.com/understanding-cross-validation-419dbd47e9bd.

Yong, Y. L., Ngo, D. C., y Lee, Y. (2015). Technical indicators for forex forecasting: a preliminary study. En: *International Conference in Swarm Intelligence*. Springer, pp. 87–97.

Zhao, C., Gao, Y., He, J., y Lian, J. (2012). Recognition of driving postures by multiwavelet transform and multilayer perceptron classifier. *Engineering Applications of Artificial Intelligence*, **25**(8): 1677–1686.

Apéndice I: Aprendizaje máquina

En este apartado se encuentra la teoría utilizada de aprendizaje máquina a lo largo del trabajo.

El aprendizaje máquina es una herramienta para convertir la información en conocimiento (Raul Garreta, 2013). En los últimos 50 años, ha habido una explosión de datos. Esta gran cantidad de datos es inútil a menos que la analicemos y encontremos los patrones ocultos en su interior. Las técnicas de aprendizaje máquina se utilizan para encontrar automáticamente patrones dentro de datos complejos que de otro modo tendríamos dificultades para descubrir. Dichos patrones ocultos y el conocimiento sobre un problema se pueden usar para predecir eventos futuros y realizar todo tipo de toma de decisiones complejas.

La mayoría desconoce que ya interactuamos con *aprendizaje máquina* todos los días (Edwards, 2018). Cada vez que buscamos en Google algo, escuchamos una canción o incluso tomamos una foto, el *Aprendizaje Máquina* o *Machine Learning*, se está convirtiendo en parte del motor detrás de él, constantemente aprendiendo y mejorando de cada interacción. También está detrás de los avances que cambian el mundo, como la detección del cáncer, la creación de nuevos medicamentos y automóviles autónomos.

Para aprender las reglas que rigen un fenómeno, las máquinas tienen que pasar por un proceso de aprendizaje, probar diferentes reglas y aprender qué tan bien funcionan (Yiu, 2020). De ahí el nombre de *Aprendizaje Máquina*.

Existen múltiples formas de aprendizaje automático (Edwards, 2018); aprendizaje supervisado, no supervisado, semi-supervisado y de refuerzo. En este trabajo se hace uso de aprendizaje supervisado, esta técnica ayuda a deducir una función a partir de datos de entrenamiento. Cada forma de aprendizaje máquina tiene enfoques diferentes, pero todos siguen el mismo proceso y teoría.

Terminología

- Conjunto de datos: conjunto de ejemplos de datos que contienen características importantes para resolver el problema (Raul Garreta, 2013).
- Características: Datos importantes que nos ayudan a comprender un problema. Estos se introducen en un algoritmo de *machine learning* para ayudarlo a aprender (Raul Garreta, 2013).
- Modelo: La representación (modelo interno) de un fenómeno que un algoritmo de *machine learning* ha aprendido (Raul Garreta, 2013). Aprende esto de los datos que se muestran durante el entrenamiento. El modelo es el resultado que se obtiene después de

entrenar un algoritmo. Por ejemplo, un algoritmo de árbol de decisión sería entrenado y produciría un modelo de árbol de decisión.

Proceso

Los algoritmos de aprendizaje automático son como un bucle finito, ya que se detienen cuando el objetivo final, que depende del tipo de algoritmos, se cumple (Raul Garreta, 2013; Kramer, 2016; Edwards, 2018; Hackeling, 2014). Dichos algoritmos llevan a cabo un procedimiento como este:

1. Recopilación de datos: Obtener los datos de los que aprenderá el algoritmo.
2. Preparación de datos: Procesar los datos para obtenerlos en el formato óptimo, extrayendo características importantes y reduciendo la dimensionalidad.
3. Entrenamiento: También conocida como la etapa de ajuste, aquí es donde el algoritmo de aprendizaje máquina realmente aprende mostrándole los datos que se han recopilado y preparado.
4. Evaluación: Prueba el modelo para ver qué tan bien funciona.
5. Ajuste: Se adecua el modelo para maximizar su rendimiento.

Selección de características

Existen tres motivos esenciales por los cuales no le damos todas las características a un algoritmo, por lo cual debemos pasar previamente por un proceso de selección (Raul Garreta, 2013; Agarwal, 2019).

- El sobreajuste. A medida que aumenta la dimensionalidad del espacio de características, las configuraciones de números pueden crecer exponencialmente y, por lo tanto, el número de configuraciones cubiertas por una observación disminuye.
- La navaja de Ockham. Queremos que nuestros modelos sean simples y explicables. Perdemos esta facilidad y simpleza cuando tenemos muchas características.
- La calidad de la entrada, es la calidad de la salida. La mayoría de las veces, tendremos muchas características no informativas. La entrada de baja calidad producirá una salida de baja calidad.

Hay muchas maneras de seleccionar características, pero la mayoría de los métodos se pueden dividir en tres grupos principales (Agarwal, 2019).

1. Basado en filtro: especificamos algunas métricas y en función de las características de ese filtro. Un ejemplo de tal métrica podría ser correlación / chi-cuadrado.

2. Basado en envoltura: Consideran la selección de un conjunto de características como un problema de búsqueda. Ejemplo: eliminación de funciones recursivas.
3. Embebidos o Incrustados: Utilizan algoritmos que tienen métodos de selección de funciones incorporados. Por ejemplo, Lasso y bosque aleatorio (también conocido como Random Forest o RF) tienen sus propios métodos de selección de funciones.

En este proyecto se usaron los siguientes seleccionadores de características: correlación, chi-cuadrada, eliminación recursiva de características, Lasso, bosque aleatorio y *Light GBM*.

Correlación

Este es un método basado en filtros (Agarwal, 2019). Correlación es un término estadístico que, en el uso común, se refiere a qué tan cerca están dos variables de tener una relación lineal entre sí (Hall, 1999). Las características con alta correlación dependen más linealmente y, por lo tanto, tienen casi el mismo efecto en la variable dependiente. Entonces, cuando dos características tienen una alta correlación, podemos descartar una de las dos características.

Antes de tratar de entender el *p valor*, necesitamos saber sobre la hipótesis nula (Hall, 1999). La hipótesis nula es una afirmación general de que no hay relación entre dos fenómenos medidos. El *p valor* o el valor de probabilidad es un valor de probabilidad para un modelo estadístico dado que, si la hipótesis nula es cierta, un conjunto de observaciones estadísticas más comúnmente conocidas como resumen estadístico es mayor o igual en magnitud a los resultados observados (Biesiada y Duch, 2007; Agarwal, 2019). En otras palabras, el *p valor* indica la probabilidad de encontrar una observación bajo el supuesto de que una hipótesis particular es verdadera. Esta probabilidad se usa para aceptar o rechazar esa hipótesis.

La eliminación de diferentes características del conjunto de datos tendrá diferentes efectos sobre el valor de probabilidad para el conjunto de datos (Hackeling, 2014; Biesiada y Duch, 2007; Hall, 1999). Podemos eliminar diferentes características y medir el *p valor* en cada caso. Estos valores de probabilidad medidos se pueden usar para decidir si mantener una característica o no.

X^2

Este es otro método basado en filtros (Agarwal, 2019). En este método, calculamos la métrica de chi-cuadrado entre el objetivo y la variable numérica y solo seleccionamos la variable con los valores máximos de chi-cuadrado (Hackeling, 2014; Kramer, 2016). La chi-cuadrada está definida de la siguiente manera:

$$\chi^2 = \sum \frac{O_i - E_i{}^2}{E_i}, \tag{20}$$

donde O_i es el número de observaciones en la clase i y E_i es el número de observaciones

esperadas en la clase i.

Eliminación recursiva de características

Este es un método basado en envoltura (Shieh y Yang, 2008; Hackeling, 2014). Los métodos de envoltura consideran la selección de un conjunto de características como un problema de búsqueda (Kramer, 2016).

El objetivo de la eliminación de características recursivas (RFE, por sus siglas en inglés) es seleccionar características considerando recursivamente conjuntos de características cada vez más pequeños (Hackeling, 2014; Shieh y Yang, 2008). Primero, el estimador se entrena en el conjunto inicial de características y la importancia de cada característica se obtiene a través de un atributo. Luego, las características menos importantes se eliminan del conjunto actual de características. Ese procedimiento se repite de forma recursiva en el conjunto actual de características hasta que finalmente se alcanza el número deseado de características para seleccionar.

Se puede usar cualquier estimador con el método (Shieh y Yang, 2008). En este caso, usamos Regresión Logística, y el RFE usará como atributo el coeficiente del objeto de Regresión Logística.

Regresión logística

Este es un método embebido o incrustado (Agarwal, 2019). Los métodos embebidos utilizan algoritmos que tienen métodos de selección de funciones incorporados. Por ejemplo, regresión logística tiene su propio método de selección de características:

$$\sum_{i=1}^{k} |w_i|, \text{ donde } w_i \text{ es el parámetro en la clase } i. \tag{21}$$

El regularizador de Lasso o simplemente Lasso, es la función incorporada en el algoritmo de regresión logística, este obliga a que muchos pesos de características sean cero, usa la norma L1 como regularizador y cuanto mayor sea el valor de alfa, menos características tendrán un valor distinto de cero (Hackeling, 2014).

Bosque aleatorio

También podemos usar bosque aleatorio para seleccionar características basadas en su importancia (Agarwal, 2019).

Para conocer el funcionamiento del algoritmo de bosque aleatorio, revísese las secciones y que hablan sobre árboles de decisión y bosque aleatorio respectivamente.

Calculamos la importancia de la característica utilizando impurezas de nodo en cada árbol

de decisión (Hackeling, 2014; Agarwal, 2019). En el bosque aleatorio, la importancia de la característica final es el promedio de toda la importancia de la característica del árbol de decisión.

LightGBM

Light Gradient Boosting Machine, o simplemente *LightGBM*, es una estructura de potenciación de gradiente que utiliza algoritmos de aprendizaje basados en árboles (Hackeling, 2014; Agarwal, 2019). Está diseñado para distribuir y ser eficiente con las siguientes ventajas (Ke *et al.*, 2017):

- Mayor velocidad de entrenamiento y mayor eficiencia.
- Menor uso de memoria.
- Mejor precisión.
- Soporte de aprendizaje paralelo y GPU.
- Capaz de manejar datos a gran escala.

LightGBM tiene el prefijo *light* debido a su alta velocidad, puede manejar gran tamaño de datos y requiere menos memoria para ejecutarse. *LightGBM* genera un creciente árbol verticalmente, mientras que otros algoritmos de aprendizaje basados en árboles crecen árboles horizontalmente (Ke *et al.*, 2017). Significa que *LightGBM* crece en forma de árbol mientras que otros algoritmos crecen en términos de nivel. Elegirá la hoja con pérdida máxima delta para crecer (*leaf-wise*, como se muestra en la Figura 37). Al cultivar la misma hoja, el algoritmo puede reducir más pérdida que un algoritmo de *level-wise* (construcción de hojas *level-wise* en la Figura 38).

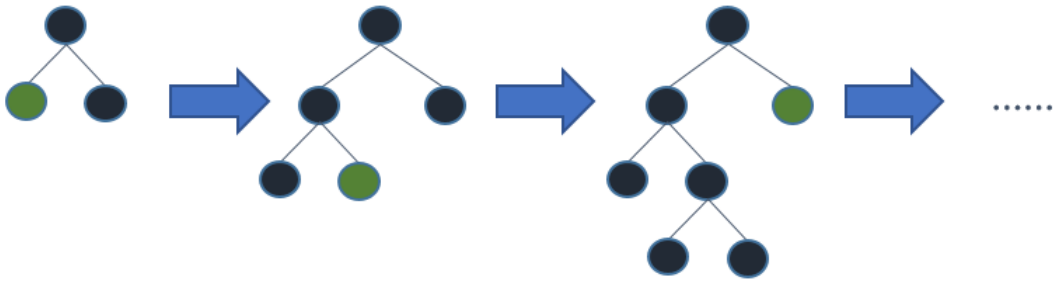

Figura 37. Algoritmo de construcción de hojas en *Light GBM*: *Leaf-Wise.*

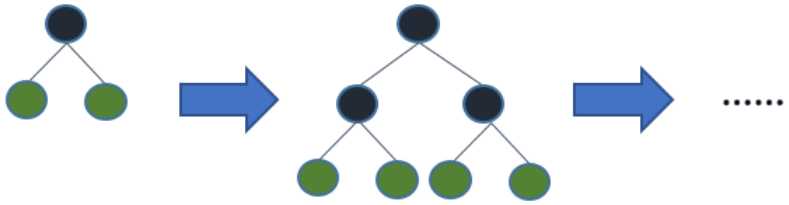

Figura 38. Algoritmo de construcción de hojas en otros algoritmos de aprendizaje basado en árboles: *Level-Wise*.

