Universidad Central de Venezuela

Facultad de Ciencias Económicas y Sociales
Escuela de Estadística y Ciencias Actuariales

Modelado de los retornos de la inversión de una estrategia automatizada en el mércado bursátil

Trabajo Final de Pregrado

PARA OPTAR POR EL TÍTULO DE:

Licenciado en Ciencias Actuariales

Rodrigo Alejandro Serrano Morales

TUTOR:
Prof. Jonattan Ramos & Prof. Eloy Eligon

Caracas, Marzo 2019

Índice general

Índice de figuras 2

Índice de tablas 3

Introduccion 4

1. El Problema 5

1.1. Planteamiento del Problema 5

1.2. Objetivo General 6

1.3. Objetivos Específicos 7

1.4. Justificación 7

2. Marco Teórico 8

2.1. Antecedentes 8

2.1.1. Trading de Cryptomonedas basado en Aprendizaje Automatico 8

2.1.2. Un enfoque de aprendizaje automático para el comercio automatizado . . 8

2.1.3. Modelos predictivos para el mercado FOREX 8

2.1.4. Un Análisis de estrategias de trading técnico 9

2.1.5. Modelos Ocultos de Markov Aplicados al Reconocimiento de Patrones del Análisis Técnico Bursátil 9

2.2. Bases Teóricas 9

2.2.1. Hipótesis del Mercado Eficiente 9

2.2.2. Análisis Técnico 10
2.2.3. Introducción al aprendizaje automático 12
2.2.4. Validación Cruzada en Series de Tiempo 13
2.2.5. Regresión Logística 13
2.2.6. Análisis de Componentes Principales 15
2.2.7. Matríz de Confusión 16
2.2.8. Medidas de Riesgo 17
2.3. Bases Legales 19
2.3.1. Superintendencia Nacional de Valores 19
2.3.2. El uso del VaR para la Regulación de Entidades Financieras 19

3. Marco Metódico **20**
3.1. Tipo de Investigación 20
3.2. Universo y Muestra 20
3.2.1. Tipo de Muestreo 21
3.3. Fuentes de Datos 21
3.4. Variables 21
3.4.1. Variable Dependiente 21
3.4.2. Variables Predictoras 22
3.5. Estrategia de Análisis 23
3.5.1. WalkForward Backtesting 23
3.5.2. Reducción de la dimensión con ACP 25
3.5.3. Modelado de los Retornos 26

4. Análisis de Resultados **27**
4.1. Análisis Exploratorio de los datos 27
4.1.1. Análisis de los Índices 27
4.1.2. Análisis de las Variables Predictoras 29

4.1.3. Resultados del ACP 33
4.2. Coeficientes del modelo 39
4.3. Resultados de la simulación 41
4.3.1. Medidas de Riesgo 44

Conclusiones y Recomendaciones **45**

5. Anexos **47**
5.1. Coeficientes de los modelos 47

Lista de Referencias **55**

Índice de figuras

2.1. Matriz de Confusión 17

2.2. Función de densidad con VaR y ES 18

3.1. Metodología WalkForward 24

3.2. Observaciones que presentan ocurrencias con sl = 2.5 %, en el primer conjunto de datos de entrenamiento (26/10/2008 - 31/12/2010) para el índice S&P500, para diferentes valores de tp y h 25

4.1. Precios de Cierre de los índices en el período de estudio (26/10/2008 - 18/01/2019) 28

4.2. Correlación entre indicadores originales calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012) 29

4.3. Correlación entre indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012) 30

4.4. Gráfico de cajas de indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012) 31

4.5. Histogramas de frecuencia de indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012) . . 32

4.6. Gráfico de densidad de indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012) 33

4.7. Eigenvalores y Porcentaje de contribución para los 10 componentes más importantes obtenidos por la matríz de datos 34

4.8. Contribución de cada variable para PC1, PC2 y el total de la contribución en ambos componentes 35

4.9. Calidad de representación medida por Cos^2 de cada variable en PC1 y PC2 . . . 36

4.10. Gráfico de Correlación entre PC1 y PC2 37

4.11. Gráfico Dispersión entre los componentes. Los puntos rojos representan las observaciones marcadas como 'buys', los triángulos azules los 'stays' 38

4.12. Clasificación de los trades 42

4.13. Retorno acumulado para cada índice 43

4.14. Gráfico de Disperción de los p-valores vs período de entrenamiento 45

Índice de tablas

4.1. Resumen del modelo para cada período de entrenamiento utilizando S&P500 . . 40
4.2. Resumen de resultados de aplicar el modelo en la data de prueba para los 5 índices 41
4.3. Resultados del test de Wald–Wolfowitz (Test de Racha) 44
4.4. VaR y ES para retornos de cada índice 44

5.1. Resumen del modelo para cada período de entrenamiento utilizando NASDAQ . 47
5.2. Resumen del modelo para cada período de entrenamiento utilizando NIKKEI 225 49
5.3. Resumen del modelo para cada período de entrenamiento utilizando FTSE 100 . 51
5.4. Resumen del modelo para cada período de entrenamiento utilizando BOVESPA . 53

Introducción

En la presente investigación se busca utilizar las técnicas de aprendizaje estadístico para realizar predicciones sobre el comportamiento de un activo financiero en el corto plazo, intentando capitalizar este conocimiento. Ahora bien, existe un debate sobre si las inversiones activas, son realmente efectivas. La principal razón de la polémica, es debido a la complejidad que involucra realizar una predicción con tan poco margen de error y durante un período de tiempo sostenido. Aunado de las múltiples fuentes de riesgo concernientes a los portafolios de inversión, sería impensable tratar de definir en una primera investigación una estrategia eficaz para el desarrollo del trading.

En las investigaciones similares a la presente, se toman algunos métodos y se intenta definir cuáles de ellos arroja mayor precisión. Por otro lado el explosivo crecimiento en los últimos años de los métodos de aprendizaje estadístico ha creado innumerables variantes de los algoritmos, por lo que abarcar varios de estos supondría un trabajo arduo, realizado probablemente por un equipo de trabajo con experiencia, tanto en el mundo financiero como computacional y estadístico. En esta investigación se abordará el aprendizaje supervisado, basado en un modelo lineal generalizado -MLG-, específicamente, una regresión logística. Se intenta contrarrestar la debilidad de asumir linealidad en el modelo, agrupando las variables con una técnica no supervisada como el Análisis de Componentes Principales.

La idea del modelo es predecir los casos en los cuales, el incurrir en una posición en largo provocaría una ganancia, asegurando que la posición se capitalice a un porcentaje t del precio inicial, sin antes haber incurrido en una pérdida de s porciento, en un horizonte de tiempo h. Para esto se identifica en un período de la data utilizada, todas las observaciones en la que esto ocurre, y se intenta generar un modelo que entienda las condiciones del activo cuando este patrón sucede.

Capítulo 1

El Problema

1.1. Planteamiento del Problema

A principios del siglo XX los mercados bursátiles eran operados manualmente, los corredores y demás participantes realizaban una serie de labores sin asistencia de sistemas informáticos, que evidentemente no existían para la época. El 8 de febrero de 1971 comienza operaciones la bolsa de valores NASDAQ -National Association of Securities Dealers Automated Quotation-, la cual marca un precedente en la industria. NASDAQ fue el primer mercado bursátil en el mundo en adoptar el uso de sistemas electrónicos para proveer cotizaciones de las acciones. Más tarde añadirían, también, la capacidad de manejar operaciones electrónicamente, ejecutando órdenes de compra y venta sin necesidad de encontrarse físicamente en el lugar destinado para las transacciones.

Este precedente aunado con el rápido crecimiento de la industria tecnológica y la creación del internet, facilitan al mercado bursátil digitalizarse rápidamente. Esto permite a los competidores y la industria en general obtener un sistema más eficiente en términos de operatividad, además de globalizar la actividad financiera y propiciar la participación de más inversores en los mercados. El globalizar la información incrementa la competitividad en el sector e incentiva estudios de privados e independientes para maximizar retornos, de donde surgen los análisis técnicos y fundamentales.

Ahora bien la iniciación al mundo bursátil trae consigo grandes riesgos si no se opera con conocimiento y planificación financiera. Es por esto que cada año los individuos pierden dinero al querer incursionar en el mundo del trading sin la preparación adecuada, lo que por otro lado, le da oportunidad a participantes ya establecidos y con conocimiento de aprovechar esta dinámica. El trading requiere de una planificación exhaustiva en cuanto al manejo de riesgo y aplicación de estrategias, es así como los algoritmos automáticos suponen un manejo ideal de los factores, estableciendo reglas sistemáticas para las operaciones.

En los últimos años y con el impulso de computadores cada vez más potentes, ha surgido un nicho dentro del mercado financiero destinado a compañías que operan con algoritmos automatizados, llamadas comúnmente “Quantitave Funds” -Fondos Cuantitativos-. Según un informe realizado por Credit Suisse Group, estos fondos son los de mayor crecimientos en los últimos años. Los fondos cuantitativos y de inversión pasiva controlan actualmente el 60 % de

los activos del mercado, incluso, tan solo el 10 % del volumen de transacciones proviene de inversores discrecionales, según JPMorgan Chase & Co. en cita de Bloomberg. De esta manera compañías como Goldman Sachs han reemplazado gran cantidad de operadores por algoritmos, como dio a conocer en el 2017. La mayoría de las grandes instituciones y proveedores del sector ya cuentan con servicios que cumplen con los requerimientos para la aplicación de los sistemas automatizados, como plataformas con puertos API y datas históricas destinadas a testear algoritmos de trading. En palabras de Frank Thermitus (2018), presidente de la Asociación de Trading Algorítmico de Argentina "No tenemos datos oficiales en Argentina, pero sabemos que la aparición de soporte para este tipo de trading está llamando a un crecimiento de la participación de agentes automatizados".

Con este panorama es evidente la utilidad que supone la aplicación de métodos de aprendizaje estadístico para el desarrollo de algoritmos automáticos. El aprendizaje estadístico, referido en la literatura anglosajona como "Machine Learning" es definido por uno de sus principales exponentes -el profesor de la universidad de Carnegie Mellon- Tom Mitchell como "el campo de estudio que da a las computadoras la habilidad de aprender sin ser explícitamente programada". En este sentido, expone que "Un algoritmo aprende de la experiencia E con respecto a una tarea T cuyo rendimiento es medido por el indicador P, si su desempeño en T como medida de P mejora con la experiencia E". Esta referencia a que la "computadora aprende", no son más que técnicas estadísticas aplicadas a distintos problemas, en donde se busca mejorar el desempeño del modelo calibrando una medida de error.

El desarrollo tecnológico, especialmente el potenciamiento de los procesadores de información y el surgimiento del software libre, han provocado un crecimiento en el número de profesionales que aplican el "Machine Learning". Esto ha ocasionado que innumerables problemas sean abordados con estas técnicas, desde transformar historiales de compras de consumidores en enfoques publicitario, ó intentar imitar el razonamiento humano en una tarea en específico, como jugar ajedrez.

Una de las principales críticas al trading discresional se debe a la complejidad de entender los movimientos del activo. Los patrones gráficos basados en indicadores técnicos buscan exponer el comportamiento del precio. La ídea de esta investigación es comprender si un modelo de aprendizaje estadístico puede, en base a estos indicadores, pronósticar el momento correcto de entrada para obtener una ganancia.

En base a la problemática, surge la interrogante: ¿El enfoque de desarrollar un estrategia de trading automatizado basada en aprendizaje estadístico es eficaz para pronósticar ganancias?

1.2. Objetivo General

Modelar los retornos de la inversión de una estrategia automatizada para pronósticar movimientos en el mercado búrsatil

1.3. Objetivos Específicos

- Identificar la variable dicotómica la cual señala la materialización de un precio objetivo o no
- Calcular las variables predictoras, en este caso, los indicadores técnicos que luego serán reconstruidos por el Análisis de Componentes Principales
- Desarrollar el marco de backtesting con la metodología WalkForward
- Simular las compras y ventas en función a la predicción del modelo
- Calcular el Valor a Riesgo y Pérdida Esperada para cada índice en base a los resultados de la simulación

1.4. Justificación

La especulación en el mercado bursátil contribuye a que los fondos de inversión controlen adecuadamente los riesgos financieros y a su vez diversifiquen portafolios. Estos algoritmos contribuyen también a la formación correcta del precio de los activos, que de no existir especuladores estarían en manos de monopolios, además generan liquidez al mercado y permiten a más participantes beneficiarse de las ganancias de grandes corporaciones.