Los experimentos de comparación en conjuntos de datos públicos muestran que *Light GBM* puede superar los marcos de refuerzo existentes tanto en eficiencia como en precisión, con un consumo de memoria significativamente menor (Ke *et al.*, 2017). Además, los experimentos paralelos muestran que *Light GBM* puede lograr una aceleración lineal mediante el uso de múltiples máquinas para el entrenamiento en entornos específicos. No es aconsejable usar LGBM en pequeños conjuntos de datos. *Light GBM* es sensible al sobreajuste y puede sobreajustar fácilmente datos pequeños.

Validación cruzada

La validación cruzada es una técnica que ayuda a estimar el rendimiento fuera de la muestra de un modelo y evita el sobreajuste (Yiu, 2020). El sobreajuste sucede cuando entrenamos tanto nuestro modelo con los datos existentes que pierde la capacidad de generalizar. Los modelos que se generalizan bien son los que pueden adaptarse con éxito a los nuevos datos, especialmente los que no se parecen a ninguna de las observaciones que el modelo ha visto hasta ahora. Entonces, si un modelo sobreajustado no puede generalizar, es muy probable que tenga un rendimiento errático cuando se pone en producción y realmente se queda sin muestra (Raul Garreta, 2013; Kramer, 2016). Las causas más comunes de sobreajuste son:

- Correlaciones espurias: Si observamos con atención, encontraremos fuertes correlaciones. Si ajustamos nuestro modelo con factores que estén falsamente correlacionados con lo que estemos tratando de predecir, no se generalizará bien. Por ejemplo, podríamos encontrar que el precio de peso mexicano está altamente correlacionado con el precio de la pizza en Ensenada. Pero lo más probable es que se deba al azar, no a nada real, y sería negligente apostar dinero en esa correlación.
- Uso excesivo del conjunto de prueba: Es realmente difícil de evitar por completo esto. Si un modelo no funciona bien en nuestro conjunto de prueba (el conjunto de prueba es la parte de nuestros datos que conservamos para poder evaluar cómo el modelo se generaliza en los nuevos datos), lo modificaremos hasta que encontremos una configuración. Eso funciona bien tanto en el conjunto de entrenamiento como en el conjunto de prueba. La implicación de hacer esto es que el conjunto de prueba ya no es una estimación imparcial del rendimiento de nuestro modelo fuera de la muestra; después de todo,

comenzamos a tomar decisiones de modelado utilizando el conjunto de prueba.

- Un conjunto de entrenamiento sesgado: Nuestros datos de entrenamiento rara vez serán verdaderamente representativos de la población que estamos tratando de modelar. Por lo tanto, debemos ser conscientes de que estamos prácticamente garantizados de encontrarnos con datos que nuestro modelo encuentra completamente desconocidos en algún momento. Y si bien debemos hacer todo lo posible para que las características de nuestra muestra coincidan con las de la población general, también debemos conocer las áreas donde la muestra no es suficiente, ya que los datos de estas áreas representan el mayor riesgo para nuestro modelo. Si nuestra muestra es representativa de solo una pequeña porción de nuestra población, entonces nuestro modelo tendrá un desempeño pobre con el tiempo.

La validación cruzada es una técnica que nos permite producir un conjunto de pruebas utilizando el conjunto de entrenamiento con el apoyo de métricas para asignar un puntaje en determinadas áreas (Yiu, 2020). Es decir, nos permite simular los efectos de “salir de la muestra” utilizando solo nuestros datos de entrenamiento, para que podamos tener una idea de qué tan bien generaliza nuestro modelo. Sin validación cruzada, el proceso de modelo tradicional se ve como en la Fig. 39.

Figura 39. División de conjunto de datos tradicional.

Entrenamos en la parte azul hasta que sentimos que nuestra modelo está listo para enfrentar la naturaleza. Luego lo calificamos en el conjunto de prueba (la parte dorada). El inconveniente de la forma tradicional es que solo tenemos una oportunidad. En el momento en que probamos nuestro modelo en el conjunto de prueba, hemos comprometido nuestros datos de prueba.

La validación cruzada simula cómo podría funcionar nuestro modelo en el conjunto de prueba sin usar realmente el conjunto de prueba (Browne, 2000).

Nos centraremos en un tipo específico de validación cruzada llamada *k*-Pliegues o *k*-iteraciones (por lo tanto, cuando se diga *validación cruzada*, nos referiremos a la validación cruzada de *k*-Pliegues) (Browne, 2000). La validación cruzada de *k*-iteraciones divide nuestros datos de entrenamiento en K pliegues o subsecciones. Luego entrenamos y probamos nuestro modelo

K veces para que todos y cada uno tengan la oportunidad de ser el conjunto de pseudo prueba, lo que llamamos el conjunto de validación (como se muestra en la Fig. 40).

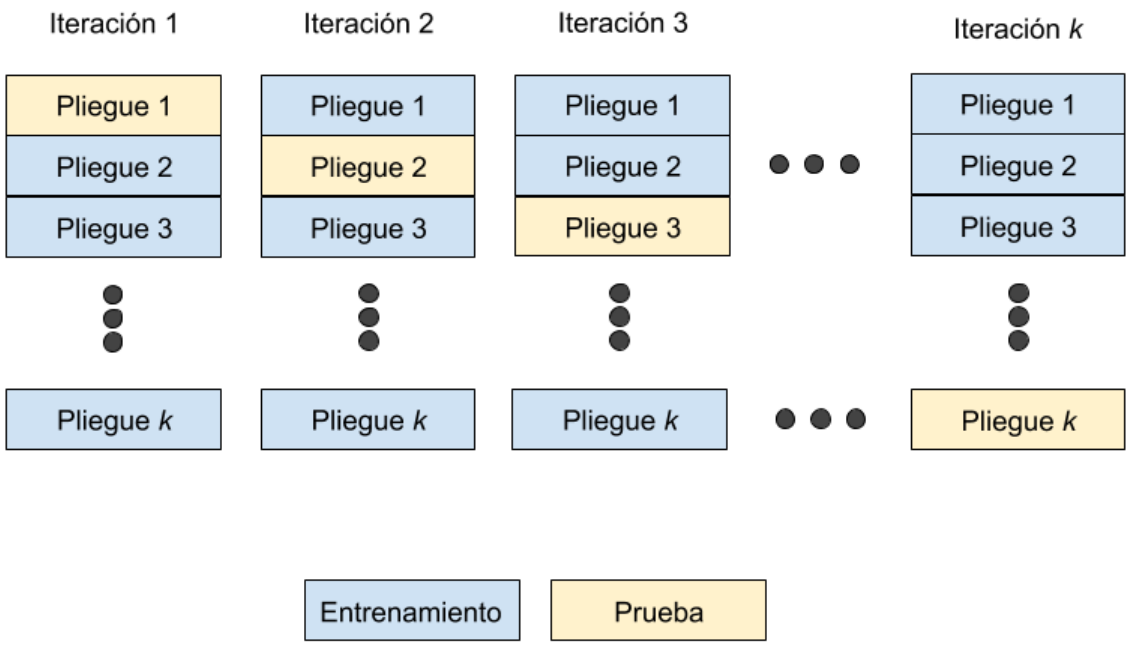

Figura 40. División de conjunto de datos en validación cruzada.

Por ejemplo, suponga que algunos datos los hemos dividido en un conjunto de entrenamiento y un conjunto de prueba. Nuestra principal preocupación es la precisión de las predicciones fuera de muestra de nuestro modelo (qué tan bien se generaliza). Por lo tanto, hemos decidido no considerar el conjunto de pruebas hasta el final (para que podamos darle a nuestro modelo una calificación intelectualmente honesta).

Pero a medida que ajustamos y refinamos nuestro modelo, todavía queremos tener una idea de cómo los cambios que estamos haciendo podrían afectar su rendimiento de muestra. Entonces se realiza la validación cruzada de la siguente manera tomando como ejemplo 3 pliegues (o bien, $K = 3$):

- Decidimos ejecutar una validación cruzada de 3 pliegues, lo que significa que dividimos nuestros datos de entrenamiento en 3 pliegues del mismo tamaño.
- En la primera ejecución de nuestra validación cruzada, el pliegue 1 se mantiene. Por lo tanto, entrenamos en los datos que no están en pliegue 1, y luego validamos en pliegue 1. Esto significa que ajustamos nuestro modelo usando los datos de entrenamiento que no son del pliegue 1, y luego calculamos y registramos qué tan bien predichas las observaciones de la variable dependiente en el pliegue 1. Crucialmente, en la ejecución

1 no utilizamos ninguno de los datos en pliegue 1 durante el entrenamiento de nuestro modelo.

- En la ejecución 2, el pliegue 2 se mantiene. Ahora el pliegue 1, que anteriormente era nuestro conjunto de validación, se ha convertido en parte de nuestro conjunto de entrenamiento. Ajustamos nuestro modelo usando los datos en el pliegue 1 y el pliegue 3, y lo calificamos usando el pliegue 2.
- Después de que concluye la ejecución 3, ahora tenemos tres valores (ya que cada pliegue tiene un turno para mantenerse). El promedio de los tres nos da una estimación decente con el puntaje respecto a alguna métrica que se haya elegido inicialmente. Esto nos da una idea de cómo se comporta fuera de la muestra nuestro modelo.

La validación cruzada se puede utilizar para comparar los resultados de diferentes procedimientos de clasificación predictiva (Browne, 2000; Kramer, 2016; Yiu, 2020). Por ejemplo, supongamos que tenemos un detector que nos determina si una cara pertenece a una mujer o a un hombre y consideramos que han sido utilizados dos métodos diferentes, por ejemplo, máquinas de soporte vectorial (SVM) y *k*-vecinos más cercanos (K-NN), ya que ambos nos permiten clasificar las imágenes. Con la validación cruzada podríamos comparar los dos procedimientos y determinar cuál de los dos es el más preciso. Esta información nos la proporciona la tasa de error que obtenemos al aplicar la validación cruzada por cada uno de los métodos planteados y las métricas.

La validación cruzada de *k*-iteraciones nos permite evaluar también modelos en los que se utilizan varios clasificadores (Browne, 2000; Yiu, 2020). Continuando con el ejemplo anterior, si tenemos un detector que nos determina si en una imagen aparece un hombre o una mujer, y éste utiliza cuatro clasificadores binarios para detectarlo, también podemos utilizar la validación cruzada para evaluar su precisión. Si tenemos un total de 20 datos (imágenes), y utilizamos el método validación cruzada con 4-iteraciones, se llevarán a cabo cuatro iteraciones, y en cada una se utilizarán unos datos de entrenamiento diferentes, que serán analizadas por cuatro clasificadores, que posteriormente evaluarán los datos de prueba. De este modo por cada muestra obtendremos cuatro resultados, y si hacemos la media entre los resultados de cada clasificador y entre las cuatro iteraciones realizadas, obtendremos el valor resultante final.

Clasificadores

La clasificación es el proceso de predecir la clase de puntos de datos dados (Hackeling, 2014). Las clases a veces son objetivos, etiquetas o categorías. Un modelo predictivo de clasificación se encarga de aproximar una función de mapeo *(f)* de variables de entrada *(X)* a variables de salida discretas *(y)*. Por ejemplo, la detección de spam en los proveedores de servicios de

correo electrónico puede identificarse como un problema de clasificación. Esta es una clasificación binaria, ya que solo hay dos clases de etiquetas: spam y no spam. Un clasificador utiliza algunos datos de entrenamiento para comprender como las variables de entrada dadas se relacionan con la clase. En este caso, los correos electrónicos no deseados y no conocidos deben usarse como datos de capacitación. Cuando el clasificador se entrena con precisión, se puede usar para detectar un correo electrónico desconocido.

La clasificación pertenece a la categoría de aprendizaje supervisado donde los objetivos también proporcionan los datos de entrada (Edwards, 2018; Kramer, 2016). Hay muchas aplicaciones en clasificación en muchos dominios, como aprobación de crédito, diagnóstico médico, marketing de objetivos, etc. Hay dos tipos de aprendizajes en la clasificación, perezosos (en inglés, *lazy learner*) y ansiosos (en inglés, *eager learner*) (Grimaldi *et al.*, 2006; Verbeke *et al.*, 2014).

Los algoritmos de aprendizaje perezoso simplemente almacenan los datos de entrenamiento y esperan hasta que aparezca un dato de prueba (Grimaldi *et al.*, 2006). Cuando lo hace, la clasificación se realiza con base en los datos más relacionados en los datos de entrenamiento almacenados. En comparación con los algoritmos de aprendizaje ansioso, los perezosos tienen menos tiempo de entrenamiento, pero más tiempo para predecir. Ejemplos de algoritmos de aprendizaje perezoso: *k*-vecino más cercano, razonamiento basado en casos.