Los beneficios del trading discrecional se ven aumentados cuando la estrategia se basa en algoritmos ya que realiza la labor de inversión de manera sistemática, removiendo el factor emocional de las personas. Apegarse a un estrategia de inversión es vital para lograr resultados consistentes en el trading. Sin embargo, es difícil encontrar estrategias que generen rendimientos positivos todo el tiempo, y aún así los factores emocionales de las personas pueden destruir cualquier plan al enfrentarse a momentos transitorios de pérdida continua, llamados 'drawdowns' en el mundo financiero. El necesitar un plan consistente y sistemático, promueve la utilización de sistemas automatizados dado que elimina cualquier aspecto emocional durante el proceso de la inversión, manteniendo la disciplina durante momentos de alta volatilidad.

La otra gran ventaja de los sistemas automatizados es que permite la posibilidad de realizar pruebas del algoritmo en data histórica para entender el comportamiento de la estrategia, este proceso es conocido como 'backtesting'. De esta manera el inversor tiene una idea de cómo se comportará el sistema en situaciones similares a las testeadas en el backtesting. El backtesting también permite optimizar los parámetros de la estrategia en busca de maximizar los retornos con el menor riesgo, sin embargo, se debe tener cuidado en no sobre optimizar el sistema.

Capítulo 2

Marco Teórico

2.1. Antecedentes

2.1.1. Trading de Cryptomonedas basado en Aprendizaje Automatico

Bach y Nielsen examinan la efectividad de diversos algoritmos de aprendizaje estadístico para operar en el mercado de cryptomonedas. Establecen un marco de trabajo en lenguaje R que les permite probar los distintos algoritmos cambiando las variables predictoras. Utilizan distintos intervalos de tiempo desde 1 minuto hasta 24 horas, diversos indicadores como ADX, MACD y RSI, así como algunas asociaciones establecidas manualmente. Consideraron cuatro algoritmos: Regresión Logística, Redes Neuronales, Gradient Boosting y Bosques aleatorios. Utilizan el mismo enfoque buscando predecir la subida del precio dado un objetivo en un determinado período de tiempo.

2.1.2. Un enfoque de aprendizaje automático para el comercio automatizado

Ning Lu explora la aplicación de varios algoritmos de aprendizaje enfocados en el trading automático, entre ellos: Regresión Logística, Naïve Bayes y Máquinas de vectores de soporte. Los algoritmos son probados en activos pertenecientes al S&P500. El enfoque de Lu es el de utilizar como variables predictoras no solo precios del activo en el cual se va a invertir, sino, usar información de precios de otros activos que puedan influir en el principal.

2.1.3. Modelos predictivos para el mercado FOREX

Huerta López utiliza dos enfoques para el trading automático en el mercado FOREX: los modelos de series de tiempo y los modelos basados en técnicas de aprendizaje automático. Por un lado aplica modelos ARIMA y métodos de automatización para su implementación basados en criterios de información -BIC y AIC-. Para el enfoque basado en aprendizaje automático implementa diversos algoritmos de KNN y Árboles de decisión, optimizando los algoritmos en la distintas horas del día utilizando datos del par EUR-USD. Señala que los modelos de

aprendizaje obtienen mejor predicción que los modelos ARIMA, acercándose a un 60 % de precisión.

2.1.4. Un Análisis de estrategias de trading técnico

Kadida Shagilla profundiza sobre las alternativas a la hipótesis del mercado eficiente y cómo se relacionan el riesgo y los retornos de una estrategia de trading. Simula tres portafolios conformados por una muestra de acciones de tres mercados Norteamericanos y algunos mercados emergentes de África El estudio demuestra cómo la relación 'book-to-market ratio', la liquidez y los acuerdos institucionales pueden explicar el exceso de ganancia a partir del análisis técnico.

2.1.5. Modelos Ocultos de Markov Aplicados al Reconocimiento de Patrones del Análisis Técnico Bursátil

Cristián Fernández introduce un sistema automático para el reconocimiento de patrones del análisis técnico basado en modelos de Markov. Desarrolla un algoritmo que extrae patrones del mercado y los clasifica en tres estados: lateral, alcista y bajista. Utiliza una clasificación basada en Árboles de decisión y apoya el reconocimiento de patrones en algunos indicadores técnicos. Concluye que es posible estimar correctamente el estado de un activo mediante una lectura automática de los patrones en su serie de tiempo. El modelo es testeado en varios activos que operan en el mercado bursátil argentino.

2.2. Bases Teóricas

2.2.1. Hipótesis del Mercado Eficiente

La Hipótesis del Mercado eficiente fue desarrollada por Eugene Fama en los años 60, en la misma argumenta que los precios de los activos reflejan toda la información disponible, es decir que siempre son transados a un valor adecuado para su riesgo, haciendo imposible para los inversores obtener retornos más elevados que los del mercado en general.

Fama sugiere tres suposiciones. Primero, el mercado eficiente requiere un gran número de competidores buscando maximizar ganancias. Segundo, la información que afecta al activo llega al mercado de manera aleatoria y cada anuncio es independiente de los demás. Tercero, todos los competidores intentarán ajustar sus posiciones lo más rápido posible conocida la información del mercado. Existen tres variantes de la hipótesis:

Eficiencia débil, en esta variante, los precios del pasado no sirven para predecir el precio futuro, es decir cualquier movimiento del activo es determinado por información no contenida en la serie de precios. Eficiencia media, en esta forma se asume que los precios se ajustan instantáneamente a la información pública, por lo que rechaza cualquier tipo de arbitraje intentando aprovechar nueva información. Eficiencia fuerte, esta última forma de la hipótesis plantea que los precios reflejan tanto información pública como privada, por lo cual incluso

obteniendo información no conocida por todos los competidores, no se pueden obtener retornos anormales al de los mercados.