Los algoritmos de aprendizaje ansioso construyen un modelo de clasificación basado en los datos de entrenamiento dados antes de recibir datos para la clasificación (Verbeke *et al.*, 2014). Debe poder comprometerse con una sola hipótesis que cubra todo el espacio de la instancia. Debido a la construcción del modelo, los algoritmos de aprendizaje ansioso tardan mucho tiempo en entrenar y menos tiempo para predecir. Ejemplos de algoritmos de aprendizaje ansioso: Árbol de decisión, Naive Bayes, Redes neuronales artificiales.

Hay muchos algoritmos de clasificación disponibles, y no es posible concluir cuál es superior a otro (Grimaldi *et al.*, 2006; Verbeke *et al.*, 2014). Depende de la aplicación y la naturaleza del conjunto de datos disponible. Por ejemplo, si las clases son linealmente separables, los clasificadores lineales como la regresión logística, y el discriminante lineal de Fisher, pueden superar a los modelos sofisticados y viceversa. En este proyecto se usaron los siguientes clasificadores: SVM lineal, bosque aleatorio, perceptrón multicapa (MLP), árbol de decisión y Naive Bayes.

Máquina de soporte vectorial

SVM o Maquina Soporte Vectorial es un modelo para problemas de clasificación y regresión (Abraham *et al.*, 2014). Puede resolver problemas lineales y no lineales. Funciona bien para muchos problemas prácticos. La idea de SVM es simple: el algoritmo crea una línea o un hi-

perplano que separa los datos en clases. De acuerdo con el algoritmo SVM, encontramos los puntos más cercanos a la línea de ambas clases. Estos puntos se denominan vectores de soporte (Abraham *et al.*, 2014; Kramer, 2016; Edwards, 2018). Ahora, calculamos la distancia entre la línea y los vectores de soporte. Esta distancia se llama margen. Nuestro objetivo es maximizar el margen. El hiperplano para el que el margen es máximo se le conoce como *hiperplano óptimo*. Por lo tanto, SVM intenta establecer un límite de decisión de tal manera que la separación entre las dos clases sea lo más amplia posible. La manera más simple de realizar la separación es mediante una línea recta, un plano recto o un hiperplano *N*-dimensional. Un ejemplo de esto se puede apreciar en la Figura 41 en la gráfica izquierda.

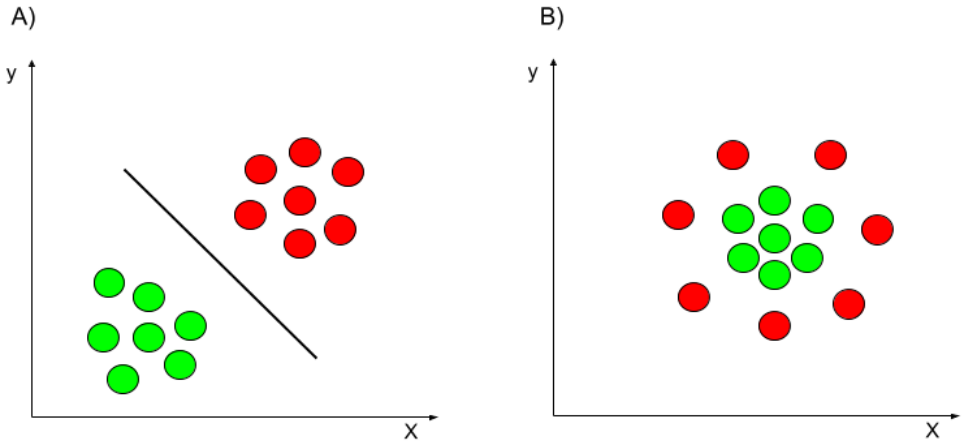

Figura 41. Conjunto de puntos linealmente separable (A) y conjunto de puntos linealmente no separable (B) por medio de SVM lineal.

Desafortunadamente los universos a estudiar no se suelen presentar en casos idílicos de dos dimensiones, sino que un algoritmo SVM debe tratar con (Abraham *et al.*, 2014; Kramer, 2016; Hackeling, 2014; Edwards, 2018; Raul Garreta, 2013):

1. Más de dos variables predictoras.
2. Curvas no lineales de separación (véase Fig. 41, gráfica derecha).
3. Casos donde los conjuntos de datos no pueden ser completamente separados.
4. Clasificaciones en más de dos categorías.

Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, éstas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real (Abraham *et al.*, 2014). La representación por medio de funciones Kernel ofrece una solución a este problema, proyectando la información a un espacio de características de mayor dimensión el cual aumenta la

capacidad computacional de las máquinas de aprendizaje lineal. Es decir, mapear el espacio de entrada *X* a un nuevo espacio de características de mayor dimensionalidad (Figura 42).

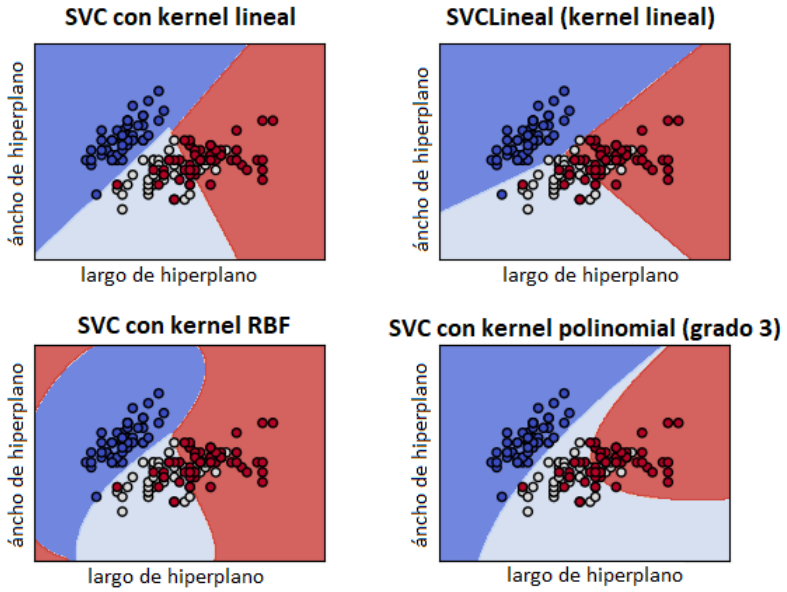

Figura 42. Diferencia de SVM usando kernel lineal, RBF y polinomial de grado 3 (imagen recuperada y adaptada de *https://scikit-learn.org/stable/modules/svm.html*).

De acuerdo a Alexandre Abraham y colaboradores (2014), las ventajas de la SVM son:

- Efectivo en espacios de altas dimensiones.
- Sigue siendo efectivo en casos donde el número de dimensiones es mayor que el número de muestras.
- Utiliza un subconjunto de puntos de entrenamiento en la función de decisión (llamados vectores de soporte), por lo que también es eficiente en la memoria.
- Versátil: se pueden especificar diferentes funciones de Kernel para la función de decisión. Se proporcionan núcleos comunes, pero también es posible especificar núcleos personalizados.

Las desventajas, de acuerdo a distintos autores, de los SVM incluyen (Abraham *et al.*, 2014; Kramer, 2016; Hackeling, 2014):

- Si el número de características es mucho mayor que el número de muestras, es recomendable evitar el ajuste excesivo al elegir las funciones del núcleo y el término de regularización es crucial.

- Los SVM no proporcionan directamente estimaciones de probabilidad, estas se calculan utilizando una costosa validación cruzada.

Árbol de decisión

Los árboles de decisión (DT, por sus siglas en inglés) son un método de aprendizaje supervisado no paramétrico utilizado para la clasificación y la regresión (Pal, 2005). Los árboles de decisión aprenden de los datos para aproximar una curva sinusoidal con un conjunto de reglas de decisión si-entonces-otro. Cuanto más profundo es el árbol, más complejas son las reglas de decisión y más se ajusta el modelo.

El árbol de decisión crea modelos de clasificación o regresión en forma de estructura de árbol (Ali *et al.*, 2012). Desglosa un conjunto de datos en subconjuntos cada vez más pequeños, mientras que al mismo tiempo se desarrolla un árbol de decisión asociado. El resultado final es un árbol con nodos de decisión y nodos hoja como se puede ver en la Figura 43. Un nodo de decisión tiene dos o más ramas. El nodo hoja representa una clasificación o decisión. El nodo de decisión superior en un árbol que corresponde al mejor predictor, es llamado nodo raíz. Los árboles de decisión pueden manejar datos categóricos y numéricos.

Figura 43. Árbol de decisión simple.

De acuerdo a Pal, Ali y colaboradores (Pal, 2005; Ali *et al.*, 2012), las ventajas de los árboles de decisión son:

- Simple de entender e interpretar. Los árboles pueden ser visualizados.
- Requiere poca preparación de datos. Otras técnicas a menudo requieren la normalización de datos, se deben crear variables ficticias y eliminar los valores en blanco.

- El costo computacional es logarítmico en la cantidad de puntos de datos utilizados para entrenar el árbol.
- Capaz de manejar datos numéricos y categóricos.
- Capaz de manejar problemas de salida múltiple.
- Utiliza un modelo de caja blanca. Si una situación dada es observable en un modelo, la explicación de la condición se explica fácilmente por la lógica booleana. Por el contrario, en un modelo de caja negra (por ejemplo, en una red neuronal artificial), los resultados pueden ser más difíciles de interpretar.
- Posible validar un modelo utilizando pruebas estadísticas.
- Funciona bien incluso si sus supuestos son violados de alguna manera por el modelo verdadero a partir del cual se generaron los datos.

Las desventajas, de acuerdo a Pal, Ali y colaboradores (Pal, 2005; Ali *et al.*, 2012), son:

- Se pueden crear árboles demasiado complejos que no generalizan bien los datos (sobreajuste).
- Los árboles de decisión pueden ser inestables. Pequeñas variaciones en los datos pueden generar un árbol completamente diferente. Este problema se mitiga mediante el uso de árboles de decisión dentro de un conjunto.
- Un árbol de decisión óptimo es NP-completo bajo varios aspectos de la optimización e incluso para conceptos simples. En consecuencia (Abraham *et al.*, 2014), los algoritmos prácticos de aprendizaje del árbol de decisiones se basan en algoritmos heurísticos donde se toman decisiones localmente óptimas en cada nodo. Tales algoritmos no pueden garantizar la devolución del árbol de decisión globalmente óptimo.
- Hay conceptos que son difíciles de aprender porque los árboles de decisión no los expresan fácilmente. Por ejemplo: XOR, problemas de paridad, etc.
- Se pueden crear árboles sesgados si algunas clases dominan. Por lo tanto, se recomienda equilibrar el conjunto de datos antes de ajustarlo con el árbol de decisión.

Bosque aleatorio

Random forest o bosque aleatorio, consiste en una gran cantidad de árboles de decisión individuales que operan en conjunto (Pal, 2005; Ali *et al.*, 2012). Cada árbol individual en el bosque aleatorio da una predicción de clase y la clase con más votos se convierte en la predicción de nuestro modelo. En la Figura 44 se puede apreciar como el bosque aleatorio está compuesto por diversos árboles de decisión.

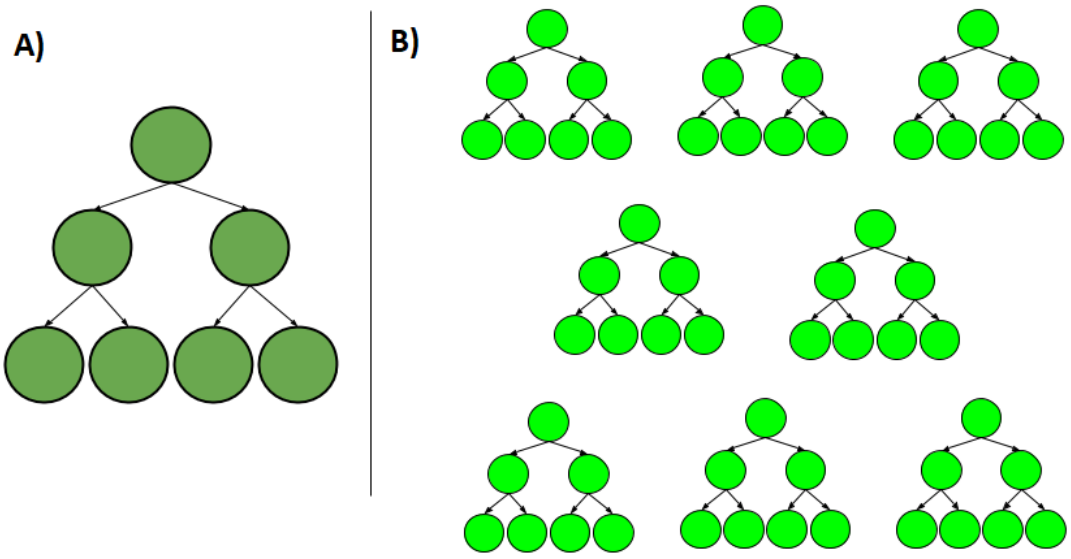

Figura 44. Comparación entre árbol de decisión (A) y bosque aleatorio (B).