Aunque esta hipótesis es la piedra angular de la teoría financiera moderna, es controversial entre la comunidad financiera y disputada frecuentemente. Gran parte de sus detractores argumentan que el precio del activo está influenciado por suposiciones sesgada de los individuos, formuladas por la manera en cómo estos responden ante nueva información.

Los inversores interpretan la información de manera distinta, por lo que generarán diferentes valuaciones de un mismo activo, lo que sugiere que la reacción del inversor a la misma noticia será distinta. Day and Wangr (2002) argumentan que si los precios son continuamente influenciados por estas interpretaciones erróneas, los movimientos contrarios del precio pueden ser predecidos estudiando la data histórica. Sugieren también que mientras más extremo sea el movimiento inicial, mayor será el ajuste de precio.

Los inversores se dejan influenciar por la tendencia del mercado, este comportamiento se ha visto a lo largo de la historia en casos de colapso del mercado como en la caída del mercado bursátil en 1987 ó la burbuja del puntocom a finales de los 90. Froot (1992) muestra cómo estos comportamientos pueden resultar en ineficiencias del mercado.

Algunos académicos como Hong y Stein's (1999) categorizan a los inversores en Informados y No informados. Los inversores que tienen acceso a la información solo operan al obtener nueva información, mientras que los no informados operan basados en el pasado reciente del activo. A medida que la información es conocida por todos los competidores, se forma el fenómeno de reversión a la media.

Es evidente la postura que se asume en la presente investigación con respecto a la hipótesis de mercado eficiente. Además de los aspectos del comportamiento de los competidores, se ha evidenciado en la historia, casos de inversores que han logrado vencer el mercado por largos períodos de tiempo, como Warren Buffet, lo cual por definición de la hipótesis es imposible. Por otro lado, Los avances tecnológicos y la capacidad de procesamiento de las computadoras en la actualidad hacen pensar que cualquier anomalía presente en el mercado por muy pequeña que sea puede ser aprovechada por sofisticados softwares automatizados. Shagilla Kadida (2006)

2.2.2. Análisis Técnico

Los inversionistas que rechazan la hipótesis del mercado eficiente buscan interpretar la situación del mercado, bien a través de noticias que afecten al activo o estudiando su movimiento intentando extraer patrones de conducta. A la primera técnica se le llama Análisis Fundamental y el segundo Análisis Técnico. El Análisis Fundamental está más asociado a estrategias de inversión pasivas a largo plazo aunque en la actualidad se han desarrollado algoritmos de compra y venta que buscan predecir la dirección del precio en función de noticas utilizando minería de texto.

El análisis técnico es aquel que busca patrones y tendencias de comportamiento en la cotización de los activos financieros, basándose en la serie de tiempo del activo, con esto intenta predecir el movimiento futuro mediante el uso de gráficos. Según J.J.Murphy (1999) existen tres fundamentos básicos en los que se basa el análisis técnico: Los movimientos del mercado lo descuentan todo, los precios se mueven por tendencias y la historia se repite.

Murphy establece que cualquier efecto que posiblemente pueda afectar al precio se ve reflejado en la cotización del mismo. Por lo que un estudio del desplazamiento del activo en un período de tiempo sería suficiente para lograr predecir su movimiento. Esto quiere decir que el análisis técnico no es más que una manera indirecta de estudiar los fundamentos del activo, suponiendo que la cotización del mismo resume toda la información que lo afecta.

El analista técnico acepta la premisa de que los mercados tienen tendencias. Buscar tendencias en las primeras etapas de su desarrollo es la razón de toda la representación gráfica dentro del análisis, con el fin de que las transacciones vayan en dirección de esa tendencia. Por otro lado, la afirmación de que la historia se repite tiene que ver con el estudio de la psicología humana. Ésta afirmación tiene también una estrecha relación en los ciclos económicos.

A continuación se establece la formulación de cada indicador utilizado en el estudio, se utilizará la notación $EMA_{(t,p)}(X)$ para referirse a una media móvil exponencial en t calculada con p observaciones anteriores con respecto a la serie de precio X:

- Retornos con respecto al precio de Cierre.

$$R_t = \frac{Pcierre_t}{Pcierre_{t-1} - 1}$$

- RSI (Relative Strength Index) de 14 períodos, el cual es un indicador de volatilidad.

$$RSI_t = 100 - \frac{100}{1 + RS_t} \qquad donde, RS_t = \frac{Ganancia\ Promedio\ en\ n\ observaciones}{Perdida\ Promedio\ en\ n\ observaciones}$$

- MACD (Moving Average Converge/Divergence) el cual es una diferencia de dos EMAs (Exponential Moving Average) de 12 y 26 períodos. Este es un indicador de tendencia que se complementa con un MA(Moving Averge) de 9 períodos.

$$MACD_t = EMA_{(t,12)}(Pcierre) - EMA_{(t,26)}(Pcierre) \qquad Signal\ Line_t = EMA_{(t,9)}(MACD_t)$$

- ATR (Average True Range), es un indicador de volatilidad calculado a partir de los máximos y mínimos de un período, en este caso 14.

$$ATR_t = EMA_{(t,14)}(TR_t) \qquad TR_t = Max\left(Pmax - Pmin, |Pmax - Pc_{t-1}|, |Pmin - Pc_{t-1}|\right)$$

- ADX (Average Directional Index), este es un indicador que utiliza dos indicadores de dirección +Di y -Di, se calculó en base a 14 períodos y mide tendencia.

$$ADX_t = 100\ EMA_{(t,14)}(|+DI - -DI|)$$

- Bandas de Bollinger, el cual es un indicador de tendencia y volatilidad, utiliza dos bandas calculadas a partir de una media móvil con desviaciones estándar. Se utilizó en base a 14 períodos y una desviación de 2.5.

$$BandaSuperior_t = EMA(t,14)(Pcierre) + 2,5\sigma \qquad BandaInferior_t = EMA(t,14)(Pcierre) - 2,5\sigma$$

2.2.3. Introducción al aprendizaje automático

Aprendizaje automático refiere a una rama de la Inteligencia Artificial, que busca crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de un conjunto de datos. Supongamos que existe una variable respuesta Y y distintos predictores $X_1, X_2, ..., X_j$. Se asume que existe una relación entre Y y $X = X_1, X_2, ..., X_j$, la cual puede ser escrita de forma general como

$$Y = f(X) + \epsilon$$

donde f es una función desconocida de X y ϵ es un término de error aleatorio, independiente de X y de media 0. En esta formulación f representa información sistemática que X proporciona sobre Y.

En escencia, el aprendizaje automático refiere a un conjunto de enfoques para estimar f

Métodos Paramétricos vs No Paramétricos

La mayoría de los métodos de aprendizaje automático pueden ser caracterizados como paramétricos o no paramétricos. Los primeros, involucran un enfoque basado en dos pasos. Primero se asume que los datos toman una forma específica, una vez asumida la forma que debe tener la función f, el problema de la estimación se simplifica. al seleccionar el modelo se procede a ajustar el modelo en la data de entrenamiento. Este es el caso de los Modelos Lineales Generalizados como la Regresión Logística. La desventaja de este enfoque paramétrico es que el modelo escogido puede no ser apropiado a la verdadera forma de f, por lo que la estimación puede ser pobre.

Por otro lado los modelo No Paramétricos no asumen ninguna forma para f, en cambio, estos modelos buscan estimar f acercandola lo más posible a los datos observados. Esto les permite evadir el problema de ajustarse a alguna forma en específico. Sin embargo, al no reducir el problema a estimar unos parámetros sino utilizar los datos directamente, se necesita un gran número de observaciones -muchas más que las necesarias por los métodos paramétricos- para obtener una estimación precisa. Además en general la interpretación del modelo se hace más difícil con estos métodos y son propensos a caer en sobreoptimización

Aprendizaje Supervisado vs No Supervisado

Se le llama aprendizaje Supervisado, a los métodos en los cuales para cada observación de las variables predictoras x_i existe un valor asociado a la variable respuesta y_i. Por lo que se ajusta un modelo que relacione la respuesta con los predictores, con el fin de predecir acertadamente respuestas futuras. Este es el caso de los modelos Lineales así como los métodos de boosting, SVM, GAM, etc.

En contraste, lo métodos No Supervisados describen una situación más complicada, en donde para cada observación, se cuenta con variables predictoras, pero no existe ninguna

variable respuesta. Lo que se busca en este tipo de modelos es buscar entender la relación entre las variables o entre las observaciones. Para esto se utilizan métodos de agrupación o cluster y métodos de reglas de asociación Los primeros intentan describir las agrupaciones subyacentes en los datos, como por ejemplo, el tipo de clientes dependiendo de su comportamiento de compra. Las reglas de asociación buscan descubrir patrones inherentes que describan el comportamiento de los datos u observaciones, por ejemplo, un grupo de clientes que compran un producto r conjunto con otro producto s.

Regresión vs Clasificación

Las variables pueden ser divididas entre cuantitativas o cualitativas -también llamadas categóricas-. Las cuantitativas toman valores numéricos mientras que las cualitativas son categorías o clases. Dependiendo del tipo de variable respuesta se realiza el enfoque del modelo. En el caso de que la variable respuesta sea cuantitativa se refiere a problemas de regresión, mientras que los que involucran una variable respuesta cualitativa, son referidos como problemas de clasificación.

2.2.4. Validación Cruzada en Series de Tiempo

La validación Cruzada es un método de validación y prueba que consiste en dividir los registros aleatoriamente en grupos de similar tamaño. El primer grupo es utilizado como validación del modelo que ha sido entrenado en el resto de los datos, este proceso se realiza k veces, y el resultado final es el promedio arrojado por cada una de las k validaciones.

Ahora bien este método asume que no existe relación entre las observaciones, es decir que son independientes. Esto no es verdad en el caso de las series de tiempo debido a la condición de autoregresión. Por lo tanto al dividir la data se debe respetar el orden temporal de cada observación.

2.2.5. Regresión Logística

Los modelos Lineales Generalizados asumen que existe una aproximada relación lineal entre la variable respuesta Y y la variable predictora X. Matemáticamente se puede describir la relación como:

$$Y \approx \beta_0 + \beta_1 X_1 + ... + \beta_j X_j$$

En donde X_j representa las variables predictoras y los coeficientes β_j cuantifican la asociación entre la variable predictora X_j y la variable respuesta Y. Por lo que se interpreta a β_j como el efecto promedio que tiene en Y un incremento de una unidad en X_j, bajo el supuesto de que todas las demás variables se mantienen constantes.

En problemas de clasificación, la variable predictora asume valores categóricos, por lo que al utilizar este enfoque se pueden obtener probabilidades fuera del intervalo [0, 1], haciendo

imposible su interpretación. Esto concluye en que se deba utilizar una función, tal que permita la generación de valores entre [0, 1], en el caso de la regresión logística esta función es:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...++\beta_j X_j}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...++\beta_j X_j}}$$

Despejando se obtiene

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...++\beta_j X_j}$$

El lado izquierdo de la ecuación puede tomar valores entre 0 e ∞, lo cual indicaría muy bajas o muy altas probabilidades, aplicando logaritmo en ambos miembros de la ecuación se obtiene la función logit

$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + +\beta_j X_j$$

Se observa que la función logit es lineal en X, por lo que incrementar una unidad de X afecta el lado izquierdo de la ecuación en β. Sin embargo dado que la relación entre $p(X)$ y X no es una línea recta, β no corresponde a un cambio en $p(X)$ asociado a una unidad de incremento en X. Se debe hacer la respectiva transformación para interpretar el coeficiente β en relación a Y.

Máxima Verosimilitud

Los coeficientes β son desconocidos, por lo que deben estimarse en la data de entrenamiento. Para esto se utiliza el método de $MximaVerosimilitud$, el cual consiste en estimar los coeficientes para los cuales la probabilidad de predicción para cada individuo, utilizando (fórmula arriba), corresponda lo más cercano posible al valor observado del individuo. Se define la función de verosimilitud como

$$l(\beta) = \prod_{i=1}^{j} P(x_i/\beta)$$

Por conveniencia se trabaja con el logaritmo, dado que esto transforma una operación de productos de probabilidades en una sumatoria, por lo que se obtiene

$$l(\beta) = \sum_{i=1}^{N} \log P(y_i/x_i;\beta)$$

Al codificar las clases en 0 y 1, la función de verosimilitud para la regresión logarítmica puede ser escrita como

$$l(\beta) = \sum_{i=1}^{N} (y_i \beta^T x_i - \log 1 + e^{\beta^T x_i})$$

Para maximizar la función, se iguala la derivada a 0

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - P(x_i; \beta)) = 0$$

Para resolver la ecuación (n arriba) se utiliza un algoritmo de optimización llamado $Newton - Raphson$

2.2.6. Análisis de Componentes Principales

Los modelos lineales tienen distintas ventajas en cuanto a interpretación y muchas veces son sorprendentemente competitivos en relación con los métodos no lineales. Existen técnicas para relajar el supuesto de que la relación entre la respuesta y los predictores es lineal, arrojando mejores predicciones e interpretabilidad.