De acuerdo a Pal, Ali y colaboradores (Pal, 2005; Ali *et al.*, 2012), las ventajas del bosque aleatorio:

- Se basa en el algoritmo de *bagging* y utiliza la técnica *Ensemble Learning*. Crea tantos árboles en el subconjunto de datos y combina la salida de todos los árboles. De esta manera, reduce el problema de sobreajuste en los árboles de decisión y también reduce la varianza y, por lo tanto, mejora la precisión.
- Puede usarse para resolver problemas de clasificación y regresión.
- Funciona bien con variables categóricas y continuas.
- Puede manejar automáticamente los valores faltantes.
- No se requiere escalamiento de características, ya que utiliza un enfoque basado en reglas en lugar del cálculo de distancia.
- Maneja los parámetros no lineales de manera eficiente.
- Suele ser robusto para los valores atípicos y puede manejarlos automáticamente.
- Es muy estable. Incluso si se introduce un nuevo punto de datos en el conjunto de datos, el algoritmo general no se ve afectado mucho ya que los nuevos datos pueden afectar a un árbol, pero es muy difícil que afecte a todos los árboles.
- Es comparativamente menos afectado por el ruido.

Las desventajas, de acuerdo a Pal, Ali y colaboradores (Pal, 2005; Ali *et al.*, 2012), de dicho algoritmo son:

- Complejidad: *Random Forest* crea muchos árboles (a diferencia de un solo árbol en caso de árbol de decisión) y combina sus resultados. Para hacerlo, este algoritmo requiere mucha más potencia y recursos computacionales. Por otro lado, el árbol de decisión es simple y no requiere tantos recursos computacionales.
- Período de entrenamiento más largo: Bosque aleatorio requiere mucho más tiempo para entrenar en comparación con los árboles de decisión, ya que genera muchos árboles y toma la decisión sobre la mayoría de los votos.

Perceptrón multicapa

Un perceptrón es un clasificador lineal; es decir, es un algoritmo que clasifica la entrada al separar dos categorías con una línea recta (Suykens y Vandewalle, 1999). Un perceptrón multicapa (MLP) es un algoritmo de aprendizaje supervisado que aprende una función $f : R^n \longrightarrow R^m$ entrenando en un conjunto de datos, donde n es el número de dimensiones de entrada y n es el número de dimensiones de salida. Dado un conjunto de características $X = x_1, x_2, ..., x_n$ y un objetivo y, puede aprenderse un aproximador de función no lineal para clasificación o regresión. Es diferente de la regresión logística, en que entre la capa de entrada y la de salida, puede haber una o más capas no lineales, llamadas capas ocultas. La Figura 45 muestra un MLP de una capa oculta con salida escalar.

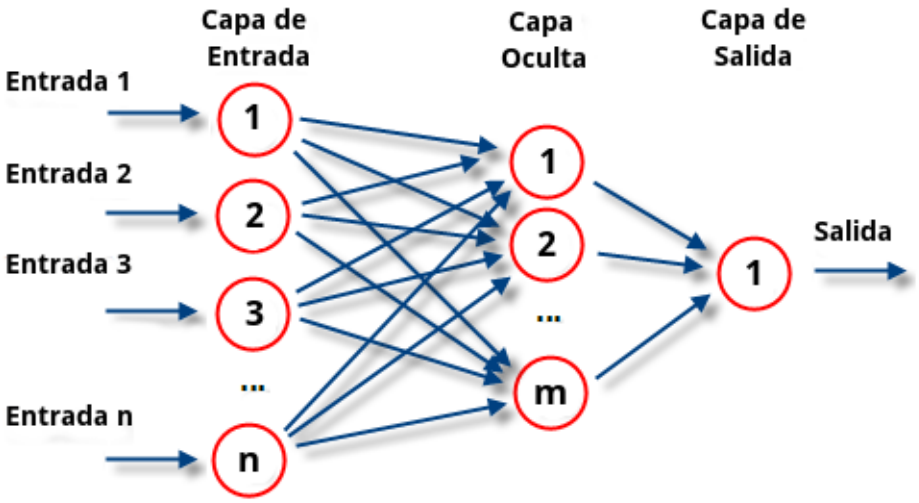

Figura 45. Diagrama de un MLP y sus capas.

La capa más a la izquierda, conocida como capa de entrada, consta de un conjunto de neuronas $x_i|x_1, x_2, ..., x_m$ que representa las características de entrada. Cada neurona de la capa oculta transforma los valores de la capa anterior con una suma lineal ponderada $w_1x_1 + w_2x_2 + ... + w_mx_m$, seguido de una función de activación no lineal $g : R \longrightarrow R$ (por ejemplo, la función *tangente hiperbólica*). La capa de salida recibe los valores de la última capa oculta y los transforma en valores de salida.

De acuerdo a Zhao y colaboradores (Zhao *et al.*, 2012), las ventajas del MLP son:

- Capacidad para aprender modelos no lineales.
- Capacidad para aprender modelos en tiempo real.

Las desventajas del MLP, dado el trabajo de Zhao, Suykens y colaboradores (Zhao *et al.*, 2012; Suykens y Vandewalle, 1999), son:

- Tiene una función de pérdida no convexa donde existe más de un mínimo local. Por lo tanto, diferentes inicializaciones de peso al azar pueden conducir a una precisión de validación diferente.
- Requiere ajustar una cantidad de hiperparámetros, como la cantidad de neuronas, capas e iteraciones ocultas.
- Es sensible al escalado de características.

Naive Bayes

Naive Bayes o Bayeasiano Ingenuo, es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales (Rish, 2001). Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de *naive*, es decir, ingenuo. En términos simples, un clasificador Naive Bayes supone que la presencia de una característica particular en una clase no está relacionada con la presencia de ninguna otra característica, representado matemáticamente en la Ecuación 22:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{22}$$

Para otros modelos de probabilidad, los clasificadores de Naive Bayes se pueden entrenar de manera muy eficiente en un entorno de aprendizaje supervisado (Rish, 2001; Mukherjee y Sharma, 2012). En muchas aplicaciones prácticas, la estimación de parámetros para los modelos Naive Bayes utiliza el método de máxima verosimilitud, en otras palabras, se puede trabajar con el modelo de Naive Bayes sin aceptar probabilidad bayesiana o cualquiera de los métodos bayesianos.

Los algoritmos ingenuos de Bayes se utilizan principalmente en el análisis de sentimientos, el filtrado de spam, los sistemas de recomendación, etc. (Mukherjee y Sharma, 2012). Son rápidos y fáciles de implementar, pero su mayor desventaja es que los predictores o características deben ser independientes (Vaidya y Clifton, 2004). En la mayoría de los casos de la vida real, los predictores son dependientes, lo que dificulta el rendimiento del clasificador. Una ventaja del clasificador de Naive Bayes es que solo se requiere una pequeña cantidad de datos

de entrenamiento para estimar los parámetros (las medias y las varianzas de las variables) necesario para la clasificación (Vaidya y Clifton, 2004; Mukherjee y Sharma, 2012; Rish, 2001). Como las variables independientes se asumen, solo es necesario determinar las varianzas de las variables de cada clase y no toda la matriz de covarianza.

Métricas y desempeño

Después de hacer selección de características, implementar un modelo y obtener algunos resultados en forma de probabilidad o clase, el siguiente paso es descubrir qué tan efectivo es el modelo basado en algunas métricas utilizando conjuntos de datos de prueba.

La elección de métricas es muy importante. Influye en como se mide y compara el rendimiento de los algoritmos de aprendizaje automático (de Vazelhes *et al.*, 2019). En este escrito revisaremos las métricas que se emplearon: Exactitud, precisión, sensibilidad, métrica f (f1) y *hamming loss*.

En un modelo binario de clasificación, podemos considerar dos clases: 'Positiva' y 'Negativa' (de Vazelhes *et al.*, 2019). Para evaluar este modelo, se puede calcular su precisión, como la proporción entre las predicciones correctas que ha hecho el modelo y el total de predicciones. Sin embargo, aunque en ocasiones resulta práctico por su facilidad de cálculo, otras veces es necesario profundizar un poco más y tener en cuenta los tipos de predicciones correctas e incorrectas que realiza el clasificador. Por esto, es importante conocer primeramente la 'Matriz de Confusión'.

La matriz de confusión de un problema de clase *n* es una matríz *nxn* en la que las filas se nombran según las clases reales y las columnas, según las clases previstas por el modelo (Hackeling, 2014). Sirve para mostrar de forma explícita cuándo una clase es confundida con otra, de ahí el nombre. Por eso, permite trabajar de forma separada con distintos tipos de error.

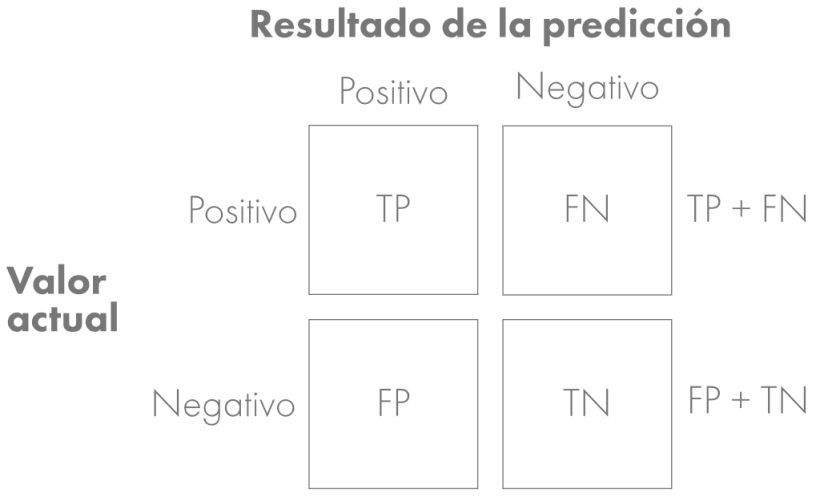

Figura 46. Matriz de confusión para un clasificador binario.

En la matriz de confusión de la Figura 46 se puede ver los nombres genéricos cuando usamos la nomenclatura inglesa: Verdaderos Negativos = True Negative (TN), Verdadeors Positivos = True Positive (TP), Falsos Positivos = False Positive (FP), Falsos Negativos = False Negative (FN).

El que sea *Positivo (Positive)* o *Negativo (Negative)* se refiere a la predicción. Si el modelo predice 1 entonces será positivo, y se predice 0 será negativo. Las etiquetas *Verdadero (True)* o *Falso (False)* se refiere si la predicción es correcta o no.

Exactitud

La métrica de *exactitud*, o en inglés, *accuracy*, indica el número de elementos clasificados correctamente en comparación con el número total de artículos (Jeni *et al.*, 2013). Sin embargo, la métrica de exactitud tiene limitaciones: no funciona bien con las clases no balanceadas, o bien, clases que pueden tener muchos elementos de la misma clase e incluir algunas otras clases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{23}$$

Precisión

La métrica de *precisión* representa el número de verdaderos positivos que son realmente positivos en comparación con el número total de valores positivos predichos (Davis y Goadrich, 2006).

$$Precision = \frac{TP}{TP + FP}. \tag{24}$$

Sensibilidad

La métrica de *sensibilidad*, *exhaustividad*, o en inglés, *recall*, muestra la cantidad de verdaderos positivos que el modelo ha clasificado en función del número total de valores positivos (Davis y Goadrich, 2006).

$$Recall = \frac{TP}{TP + FN}. \tag{25}$$

Métrica-f

La métrica f, o f1, se utiliza para combinar las medidas de *precisión* y *sensibilidad* en un sólo valor (Davis y Goadrich, 2006). Esto es práctico porque hace más fácil comparar el rendimiento combinado de la precisión y la sensibilidad (Jeni *et al.*, 2013). f1 se calcula haciendo la media armónica entre la *precisión* y la *sensibilidad*:

$$f1 = 2 * \frac{precision * recall}{precision + recall}. \tag{26}$$

Pérdida de Hamming

La *pérdida de hamming*, o *hamming loss* en inglés, indica el número de elementos clasificados erróneamente en comparación con el número total de artículos (Dembczyński *et al.*, 2010).

$$HL = 1 - accuracy. \tag{27}$$

Métricas para clasificación multiclase

Algunas métricas se definen esencialmente para tareas de clasificación binaria (por ejemplo, f1). En estos casos, de forma predeterminada solo se evalúa la etiqueta positiva (Hackeling, 2014). Al extender una métrica binaria a problemas multiclase o multietiqueta, los datos se tratan como una colección de problemas binarios, uno para cada clase (de Vazelhes *et al.*, 2019). Hay varias maneras de promediar cálculos métricos binarios en el conjunto de clases, cada uno de los cuales puede ser útil en algún escenario. En este trabajo nos enfocaremos en los promedios de *micro* y *macro*.

El promedio macro simplemente calcula la media de las métricas binarias, dando el mismo peso a cada clase (Hackeling, 2014). Por ejemplo, la sensibilidad con promedio macro se da de la siguiente manera:

$$Recall_{macro} = \frac{recall_1 + recall_2}{2}. \tag{28}$$

En problemas donde las clases poco frecuentes son importantes, el promedio macro puede ser un medio para resaltar su desempeño (Hackeling, 2014). Por otro lado, la suposición de que todas las clases son igualmente importantes a menudo es falsa, de modo que el promedio de macro enfatizará en exceso el rendimiento típicamente bajo en una clase poco frecuente.

El promedio micro le da a cada par de clase de muestra una contribución igual a la métrica general (excepto como resultado del peso de la muestra) (Hackeling, 2014). Por ejemplo, la sensibilidad con promedio micro se da de la siguiente manera:

$$Recall_{micro} = \frac{TP_1 + TP_2}{TP_1 + FN_1 + TP_2 + FN_2}. \tag{29}$$

En lugar de sumar la métrica por clase, esto suma los dividendos y divisores que componen las métricas por clase para calcular un cociente global (Hackeling, 2014). Se puede preferir el promedio micro en entornos de multietiquetas y multiclase donde se debe ignorar una clase mayoritaria.

Media armónica

La media armónica (representada como H) de una cantidad n finita de números es igual al

inverso (Uribe y Álvarez, 1978), de la media aritmética de los inversos de dichos valores. Esta media se caracteriza por ser sensible a la dispersión de los datos. Además, no está definida en el caso de que exista algún valor nulo. La media armónica es representada de manera general con la ecuación siguiente:

$$H = \frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}}. \tag{30}$$

Por ejemplo, al tener dos pares de calificaciones $x_1 = 1$, $y_1 = 100$, $x_2 = 50.5$ y $y_2 = 50.5$, la media aritmética de x_1 y y_1, se puede calcular con la Ecuación 31, es 50.5, mientras que la media aritmética del par x_2 y y_2 también lo es. Ahora bien, la media armónica del par x_1 y y_1, que se puede calcular con la Ecuación 30, es 1.98, y la media armónica del segundo par es 50.5. Lo cual nos da un resultado acorde a la dispersión de las calificaciones.

$$\text{media aritmética} = \frac{x_1 + x_2 + \ldots + x_n}{n}. \tag{31}$$

Apéndice II

En este apartado se encuentran los acrónimos usados en el documento.

Tabla 15. Tabla de acrónimos de los indicadores técnicos manejados en el presente trabajo.

Indicador técnico	Acrónimo
Moving Average Convergence Divergence	MACD
Average Directional Movement Index	ADX
Vortex Indicator	VI
Trix	TRIX
Mass Index	MI
Commodity Channel Index	CCI
Detrended Price Oscillator	DPO
KST Oscillator	KST
Ichimoku Kinkō Hyō	Ichimoku
Parabolic Stop And Reverse	Parabolic SAR
Money Flow Index	MFI
Relative Strength Index	RSI
True strength index	TSI
Ultimate Oscillator	UO
Stochastic Oscillator	SR
Williams %R	WR
Awesome Oscillator	AO
Kaufman’s Adaptive Moving Average	KAMA
Rate of Change	ROC
Average True Range	ATR
Bollinger Bands	BB
Keltner Channel	KC
Donchian Channel	DC
Accumulation/Distribution Index	ADI
On-Balance Volume	OBV
Chaikin Money Flow	CMF
Force Index	FI
Ease of Movement	EoM, EMV
Volume-price Trend	VPT
Negative Volume Index	NVI
Volume Weighted Average Price	VWAP
Daily Return	DR
Daily Log Return	DLR
Cumulative Return	CR

Los acrónimos de las divisas pueden combinarse para formar un "par de divisas", por ejemplo: *NZDUSD* que se lee *dólar neozelandés-dólar americano*.

Tabla 16. Tabla de acrónimos de cada una de las divisas manejadas en el presente trabajo.

Divisa	Acrónimo
Dólar americano	USD
Dólar neozelandés	NZD
Dólar australiano	AUD
Libra esterlina	GBP
Dólar canadiense	CAD
Franco suizo	CHF
Yen japonés	JPY
Dólar de singuapur	SGD
Euro	EUR

Apéndice III: Escenario 1

En este apartado se encuentran los resultados detallados de los entrenamientos y pruebas del primer escenario de la fase de experimentación.

Tabla 17. Conjuntos de características generados para el primer escenario.

Conjunto	Tipo de indicador	Indicador	Versión
Correlación (21)	Tendencia	ADX	0, 2, 3 y 4
	Tendencia	Parabolic SAR	1
	Volatilidad	Bollinger Bands	0, 4 y 5
	Volatilidad	Donchian Channel	0 al 5
X^2 (21)	Volatilidad	Bollinger Bands	0, 2, 3, 4 y 5
	Volatilidad	Donchian Channel	0 al 5
R.F.E. (14)	Volatilidad	Bollinger Bands	0, 4 y 5
	Volatilidad	Dochian Channel	0, 1, 4 y 5
Regresión	Tendencia	MACD	0
logística (56)	Tendencia	Mass Index	0
	Tendencia	ADX	0, 2, 3 y 4
	Tendencia	Trix	1, 4 y 5
	Tendencia	DPO	0, 4 y 5
	Tendencia	Parabolic SAR	0, 2, 3 y 5
	Oscilador	Stochastic Oscillator	0
	Oscilador	Williams %R	0
	Oscilador	KAMA	0
	Volatilidad	ATR	0 al 5
	Volatilidad	Donchian Channel	0 al 5
	Volatilidad	Bollinger Bands	0, 2, 3, 4 y 5
	Volumen	ADI	0
	Volumen	Ease of Movement	0

Conjunto	Tipo de indicador	Indicador	Versión
	Volumen	Volume-price Trend	0
Bosque aleatorio (44)	Tendencia	MACD	1, 2, 3 y 5
	Tendencia	Trix	2
	Tendencia	CCI	1, 3 y 4
	Tendencia	Parabolic SAR	0, 1, 2, 4 y 5
	Oscilador	Stochastic Oscillator	0
	Oscilador	RSI	0 al 5
	Volatilidad	Bollinger Bands	1 y 2
	Volatilidad	Keltner Channel	1, 2, y 5
	Volatilidad	Donchian Channel	0, 1, 2, 4 y 5
	Volumen	Volume-price Trend	0
	Otros	Daily Return	0
	Otros	Daily Log Return	0
LightGBM (97)	Tendencia	MACD	0
	Tendencia	Mass Index	0
	Tendencia	KST Oscillator	0
	Tendencia	Ichimoku Kinkō Hyō	0
	Tendencia	Vortex Indicator	0, 1, 2, 4 y 5
	Tendencia	Parabolic SAR	0, 2, 3 y 4
	Tendencia	ADX	0 al 5
	Tendencia	Trix	0 al 5
	Tendencia	CCI	0 al 5
	Tendencia	DPO	0 al 5
	Oscilador	True Strenght Index	0
	Oscilador	Stochastic Oscillator	0
	Oscilador	Williams %R	0
	Oscilador	Awesome Oscillator	0
	Oscilador	Money Flow Index	0
	Oscilador	RSI	0 al 5
	Volatilidad	Bollinger Bands	1
	Volatilidad	Donchian Channel	1
	Volatilidad	ATR	0 al 5
	Volumen	ADI	0
	Volumen	On-Balance Volume	0
	Volumen	Chaikin Money Flow	0
	Volumen	Force Index	0
	Volumen	Volume-price Trend	0

Conjunto	Tipo de indicador	Indicador	Versión
	Volumen	Negative Volume Index	0
	Volumen	Ease of Movement	0, 2, 3, 4 y 5
	Otros	Daily Return	0
	Otros	Daily Log Return	0
	Otros	Cumulative Return	0

Tabla 18. Resultados de entrenamiento en escenario 1 con el conjunto de correlación.

Conjunto: CORRELACIÓN								
Alg.	f1		Sensibilidad		Precisión		Exactitud	
	μ	σ	μ	σ	μ	σ	μ	σ
SVM (Lineal)	0.6399	0.2933	0.755	0.3794	0.6331	0.2349	0.676	0.0938
B.A. 100	0.7374	0.0065	0.7749	0.0163	0.7035	0.0075	0.7241	0.0053
B.A. 200	0.7396	0.0066	0.7802	0.0162	0.7033	0.0069	0.7254	0.0053
M.L.P. 50	0.7485	0.031	0.8135	0.0744	0.6955	0.0119	0.728	0.0201
M.L.P. 100	0.7399	0.0184	0.7832	0.0488	0.7033	0.0141	0.7256	0.0111
A.D. (Mejor)	0.6785	0.0062	0.6678	0.0133	0.6898	0.0065	0.6837	0.0044
A.D. (Aleatorio)	0.678	0.0077	0.6659	0.0136	0.6908	0.0075	0.6838	0.0062
N. Bayes	0.7214	0.0093	0.6931	0.0176	0.7524	0.0052	0.7325	0.0059

Tabla 19. Resultados de entrenamiento en escenario 1 con el conjunto de X^2.

Conjunto: CHI-CUADRADA								
Alg.	f1		Sensibilidad		Precisión		Exactitud	
	μ	σ	μ	σ	μ	σ	μ	σ
SVM (Lineal)	0.7225	0.008	0.7967	0.0245	0.6612	0.0069	0.6941	0.0042
B.A. 100	0.7222	0.008	0.7962	0.0235	0.6614	0.007	0.694	0.0049
B.A. 200	0.7222	0.0068	0.796	0.0221	0.6614	0.007	0.694	0.0038
M.L.P. 50	0.727	0.0268	0.7867	0.0508	0.6776	0.0284	0.705	0.026
M.L.P. 100	0.7294	0.0318	0.7994	0.0585	0.6713	0.0288	0.7036	0.0286
A.D. (Mejor)	0.7217	0.0067	0.7949	0.022	0.6613	0.007	0.6937	0.0037
A.D. (Aleatorio)	0.7217	0.0067	0.795	0.022	0.6613	0.007	0.6937	0.0037
N. Bayes	0.7199	0.0075	0.6943	0.0181	0.748	0.0126	0.73	0.0056

Tabla 20. Resultados de entrenamiento en escenario 1 con el conjunto de R.F.E.

Conjunto: R.F.E.								
Alg.	f1		Sensibilidad		Precisión		Exactitud	
	μ	σ	μ	σ	μ	σ	μ	σ
SVM (Lineal)	0.795	0.0049	0.9996	0.0006	0.66	0.0068	0.7422	0.0076
B.A. 100	0.914	0.009	0.9665	0.0117	0.8673	0.0183	0.909	0.0101
B.A. 200	0.9145	0.0095	0.968	0.0115	0.867	0.0187	0.9094	0.0106
M.L.P. 50	0.8622	0.0577	0.9555	0.0893	0.7915	0.0557	0.8469	0.0605
M.L.P. 100	0.8635	0.0255	0.8997	0.0765	0.8349	0.0284	0.861	0.0198
A.D. (Mejor)	0.8666	0.0082	0.8586	0.0137	0.8753	0.0169	0.8679	0.0085
A.D. (Aleatorio)	0.8625	0.0101	0.8518	0.0182	0.8741	0.0174	0.8643	0.0097
N. Bayes	0.2885	0.0198	0.2003	0.0174	0.5213	0.0536	0.5068	0.0151

Tabla 21. Resultados de entrenamiento en escenario 1 con el conjunto de regresión logística.