Una clase de métodos es el enfoque de Reducción de la Dimensión, el cual involucra proyectar los p predictores en M-dimensiones o componentes, donde $M < p$. Esto se logra transformando los predictores en combinaciones lineales que recogen parte de la información, estas M componentes o dimensiones son entonces utilizadas como nuevos predictores en el modelo de regresión. Esto es:

$$Z_m = \sum_{i=1}^{p} \phi_{im} X_i$$

Para cualquier constante $\phi_{1m}, \phi_{2m}, ..., \phi_{pm}, m = 1, ..., M$. Se ajusta el modelo

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \qquad i = 1, ..., n$$

En situaciones donde p es relativamente grande con relación a n, seleccionar un valor de $M << p$ puede reducir considerablemente la varianza de los coeficientes. Es de notar que si $M = p$, ajustar el modelo con las combinaciones lineales de los coeficientes originales es equivalente a ajustar el modelo original.

El Análisis de Componentes Principales (ACP) es una técnica que reduce la dimensión de una matriz de datos. La dirección del primer componente principal es aquella en la cual exista mayor variación entre las observaciones, es decir

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + ... + \phi_{p1} X_p$$

donde, $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Los elementos de ϕ son llamados *loadings*, y el subíndice representa el número de componente. Juntos, los *loadings* forman el vector de *loadings* $\phi_1 = (\phi_{11}\phi_{21}...\phi_{p1})^T$.

Dado una matriz de datos X de n x p, se asume que cada variable en X está normalizada -tiene media 0-, entonces se obtiene la combinación lineal de los valores de los predictores que contiene la mayor varianza, llamadas *scores*.

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + ... + \phi_{p1}x_{ip}$$

El primer vector de loadings de componente principal resuelve el problema de optimización

$$\max_{\phi_{11},...,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2 \qquad sujeto\ a \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

El problema de maximización en (fórmula de arriba) se soluciona mediante la descomposición de los eigenvalores. Luego de determinar el primer componente Z_1, se procede a encontrar el segundo componente Z_2, el cual es una combinación lineal de $X_1, ..., X_p$ que tiene la maximiza varianza de todas las combinaciones lineales que no están correlacionadas con Z_1. Así los scores del segundo componente principal toman la forma

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2+...+\phi_{p2}x_{ip}}$$

donde ϕ_2 es el segundo vector loading del componente principal. Es de notar que restringir Z_2 a no ser correlacionada con Z_1 es equivalente a restringir la dirección de ϕ_2 a ser ortogonal a la dirección de ϕ_1.

El utilizar la técnica de componentes principales en el modelo de regresión también soluciona el tema de la multicolinealidad entre las variables.

2.2.7. Matríz de Confusión

En los problemas de clasificación se utiliza la matriz de confusión para evaluar el desempeño del modelo. La misma es una tabla que categoriza las predicciones realizadas por el modelo de acuerdo a la coincidencia con los valores reales.

La estrategia solo toma la señal cuando el modelo predice un incremento en el precio, la venta por el contrario no depende del modelo, sino de los parámetros predefinidos (porcentaje de Stop Loss y Horizonte de tiempo). Esta característica implica que el valor a maximizar es la predicción de los verdaderos positivos, conocido como Precisión.

$$Precisin = \frac{VP}{VP + FN}$$

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 2.1: Matriz de Confusión

2.2.8. Medidas de Riesgo

Valor en Riesgo

El valor en riesgo ó VaR por sus siglas en inglés -Value at Risk- es una medida común del riesgo implementado en instituciones financieras. Se define como la pérdida en un portafolio, tal que existe una probabilidad p que las pérdidas sean iguales o mayores que el VaR y una probabilidad $(1 - p)$ que sean menores que el VaR, en un tiempo determinado. Este valor se corresponde con el cuantil de la distribución de pérdida.

$$Pr(Q \leq -VaR(p)) = p$$

ó

$$p = \int_{-\infty}^{-VaR(p)} f_q(x)\ dx$$

siendo $f_q(x)$ la función de densidad de la variable aleatoria pérdida/ganancia del portafolio denotada por Q.

Aunque el VaR es utilizado comúnmente en instituciones financieras e incluso exigido en algunas regulaciones como Basilea, existen varias críticas en cuanto a su implementación. Una de las críticas es su inconclusividad en cuanto al tamaño de la pérdida, en este sentido, el VaR es un cuantil de la distribución que establece un umbral en cuanto a la posible pérdida en un período, dado un nivel de significancia. Jon Danielsson (2011)

Pérdida Esperada

La pérdida esperada mejor conocida conocida como 'Expected Shortfall' -ES- es una medida alternativa al VaR que responde la pregunta ¿Cuál es la pérdida esperada cuando éstas exceden el VaR?. Formalmente la pérdida esperada se define cómo

$$ES = E(Q/Q \leq VaR(p))$$

ó, utilizando la formulación matemática de esperanza

$$ES = \int_{-\infty}^{-VaR(p)} x f_{VaR}(x)\, dx$$

Es importante acotar que para el ES existen dos fuentes de error ya que primero se debe estimar el VaR y luego obtener la esperanza de la cola de la distribución. Sin embargo es una medida que complementa el VaR. En la figura 2.2 se observa la función de densidad de una distribución normal (0,1) donde se representa el VaR y ES, el VaR vendría siendo la frontera entre las dos áreas mientras que el ES la esperanza del área azul o cola superior. Es importante acotar que tanto el VaR como ES a pesar ser medidas de pérdidas potenciales, pueden ser referidas con signo negativo ó positivo.