Conjunto: REGRESIÓN LOGISTICA								
Alg.	f1		Sensibilidad		Precisión		Exactitud	
	μ	σ	μ	σ	μ	σ	μ	σ
SVM (Lineal)	0.7641	0.2866	0.8921	0.3466	0.6778	0.1259	0.726	0.0924
B.A. 100	0.9268	0.0082	0.9606	0.0111	0.8956	0.0168	0.9241	0.0089
B.A. 200	0.9274	0.008	0.9627	0.0118	0.895	0.0164	0.9246	0.0086
M.L.P. 50	0.7364	0.0283	0.8088	0.0672	0.69	0.0135	0.7176	0.0186
M.L.P. 100	0.7006	0.0292	0.7154	0.0464	0.7009	0.0227	0.7039	0.0278
A.D. (Mejor)	0.8818	0.0063	0.8699	0.011	0.8945	0.0131	0.8834	0.0064
A.D. (Aleatorio)	0.8853	0.0064	0.8758	0.0169	0.8954	0.0128	0.8866	0.0057
N. Bayes	0.0188	0.0416	0.01	0.0225	0.8775	0.2029	0.5048	0.0112

Tabla 22. Resultados de entrenamiento en escenario 1 con el conjunto de bosque aleatorio.

Conjunto: BOSQUE ALEATORIO								
Alg.	f1		Sensibilidad		Precisión		Exactitud	
	μ	σ	μ	σ	μ	σ	μ	σ
SVM (Lineal)	0.6879	0.3066	0.7502	0.3606	0.662	0.2516	0.7114	0.1371
B.A. 100	0.9311	0.0084	0.9662	0.0128	0.8987	0.0163	0.9284	0.0089
B.A. 200	0.9316	0.008	0.968	0.0124	0.8981	0.016	0.9289	0.0085
M.L.P. 50	0.8091	0.0236	0.8428	0.0528	0.7809	0.0168	0.8024	0.0179
M.L.P. 100	0.8008	0.0237	0.8438	0.0581	0.7628	0.014	0.7907	0.0181
A.D. (Mejor)	0.8889	0.0074	0.8792	0.0167	0.8993	0.0153	0.8902	0.007
A.D. (Aleatorio)	0.8901	0.0061	0.882	0.0155	0.8989	0.0138	0.8912	0.0058
N. Bayes	0.4103	0.0894	0.2901	0.0926	0.7169	0.0498	0.587	0.0424

Tabla 23. Resultados de entrenamiento en escenario 1 con el conjunto de lightGBM.

Conjunto: LIGHTGBM								
Alg.	f1		Sensibilidad		Precisión		Exactitud	
	μ	σ	μ	σ	μ	σ	μ	σ
SVM (Lineal)	0.5136	0.2352	0.6541	0.3295	0.533	0.0932	0.543	0.0602
B.A. 100	0.922	0.0108	0.9473	0.0128	0.8983	0.0177	0.9198	0.0112
B.A. 200	0.9226	0.0109	0.9492	0.0132	0.8977	0.0175	0.9203	0.0113
M.L.P. 50	0.1574	0.287	0.2117	0.4398	0.4698	0.0568	0.4996	0.0034
M.L.P. 100	0.4139	0.2869	0.6014	0.455	0.4749	0.0475	0.4998	0.0048
A.D. (Mejor)	0.8789	0.0075	0.8662	0.0132	0.8922	0.0162	0.8806	0.0076
A.D. (Aleatorio)	0.8786	0.0071	0.8689	0.0148	0.8889	0.0074	0.88	0.0061
N. Bayes	0.0248	0.0378	0.013	0.0204	0.7371	0.1186	0.5053	0.0101

Tabla 24. Resultados de entrenamiento en escenario 1 con el conjunto de todas las características.

Conjunto: TODAS								
Alg.	f1		Sensibilidad		Precisión		Exactitud	
	μ	σ	μ	σ	μ	σ	μ	σ
SVM (Lineal)	0.6345	0.1836	0.6604	0.2576	0.6692	0.0314	0.666	0.0609
B.A. 100	0.9256	0.0095	0.9567	0.0119	0.8968	0.0172	0.9231	0.0101
B.A. 200	0.9266	0.01	0.9589	0.0121	0.8966	0.0176	0.924	0.0106
M.L.P. 50	0.4036	0.2853	0.5967	0.4404	0.4875	0.0473	0.4998	0.0038
M.L.P. 100	0.2129	0.2824	0.3052	0.4334	0.4641	0.0723	0.4991	0.0059
A.D. (Mejor)	0.8818	0.0078	0.8681	0.0142	0.8963	0.0157	0.8837	0.0078
A.D. (Aleatorio)	0.8846	0.0068	0.8734	0.0156	0.8966	0.0148	0.8862	0.0065
N. Bayes	0.0248	0.0378	0.013	0.0204	0.7371	0.1186	0.5053	0.0101

Tabla 25. Resultados de prueba en escenario 1 con el conjunto de correlación.

Correlación					
Alg.	f1	Sensibilidad	Precisión	Exactitud	Minutos
SVM	0.6397	0.7683	0.6376	0.6967	4.3343
B.A. 100	0.7367	0.7718	0.7049	0.7243	3.1998
B.A. 200	0.7379	0.7756	0.7039	0.7246	6.5993
MLP 50	0.7488	0.8217	0.6909	0.7266	8.5842
MLP 100	0.74	0.7778	0.7077	0.7276	16.0947
A.D. (Mejor)	0.6773	0.6661	0.6891	0.6827	0.199
A.D. (Aleatorio)	0.6786	0.666	0.6919	0.6847	0.0417
N. Bayes	0.7211	0.691	0.7543	0.7329	0.0169

Tabla 26. Resultados de prueba en escenario 1 con el conjunto de X^2.

Chi-Cuadrada					
Alg.	f1	Sensibilidad	Precisión	Exactitud	Minutos
SVM	0.7219	0.7928	0.663	0.6948	1.2899
B.A. 100	0.7233	0.7946	0.6641	0.6962	0.5843
B.A. 200	0.7251	0.7984	0.6645	0.6975	1.1547
MLP 50	0.7287	0.7863	0.6811	0.7079	2.2826
MLP 100	0.7212	0.7639	0.6853	0.7057	3.2005
A.D. (Mejor)	0.7249	0.7981	0.6645	0.6974	0.0201
A.D. (Aleatorio)	0.7249	0.7981	0.6645	0.6974	0.0177
N. Bayes	0.7207	0.6931	0.7511	0.7315	0.0161

Tabla 27. Resultados de prueba en escenario 1 con el conjunto de R.F.E.

R.F.E.					
Alg.	f1	Sensibilidad	Precisión	Exactitud	Minutos
SVM	0.7959	0.9996	0.6612	0.7437	1.7268
B.A. 100	0.9109	0.9633	0.8642	0.9057	5.3206
B.A. 200	0.9117	0.9659	0.8636	0.9064	10.6087
MLP 50	0.8527	0.9609	0.7705	0.8348	48.0992
MLP 100	0.883	0.964	0.8194	0.8738	49.0802
A.D. (Mejor)	0.8596	0.8498	0.87	0.8612	0.318
A.D. (Aleatorio)	0.8574	0.8451	0.8707	0.8595	0.0338
N. Bayes	0.2732	0.1877	0.508	0.5014	0.0122

Tabla 28. Resultados de prueba en escenario 1 con el conjunto de regresión logística.

Regresión logística					
Alg.	f1	Sensibilidad	Precisión	Exactitud	Minutos
SVM	0.606	0.6843	0.6201	0.673	7.7057
B.A. 100	0.9218	0.9542	0.8918	0.919	11.5057
B.A. 200	0.9219	0.9551	0.8912	0.919	22.9544
MLP 50	0.7432	0.7837	0.7097	0.731	16.7799
MLP 100	0.7491	0.7846	0.7176	0.7379	20.8437
A.D. (Mejor)	0.8766	0.8647	0.8891	0.8783	1.4459
A.D. (Aleatorio)	0.8784	0.8662	0.8914	0.8801	0.1273
N. Bayes	0.02	0.0106	0.67	0.5044	0.0349

Tabla 29. Resultados de prueba en escenario 1 con el conjunto de bosque aleatorio.

Bosque aleatorio					
Alg.	f1	Sensibilidad	Precisión	Exactitud	Minutos
SVM	0.6265	0.7083	0.5781	0.6871	6.9343
B.A. 100	0.9257	0.9604	0.8936	0.9228	9.1021
B.A. 200	0.9262	0.9615	0.8937	0.9234	18.1304
MLP 50	0.7884	0.8367	0.7474	0.7764	16.0208
MLP 100	0.7946	0.8539	0.7451	0.7804	19.9826
A.D. (Mejor)	0.8835	0.8724	0.8953	0.885	0.993
A.D. (Aleatorio)	0.8841	0.8743	0.8945	0.8854	0.0979
N. Bayes	0.499	0.3801	0.7466	0.6273	0.0289

Tabla 30. Resultados de prueba en escenario 1 con el conjunto de lightGBM.

LightGBM					
Alg.	f1	Sensibilidad	Precisión	Exactitud	Minutos
SVM	0.3971	0.4142	0.514	0.5337	10.4175
B.A. 100	0.9171	0.9395	0.8961	0.9151	19.7842
B.A. 200	0.9172	0.9406	0.8951	0.915	39.6135
MLP 50	0.2255	0.3078	0.4747	0.5002	17.4323
MLP 100	0.4267	0.598	0.4987	0.5015	22.1818
A.D. (Mejor)	0.8745	0.8596	0.8902	0.8766	4.0082
A.D. (Aleatorio)	0.8705	0.8596	0.8819	0.8722	0.3366
N. Bayes	0.0239	0.0125	0.7208	0.5051	0.0591

Tabla 31. Resultados de prueba en escenario 1 con el conjunto de todas las características.

Todas las características					
Alg.	f1	Sensibilidad	Precisión	Exactitud	Minutos
SVM	0.6744	0.8243	0.6186	0.6568	14.8472
B.A. 100	0.9201	0.9491	0.893	0.9175	24.9748
B.A. 200	0.9205	0.9511	0.8921	0.9178	50.2164
MLP 50	0.2285	0.3106	0.4802	0.4989	23.785
MLP 100	0.4759	0.6856	0.5183	0.5024	29.3158
A.D. (Mejor)	0.8783	0.8638	0.8937	0.8803	6.4375
A.D. (Aleatorio)	0.878	0.8649	0.892	0.8799	0.5887
N. Bayes	0.0239	0.0125	0.7208	0.5051	0.1237

Apéndice IV: Escenario 2

En este apartado se encuentran los resultados detallados de los entrenamientos y pruebas del segundo escenario de la fase de experimentación.

Tabla 32. Conjuntos de características generados para el segundo escenario.

Conjunto	Tipo de indicador	Indicador	Versión
Correlación (5)	Tendencia	CCI	1
	Tendencia	Parabolic SAR	1
	Oscilador	RSI	1
	Volatilidad	Donchian Channel	1
X^2 (5)	Volatilidad	Donchian Channel	0, 1, 2, 4 y 5
R.F.E. (38)	Tendencia	MACD	0, 1, 2, 4 y 5
	Tendencia	CCI	1, 3 y 4
	Tendencia	Parabolic SAR	0, 1, 3 y 5
	Oscilador	RSI	0 al 5
	Volatilidad	Bollinger Band	1 y 2
	Volatilidad	Keltner Channel	0, 1, 2 y 5
	Volatilidad	Donchian Channel	0, 1, 2, 4 y 5
	Otros	Daily Return	0
	Otros	Daily Long Return	0
Regresión logística (23)	Tendencia	MACD	1, 2 y 5
	Tendencia	Parabolic SAR	1
	Volatilidad	Bollinger Band	1 y 2
	Volatilidad	Keltner Channel	0, 1, 2 y 5
	Volatilidad	Donchian Channel	0 y 1
	Volumen	Volume-price Trend	0
	Otros	Daily Return	0
	Otros	Daily Long Return	0
Bosque aleatorio (7)	Tendencia	MACD	1, 2 y 5
	Volatilidad	Bollinger Bands	1 y 2
	Volatilidad	Donchian Channel	1
LightGBM (12)	Tendencia	MACD	1 y 2
	Volatilidad	Bollinger Bands	1 y 2
	Volatilidad	Keltner Channel	0, 1 y 2
	Volatilidad	Donchian Channel	1

Tabla 33. Resultados de entrenamiento en escenario 2 con el conjunto de correlación.

Conjunto: CORRELACIÓN										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.9953	0.012863	0.9929	0.021436	0.9979	0.003945	0.9954	0.01248	0.0046	0.01248
B.A. 200	0.9953	0.012529	0.9929	0.020831	0.9979	0.003914	0.9955	0.012169	0.0045	0.012169
A.D. (Mejor)	0.9955	0.011855	0.9932	0.01943	0.9979	0.004006	0.9956	0.011543	0.0044	0.011543
A.D. (Aleatorio)	0.9957	0.014602	0.9938	0.024772	0.9977	0.003978	0.9958	0.014084	0.0042	0.014084
N. Bayes	0.9985	0.001417	1	0.000756	0.997	0.002308	0.9985	0.00142	0.0015	0.00142

Tabla 34. Resultados de entrenamiento en escenario 2 con el conjunto de X^2.

Conjunto: CHI-CUADRADA										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.9906	0.025491	0.9841	0.047375	0.9985	0.001779	0.9913	0.02345	0.0087	0.02345
B.A. 200	0.9906	0.025486	0.9841	0.047375	0.9985	0.001763	0.9913	0.023446	0.0087	0.023446
A.D. (Mejor)	0.9906	0.025486	0.9841	0.047375	0.9985	0.001763	0.9913	0.023446	0.0087	0.023446
A.D. (Aleatorio)	0.9906	0.025486	0.9841	0.047375	0.9985	0.001763	0.9913	0.023446	0.0087	0.023446
N. Bayes	0.9905	0.025797	0.984	0.048021	0.9984	0.001944	0.9912	0.023709	0.0088	0.023709

Tabla 35. Resultados de entrenamiento en escenario 2 con el conjunto de R.F.E.

Conjunto: R.F.E.										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	1	0	1	0	1	0	1	0	0	0
B.A. 200	1	0	1	0	1	0	1	0	0	0
A.D. (Mejor)	1	0	1	0	1	0	1	0	0	0
A.D. (Aleatorio)	1	0.000157	1	0.000227	1	0.000258	1	0.000157	0	0.000157
N. Bayes	0.9905	0.025753	0.9841	0.047946	0.9984	0.001944	0.9912	0.023671	0.0088	0.023671

Tabla 36. Resultados de entrenamiento en escenario 2 con el conjunto de regresión logística.

Conjunto: REGRESIÓN LOGISTICA										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	1	0	1	0	1	0	1	0	0	0
B.A. 200	1	0	1	0	1	0	1	0	0	0
A.D. (Mejor)	1	0	1	0	1	0	1	0	0	0
A.D. (Aleatorio)	1	0	1	0	1	0	1	0	0	0
N. Bayes	0.9904	0.025927	0.9839	0.04824	0.9984	0.001944	0.9911	0.023818	0.0089	0.023818

Tabla 37. Resultados de entrenamiento en escenario 2 con el conjunto de bosque aleatorio.

Conjunto: BOSQUE ALEATORIO										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	1	0	1	0	1	0	1	0	0	0
B.A. 200	1	0	1	0	1	0	1	0	0	0
A.D. (Mejor)	1	0	1	0	1	0	1	0	0	0
A.D. (Aleatorio)	1	0	1	0	1	0	1	0	0	0
N. Bayes	1	0	1	0	1	0	1	0	0	0

Tabla 38. Resultados de entrenamiento en escenario 2 con el conjunto de lightGBM.

Conjunto: LIGHTGBM										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	1	0	1	0	1	0	1	0	0	0
B.A. 200	1	0	1	0	1	0	1	0	0	0
A.D. (Mejor)	1	0	1	0	1	0	1	0	0	0
A.D. (Aleatorio)	1	0	1	0	1	0	1	0	0	0
N. Bayes	0.9985	0.002242	1	0	0.997	0.004451	0.9985	0.00226	0.0015	0.00226

Tabla 39. Resultados de entrenamiento en escenario 2 con el conjunto de todas las características.

Conjunto: TODAS										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	1	0	1	0	1	0	1	0	0	0
B.A. 200	1	0	1	0	1	0	1	0	0	0
A.D. (Mejor)	1	0	1	0	1	0	1	0	0	0
A.D. (Aleatorio)	1	7.56E-05	1	0.000151	1	0	1	7.56E-05	0	7.56E-05
N. Bayes	0.9905	0.024854	0.9841	0.046662	0.9982	0.001947	0.9912	0.022881	0.0088	0.022881

Tabla 40. Resultados de la primera prueba en escenario 2 con el conjunto de correlación.

Correlación						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.995033068	0.992714898	0.997527822	0.995160071	0.004839929	0.270365973
B.A. 200	0.995116168	0.992916562	0.997480403	0.995235694	0.004764306	0.523188754
A.D. (Mejor)	0.995384174	0.993244265	0.997661344	0.995487774	0.004512226	0.007964428
A.D. (Aleatorio)	0.994358005	0.991630955	0.997338097	0.994529871	0.005470129	0.00272543
N. Bayes	0.998578614	0.99974792	0.997413937	0.99857575	0.00142425	0.003412096

Tabla 41. Resultados de la primera prueba en escenario 2 con el conjunto de X^2.

Chi-Cuadrada						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.990472934	0.983816486	0.998492467	0.991152004	0.008847996	0.118890011
B.A. 200	0.990460335	0.983816486	0.99846726	0.9911394	0.0088606	0.234198101
A.D. (Mejor)	0.990460335	0.983816486	0.99846726	0.9911394	0.0088606	0.002920262
A.D. (Aleatorio)	0.990460335	0.983816486	0.99846726	0.9911394	0.0088606	0.00272071
N. Bayes	0.990344143	0.983992942	0.998091648	0.991038568	0.008961432	0.003403123

Tabla 42. Resultados de la primera prueba en escenario 2 con el conjunto de R.F.E.

R.F.E.						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	1	1	1	1	0	0.485184956
B.A. 200	1	1	1	1	0	0.983155564
A.D. (Mejor)	1	1	1	1	0	0.037851457
A.D. (Aleatorio)	0.999899174	0.999924376	0.999874043	0.999899168	0.000100832	0.008656394
N. Bayes	0.990359033	0.98401815	0.998091648	0.991051172	0.008948828	0.012201405

Tabla 43. Resultados de la primera prueba en escenario 2 con el conjunto de regresión logística.

Regresión logística						
Alg.	**f1**	**Sensiblidad**	**Precisión**	**Exactitud**	**P. Hamming**	**Minutos**
B.A. 100	1	1	1	1	0	0.380397896
B.A. 200	1	1	1	1	0	0.746806669
A.D. (Mejor)	1	1	1	1	0	0.026281734
A.D. (Aleatorio)	1	1	1	1	0	0.005252039
N. Bayes	0.990286842	0.98389211	0.998091629	0.990988152	0.009011848	0.008183149

Tabla 44. Resultados de la primera prueba en escenario 2 con el conjunto de bosque aleatorio.

Bosque aleatorio						
Alg.	**f1**	**Sensiblidad**	**Precisión**	**Exactitud**	**P. Hamming**	**Minutos**
B.A. 100	1	1	1	1	0	0.254747049
B.A. 200	1	1	1	1	0	0.508804349
A.D. (Mejor)	1	1	1	1	0	0.011168031
A.D. (Aleatorio)	1	1	1	1	0	0.003535553
N. Bayes	1	1	1	1	0	0.004043885

Tabla 45. Resultados de la primera prueba en escenario 2 con el conjunto de lightGBM.

LightGBM						
Alg.	**f1**	**Sensiblidad**	**Precisión**	**Exactitud**	**P. Hamming**	**Minutos**
B.A. 100	1	1	1	1	0	0.319177159
B.A. 200	1	1	1	1	0	0.630821323
A.D. (Mejor)	1	1	1	1	0	0.016362035
A.D. (Aleatorio)	1	1	1	1	0	0.00378406
N. Bayes	0.998256169	1	0.996528372	0.998248046	0.001751954	0.004930949

Tabla 46. Resultados de la primera prueba en escenario 2 con el conjunto de todas las características.

Todas las características						
Alg.	**f1**	**Sensiblidad**	**Precisión**	**Exactitud**	**P. Hamming**	**Minutos**
B.A. 100	1	1	1	1	0	0.496263723
B.A. 200	1	1	1	1	0	0.975209355
A.D. (Mejor)	1	1	1	1	0	0.042506421
A.D. (Aleatorio)	0.999974786	0.999949584	1	0.999974792	2.5208E-05	0.01003956
N. Bayes	0.98990499	0.984194605	0.996932049	0.990559617	0.009440383	0.013602626

Tabla 47. Resultados de la segunda prueba en escenario 2 con el conjunto de correlación.

Correlación						
Alg.	**f1**	**Sensiblidad**	**Precisión**	**Exactitud**	**P. Hamming**	**Minutos**
B.A. 100	0.995	0.9927	0.9975	0.9952	0.0048	0.2704
B.A. 200	0.9951	0.9929	0.9975	0.9952	0.0048	0.5232
A.D. (Mejor)	0.9954	0.9932	0.9977	0.9955	0.0045	0.008
A.D. (Aleatorio)	0.9944	0.9916	0.9973	0.9945	0.0055	0.0027
N. Bayes	0.9986	0.9997	0.9974	0.9986	0.0014	0.0034

Tabla 48. Resultados de la segunda prueba en escenario 2 con el conjunto de X^2.

Chi-Cuadrada						
Alg.	**f1**	**Sensiblidad**	**Precisión**	**Exactitud**	**P. Hamming**	**Minutos**
B.A. 100	0.9905	0.9838	0.9985	0.9912	0.0088	0.1189
B.A. 200	0.9905	0.9838	0.9985	0.9911	0.0089	0.2342
A.D. (Mejor)	0.9905	0.9838	0.9985	0.9911	0.0089	0.0029
A.D. (Aleatorio)	0.9905	0.9838	0.9985	0.9911	0.0089	0.0027
N. Bayes	0.9903	0.984	0.9981	0.991	0.009	0.0034

Tabla 49. Resultados de la segunda prueba en escenario 2 con el conjunto de R.F.E.

R.F.E.						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	1	1	1	1	0	0.4852
B.A. 200	1	1	1	1	0	0.9832
A.D. (Mejor)	1	1	1	1	0	0.0379
A.D. (Aleatorio)	0.9999	0.9999	0.9999	0.9999	0.0001	0.0087
N. Bayes	0.9904	0.984	0.9981	0.9911	0.0089	0.0122

Tabla 50. Resultados de la segunda prueba en escenario 2 con el conjunto de regresión logística.

Regresión logística						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	1	1	1	1	0	0.3804
B.A. 200	1	1	1	1	0	0.7468
A.D. (Mejor)	1	1	1	1	0	0.0263
A.D. (Aleatorio)	1	1	1	1	0	0.0053
N. Bayes	0.9903	0.9839	0.9981	0.991	0.009	0.0082

Tabla 51. Resultados de la segunda prueba en escenario 2 con el conjunto de bosque aleatorio.

Bosque aleatorio						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	1	1	1	1	0	0.2547
B.A. 200	1	1	1	1	0	0.5088
A.D. (Mejor)	1	1	1	1	0	0.0112
A.D. (Aleatorio)	1	1	1	1	0	0.0035
N. Bayes	1	1	1	1	0	0.004

Tabla 52. Resultados de la segunda prueba en escenario 2 con el conjunto de lightGBM.

LightGBM						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	1	1	1	1	0	0.3192
B.A. 200	1	1	1	1	0	0.6308
A.D. (Mejor)	1	1	1	1	0	0.0164
A.D. (Aleatorio)	1	1	1	1	0	0.0038
N. Bayes	0.9983	1	0.9965	0.9982	0.0018	0.0049

Tabla 53. Resultados de la segunda prueba en escenario 2 con el conjunto de todas las características.

Todas las características						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	1	1	1	1	0	0.4963
B.A. 200	1	1	1	1	0	0.9752
A.D. (Mejor)	1	1	1	1	0	0.0425
A.D. (Aleatorio)	1	0.9999	1	1	0	0.01
N. Bayes	0.9899	0.9842	0.9969	0.9906	0.0094	0.0136

Apéndice V: Escenario 3

En este apartado se encuentran los resultados detallados de los entrenamientos y pruebas del tercer escenario de la fase de experimentación.

Tabla 54. Conjuntos de características generados para el tercer escenario.