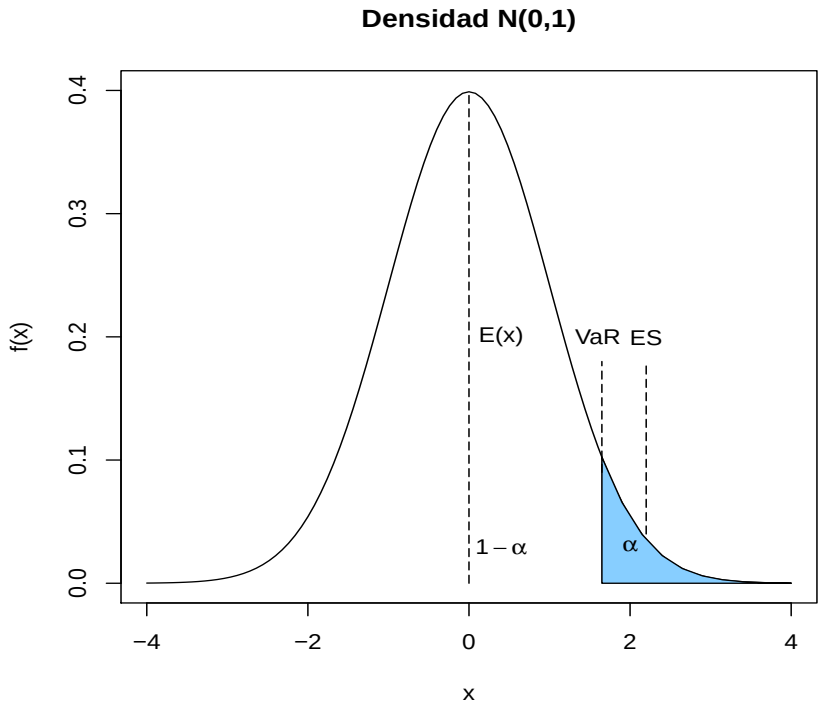

Figura 2.2: Función de densidad con VaR y ES

2.3. Bases Legales

2.3.1. Superintendencia Nacional de Valores

La Ley que rige las entidades de inversión en Venezuela entró en vigencia el 22 de agosto de 1996 y fue publicada en Gaceta Oficial Número 36.027. Esta se centra en entidades con capacidad de manejar importantes flujos de recurso de inversionistas hacia el mercado de capitales. El órgano encargado de autorizar, regular, controlar, vigilar y supervisar a las Entidades de Inversión y sus sociedades administradoras es la Superintendencia Nacional de Valores

2.3.2. El uso del VaR para la Regulación de Entidades Financieras

La utilización del VaR para la regulación en riesgo financiero se inicia en 1988 con el acuerdo de Basilea, principalmente dada las correlaciones entre los elementos de los portafolios. En 1995 se presentó un anexo sobre el modelo de riesgo de mercado, el cual autorizaba la elaboración propia del modelo de riesgo para determinar requerimientos de capital. Recomendando para la aplicación del VaR los siguientes parámetros:

- Horizonte de 10 días de operación o de dos semanas de calendario
- Intervalo de confianza de 98 %
- Un período de observación basado en un año de datos históricos actualizados, al menos una vez por trimestre

Capítulo 3

Marco Metódico

En el presente capítulo se describen los métodos utilizados en la investigación, con la finalidad de dar respuesta a los objetivos planteados, se indica el tipo de investigaciión, universo y muestra, la fuente donde se extraen los datos a usar y otros detalles referentes al método de estudio. De igual manera se expone la metodología para el análisis de los resultados.

3.1. Tipo de Investigación

Para definir el tipo de investigación a realizar se establece como referencia la clasificación expuesta por Balestrini (1997). La autora señala varias categorías: formulativo o exploratorio, descriptivo, diagnóstico, evaluativo, experimental ó proyecto factible. De acuerdo al problema planteado, la presente investigación es de tipo descriptivo donde se verifica que se alcancen los objetivos planteados con las conclusiones derivadas de los resultados de la simulación

3.2. Universo y Muestra

Balestrini (1993) define al universo ó población como cualquier conjunto de elementos de los que se pretende investigar y conocer sus caracteristicas, para el cuál serán válidas las conclusiones obtenidas en la investigación. Así mismo define la muestra como una parte de la población seleccionados científicamente, obtenida con el fin de investigar, a partir del conocim¡ento de sus características particulares, las propiedades de esa población.

El universo de estudio está representado por los índices bursátiles de los mercados financieros existentes entre el período 26/10/2008 - 18/01/2019. Un índice bursátil es un promedio de los precios de los activos que representan un mercado o sector determinado. Los mismos sirven como 'benchmark' o referencia de la economía de un país, sector financiero, etc. En el ámbito de los 'hedge funds' son una referencia para medir la rentabilidad de una estrategia de inversión y el riesgo del mercado.

En la presente investigación se utilizan los índices como reflejo del comportamiento de varios activos, de esta manera, se mide la estrategia en un sector y no en un instrumento en

específico. Otras de las ventajas de utilizar los índices es que al representar un promedio de varios activos, sus variaciones son menos drásticas. La muestra está constituida por 5 índices bursátiles que representan distintos mercados del mundo: NASDAQ, NIKKEI, FTSE 100, BOVESPA y SP500.

3.2.1. Tipo de Muestreo

Seijas, G. (1993) define dos tipos de muestreo: Probabilístico y No Probabilístico. El muestreo es probabilístico cuando se puede determinar de antemano la probabilidad de selección de cada uno de los elementos de la población siendo ésta distinta de cero; por lo tanto, calcular con antelación la probabilidad de obtener cada una de las muestras posibles.

En este caso, se realizó un muestreo no probabilístico de clase opinática, ya que la selección de los elementos de la muestra se debe a su nivel de representatividad dentro de los mercados bursátiles, por ende, la totalidad de los índices bursátiles no fueron equiprobables en su selección.