Conjunto	Tipo de indicador	Indicador	Versión
Correlación (21)	Tendencia	CCI	0 al 5
	Tendencia	Parabolic SAR	1 y 2
	Tendencia	RSI	0 al 5
	Volatilidad	Bollinger Bands	5
	Volatilidad	Donchian Channel	0 al 5
X^2 (21)	Tendencia	Parabolic SAR	1 y 2
	Volatilidad	Bollinger Bands	0, 4 y 5
	Volatilidad	Donchian Channel	0, 1, 3, 4 y 5
R.F.E. (32)	Tendencia	MACD	1 y 2
	Tendencia	Mass Index	0
	Tendencia	DPO	0, 3 y 4
	Oscilador	RSI	3 y 4
	Oscilador	KAMA	0
	Volatilidad	Bollinger Bands	1 y 2
	Volatilidad	Dochian Channel	1 y 2
	Volatilidad	Keltner Channel	0, 1, 2, y 3
	Volumen	ADI	0
	Volumen	Ease of Movement	0
	Volumen	Volume-price Trend	0
Regresión logística (108)	Tendencia	MACD	0
	Tendencia	Mass Index	0
	Tendencia	KST	0
	Tendencia	Ichimoku	0
	Tendencia	Trix	0, 2, 3, 4 y 5
	Tendencia	DPO	0, 1, 3, 4 y 5
	Tendencia	Parabolic SAR	0, 1 y 2
	Oscilador	RSI	0 al 5
	Oscilador	True Strenght Idex	0
	Oscilador	Stochastic Oscillator	0
	Oscilador	Williams %R	0
	Oscilador	Awesome Oscillator	0

Conjunto	Tipo de indicador	Indicador	Versión
	Oscilador	KAMA	0
	Volatilidad	Bollinger Bands	0 al 5
	Volatilidad	Dochian Channel	0 al 5
	Volatilidad	Keltner Channel	0 al 5
	Volumen	ADI	0
	Volumen	On-Balance Volume	0
	Volumen	Force Index	0
	Volumen	Negative Volume Index	0
	Volumen	Volume-price Trend	0
Bosque aleatorio (29)	Tendencia	MACD	1 y 2
	Tendencia	CCI	1, 2 y 4
	Tendencia	Parabolic SAR	1 y 3
	Oscilador	RSI	0, 1, 2 y 4
	Oscilador	Awesome Oscillator	0
	Volatilidad	Bollinger Bands	1 y 2
	Volatilidad	Keltner Channel	0, 1 y 2
	Volatilidad	Donchian Channel	0, 1, 3 y 4
	Otros	Daily Log Return	0
LightGBM (100)	Tendencia	MACD	0
	Tendencia	Mass Index	0
	Tendencia	KST Oscillator	0
	Tendencia	Ichimoku	0
	Tendencia	ADX	0 al 5
	Tendencia	Trix	0 al 5
	Tendencia	CCI	0 al 5
	Tendencia	DPO	0 al 5
	Tendencia	Parabolic SAR	0, 2, 3, 4 y 5
	Oscilador	True Strenght Index	0
	Oscilador	Ultimate Oscillator	0
	Oscilador	Stochastic Oscillator	0
	Oscilador	Williams %R	0
	Oscilador	Awesome Oscillator	0
	Oscilador	Money Flow Index	0 al 5
	Oscilador	RSI	0 al 5
	Volatilidad	Bollinger Bands	1
	Volatilidad	Donchian Channel	1 y 5
	Volatilidad	ATR	0 al 5

Conjunto	Tipo de indicador	Indicador	Versión
	Volumen	ADI	0
	Volumen	On-Balance Volume	0
	Volumen	Chaikin Money Flow	0
	Volumen	Force Index	0
	Volumen	Volume-price Trend	0
	Volumen	Negative Volume Index	0
	Volumen	Ease of Movement	0 al 5
	Otros	Daily Return	0
	Otros	Daily Log Return	0
	Otros	Cumulative Return	0

Tabla 55. Resultados de entrenamiento en escenario 3 con el conjunto de correlación.

Conjunto: CORRELACIÓN										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.8128	0.0042	0.8222	0.0046	0.8241	0.0054	0.8222	0.0046	0.1778	0.0046
B.A. 200	0.8136	0.0034	0.8233	0.0038	0.826	0.0048	0.8233	0.0038	0.1767	0.0038
A.D. (Mejor)	0.7479	0.0103	0.7476	0.0114	0.7486	0.0078	0.7476	0.0114	0.2524	0.0114
A.D. (Aleatorio)	0.7467	0.0083	0.7463	0.0094	0.7476	0.0062	0.7463	0.0094	0.2537	0.0094
N. Bayes	0.7187	0.0165	0.7404	0.0188	0.7411	0.0202	0.7404	0.0188	0.2596	0.0188

Tabla 56. Resultados de entrenamiento en escenario 3 con el conjunto de X^2.

Conjunto: CHI-CUADRADA										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.8088	0.0035	0.8276	0.0033	0.8618	0.0031	0.8276	0.0033	0.1724	0.0033
B.A. 200	0.8088	0.0035	0.8276	0.0033	0.8619	0.0032	0.8276	0.0033	0.1724	0.0033
A.D. (Mejor)	0.8087	0.0036	0.8274	0.0033	0.8613	0.0034	0.8274	0.0033	0.1726	0.0033
A.D. (Aleatorio)	0.8087	0.0036	0.8275	0.0034	0.8613	0.0034	0.8275	0.0034	0.1725	0.0034
N. Bayes	0.8013	0.0111	0.8208	0.0132	0.8506	0.0247	0.8208	0.0132	0.1792	0.0132

Tabla 57. Resultados de entrenamiento en escenario 3 con el conjunto de R.F.E.

Conjunto: R.F.E.										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.9461	0.004	0.9471	0.0037	0.9494	0.0029	0.9471	0.0037	0.0529	0.0037
B.A. 200	0.9465	0.0042	0.9475	0.0039	0.9499	0.003	0.9475	0.0039	0.0525	0.0039
A.D. (Mejor)	0.913	0.0033	0.9128	0.0032	0.9138	0.004	0.9128	0.0032	0.0872	0.0032
A.D. (Aleatorio)	0.913	0.0038	0.9128	0.004	0.914	0.0036	0.9128	0.004	0.0872	0.004
N. Bayes	0.169	0.002	0.3337	0.0006	0.3689	0.042	0.3337	0.0006	0.6663	0.0006

Tabla 58. Resultados de entrenamiento en escenario 3 con el conjunto de regresión logística.

Conjunto: REGRESIÓN LOGISTICA										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.949	0.0054	0.9499	0.0051	0.9523	0.0042	0.9499	0.0051	0.0501	0.0051
B.A. 200	0.9491	0.0058	0.95	0.0055	0.9524	0.0046	0.95	0.0055	0.05	0.0055
A.D. (Mejor)	0.9164	0.0028	0.9162	0.0031	0.9173	0.0021	0.9162	0.0031	0.0838	0.0031
A.D. (Aleatorio)	0.9187	0.0044	0.9186	0.0048	0.9196	0.0031	0.9186	0.0048	0.0814	0.0048
N. Bayes	0.1726	0.0238	0.3346	0.0118	0.3666	0.0601	0.3346	0.0118	0.6654	0.0118

Tabla 59. Resultados de entrenamiento en escenario 3 con el conjunto de bosque aleatorio.

Conjunto: BOSQUE ALEATORIO										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.9477	0.0056	0.9486	0.0052	0.951	0.0044	0.9486	0.0052	0.0514	0.0052
B.A. 200	0.9478	0.0052	0.9488	0.0049	0.9512	0.004	0.9488	0.0049	0.0512	0.0049
A.D. (Mejor)	0.9163	0.0041	0.9161	0.0045	0.9172	0.0027	0.9161	0.0045	0.0839	0.0045
A.D. (Aleatorio)	0.9151	0.0033	0.9149	0.0034	0.9161	0.0032	0.9149	0.0034	0.0851	0.0034
N. Bayes	0.8465	0.0154	0.8558	0.0166	0.869	0.0226	0.8558	0.0166	0.1442	0.0166

Tabla 60. Resultados de entrenamiento en escenario 3 con el conjunto de lightGBM.

Conjunto: LIGHTGBM										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.9441	0.0047	0.9449	0.0044	0.9465	0.0039	0.9449	0.0044	0.0551	0.0044
B.A. 200	0.9444	0.0044	0.9452	0.0041	0.9469	0.0036	0.9452	0.0041	0.0548	0.0041
A.D. (Mejor)	0.9072	0.0029	0.907	0.0031	0.9079	0.0026	0.907	0.0031	0.093	0.0031
A.D. (Aleatorio)	0.9032	0.0091	0.903	0.0096	0.904	0.0066	0.903	0.0096	0.097	0.0096
N. Bayes	0.1718	0.0045	0.3345	0.0017	0.3672	0.0918	0.3345	0.0017	0.6655	0.0017

Tabla 61. Resultados de entrenamiento en escenario 3 con el conjunto de todas las características.

Conjunto: TODAS										
Alg.	f1		Sensibilidad		Precisión		Exactitud		Pérdida de Hamm.	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
B.A. 100	0.9487	0.0044	0.9496	0.0041	0.9519	0.0033	0.9496	0.0041	0.0504	0.0041
B.A. 200	0.9489	0.0049	0.9498	0.0046	0.9522	0.0037	0.9498	0.0046	0.0502	0.0046
A.D. (Mejor)	0.9155	0.0044	0.9153	0.0048	0.9162	0.0033	0.9153	0.0048	0.0847	0.0048
A.D. (Aleatorio)	0.9164	0.0049	0.9163	0.0052	0.9173	0.0038	0.9163	0.0052	0.0837	0.0052
N. Bayes	0.1718	0.0045	0.3345	0.0017	0.3672	0.0918	0.3345	0.0017	0.6655	0.0017

Tabla 62. Resultados de pruebas en escenario 3 con el conjunto de correlación.

Correlación						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.8063	0.8154	0.8155	0.8154	0.1846	4.7343
B.A. 200	0.8067	0.8161	0.8169	0.8161	0.1839	9.4639
A.D. (Mejor)	0.7409	0.7403	0.7421	0.7403	0.2597	0.3766
A.D. (Aleatorio)	0.7405	0.7399	0.7417	0.7399	0.2601	0.0433
N. Bayes	0.7233	0.7443	0.7463	0.7443	0.2557	0.0117

Tabla 63. Resultados de pruebas en escenario 3 con el conjunto de X^2.

Chi-Cuadrada						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.8083	0.8274	0.8617	0.8274	0.1726	0.4636
B.A. 200	0.8082	0.8274	0.8617	0.8274	0.1726	0.9144
A.D. (Mejor)	0.8079	0.827	0.8607	0.827	0.173	0.0168
A.D. (Aleatorio)	0.8079	0.827	0.8606	0.827	0.173	0.0138
N. Bayes	0.8004	0.8201	0.8503	0.8201	0.1799	0.0117

Tabla 64. Resultados de pruebas en escenario 3 con el conjunto de R.F.E.

R.F.E.						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.9434	0.9444	0.9467	0.9444	0.0556	8.1317
B.A. 200	0.9431	0.9441	0.9465	0.9441	0.0559	16.4889
A.D. (Mejor)	0.9085	0.9082	0.9092	0.9082	0.0918	0.8433
A.D. (Aleatorio)	0.909	0.9088	0.9099	0.9088	0.0912	0.0548
N. Bayes	0.1712	0.3335	0.3563	0.3335	0.6665	0.016

Tabla 65. Resultados de pruebas en escenario 3 con el conjunto de regresión logística.

Regresión logística						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.9454	0.9464	0.9487	0.9464	0.0536	16.631
B.A. 200	0.9457	0.9467	0.9492	0.9467	0.0533	33.2513
A.D. (Mejor)	0.9118	0.9116	0.9127	0.9116	0.0884	3.3267
A.D. (Aleatorio)	0.9126	0.9124	0.9135	0.9124	0.0876	0.2265
N. Bayes	0.1781	0.3379	0.3575	0.3379	0.6621	0.0509

Tabla 66. Resultados de pruebas en escenario 3 con el conjunto de bosque aleatorio.

Bosque aleatorio						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.9446	0.9456	0.948	0.9456	0.0544	6.5498
B.A. 200	0.9447	0.9457	0.9483	0.9457	0.0543	13.0892
A.D. (Mejor)	0.9124	0.9122	0.9133	0.9122	0.0878	0.6139
A.D. (Aleatorio)	0.911	0.9107	0.9119	0.9107	0.0893	0.0458
N. Bayes	0.8462	0.8555	0.869	0.8555	0.1445	0.015

Tabla 67. Resultados de pruebas en escenario 3 con el conjunto de lightGBM.

LightGBM						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.9409	0.9418	0.9432	0.9418	0.0582	17.9697
B.A. 200	0.941	0.9419	0.9434	0.9419	0.0581	35.9633
A.D. (Mejor)	0.9037	0.9035	0.9045	0.9035	0.0965	3.6504
A.D. (Aleatorio)	0.8955	0.8952	0.8966	0.8952	0.1048	0.255
N. Bayes	0.1742	0.3352	0.3586	0.3352	0.6648	0.0541

Tabla 68. Resultados de pruebas en escenario 3 con el conjunto de todas las características.

Todas las características						
Alg.	f1	Sensiblidad	Precisión	Exactitud	P. Hamming	Minutos
B.A. 100	0.9461	0.9471	0.9495	0.9471	0.0529	21.4153
B.A. 200	0.946	0.947	0.9494	0.947	0.053	42.7351
A.D. (Mejor)	0.9096	0.9093	0.9106	0.9093	0.0907	6.5074
A.D. (Aleatorio)	0.9105	0.9103	0.9115	0.9103	0.0897	0.4902
N. Bayes	0.1742	0.3352	0.3586	0.3352	0.6648	0.098