3.3. Fuentes de Datos

La estructura de los datos utilizados en el trabajo es de tipo OHLC por sus siglas en inglés Open, High, Low, Close. La misma, resume en 4 registros el comportamiento del precio del activo (Apertura, Cierre, Mínimo y Máximo) en un intervalo de tiempo. En el caso de la presente investigación, de un día. Este tipo de dato provee la información necesaria para cubrir las exigencia del modelo, tanto para la creación de la variable dependiente como para el cálculo de los indicadores técnicos.

Los datos fueron extraídos del portal www.investing.com, uno de los portales financieros con mayor prestigio en el mundo. Fue fundado en 2007 y es conocido por su calendario económico y directorio de brokers.

3.4. Variables

3.4.1. Variable Dependiente

Las decisiones de entrada en el trading pueden ser producto de muchos factores, en la presente investigación se analiza el enfoque donde se define un porcentaje objetivo de ganancia y se intenta predecir si dicho objetivo se materializa en un futuro cercano, sin que se haya concretado una venta por Stop Loss. Este enfoque reduce la toma de decisión en una variable tal que:

$$P_X(x) = \begin{cases} p\ ; & x = c \\ \\ 1-p\ ; & x = -d \end{cases}$$

Dado los datos OHLC del activo es posible identificar los períodos en donde se materializa la variable dependiente. la identificación se realiza, comparando el precio de cierre con los precios máximos y mínimos de las siguientes h observaciones, donde h es el número de períodos, en este caso días en los cuales se desea evaluar la condición.

En la práctica se identifica los registros que cumplen con esta condición añadiendo una columna a la data donde incluimos 'buy' para identificar los registros donde se da la señal y 'stay' en caso de que no haya ocurrido o hubiese ocurrido primero el retroceso del precio.

3.4.2. Variables Predictoras

Los indicadores a utilizar fueron seleccionados buscando recoger la mayor información posible sobre el precio del activo, se pueden resumir en tres categorías: tendencia, momentum y volatilidad.

No es de interés en la presente investigación describir cómo funciona cada indicador para la toma de decisiones en el trading basado en fundamentos técnicos. Cada indicador puede utilizarse de distintas maneras, calcularse con distintos parámetros y asociarse a discreción del trader, lo que conlleva a un sin fin de reglas de asociación

Lo que busca la investigación es utilizar la relación entre los indicadores como variables independientes que ayuden al modelo a predecir oportunidades de entradas. En este sentido se asume la existencia de una dinámica local del mercado que puede ser predecida con ayuda de estos indicadores. Algunos de los indicadores inicialmente escogidos para el modelo fueron excluidos debido a la alta correlación que presentan. A continuación se exponen los indicadores utilizados:

- Retornos con respecto al precio de Cierre.
- RSI (Relative Strength Index) de 14 períodos, el cual es un indicador de volatilidad.
- MACD (Moving Average Converge/Divergence) el cual es una diferencia de dos EMAs (Exponential Moving Average) de 12 y 26 períodos. Este es un indicador de tendencia que se complementa con un EMA de 9 períodos.
- ADX (Average Directional Index), este es un indicador que utiliza dos indicadores de dirección +Di y -Di, se calculó en base a 14 períodos y mide tendencia.
- Bandas de Bollinger, el cual es un indicador de tendencia y volatilidad, utiliza dos bandas calculadas a partir de una media móvil con desviaciones estándar. Se utilizó en base a 14 períodos y una desviación de 2.5.
- ATR (Average True Range), es un indicador de volatilidad calculado a partir de los máximos y mínimos de un período, en este caso 14.

Estos indicadores están fuertemente correlacionados por lo que se decidió, disminuir el número de variables dejando solo las más representativas de cada indicador.

Ahora bien, la idea base de la investigación era utilizar los valores de cada indicador como variables predictora. Dado que el cálculo de todos los indicadores provienen de la misma variable -precio del activo, en la mayoría de los casos precio de cierre-, existe una alta colinealidad entre ellos, la idea de utilizar el ACP es precisamente para enfrentar este problema como se detalla más adelante. Sin embargo, es de notar que los valores de los indicadores por sí solos no proveen un poder predictivo, lo que realmente usa el trader son las asociaciones entre indicadores para encontrar patrones.

Se decidió entonces, utilizar como predictores no los indicadores por si solos, sino, las relaciones entre cada uno de ellos. Esto se abordó agregando al modelo las interacciones entre todos los indicador, y removiendo los valores de los indicadores por sí solos. De esta manera el hecho de utilizar ACP, no solo es visto ahora como una manera de remover la colinealidad entre predictores sino como método de reducción de variables, ya que el modelo pasó de tener 10 predictores -incluyendo el precio de cierre- a 45.

3.5. Estrategia de Análisis

3.5.1. WalkForward Backtesting

Al principio de la investigación se implementó el método de entrenamiento, validación y prueba comúnmente utilizado, en donde la mayor parte de la data es destinada a entrenamiento del modelo, otra sección es destinada a validación, para elegir los parámetros óptimos, y finalmente se aplica el modelo en la data de prueba. Sin embargo este tipo de metodología en opinión del investigador no es el más óptimo para desarrollar el presente modelo, ya que el dinamismo de los mercados bursátiles no permite al algoritmo permanecer sin cambios en el tiempo.

Para contrarrestar esta situación se optó por el método de backtesting Walkforward, el cual consiste en entrenar el modelo en un período base de data, en este caso los primeros 4 años de estudio, posteriormente se aplica la estrategia en el año siguiente y se obtiene los primeros resultados. Luego este año de aplicación es incluído en la data de entrenamiento -es decir, la data de entrenamiento pasa a ser de 5 años- y se evalúa el modelo en el siguiente año. De esta manera, contemplamos el dinamismo del mercado permitiéndole al modelo -y por ende a la estrategia- utilizar el período más reciente con respecto al cual será implementado. En la presente figura 3.1 se ilustra la metodología implementada.

Otra de las características de la metodología que se modificó fue la elección de los valores de los hiperparámetros -target, stop y horizonte-. Previamente se utilizaba la data de validación para buscar la combinación de parámetros óptima. Ahora bien en la metodología de Walkforward se utilizan los mismos parámetros durante todo el período de estudio. A opinión del investigador al buscar los mejores parámetros se estaría incurriendo en un posible sesgo de sobreoptimización. El hecho de que en un año determinado unas configuraciones den los mejores resultados no asegura que se replique en el siguiente año.

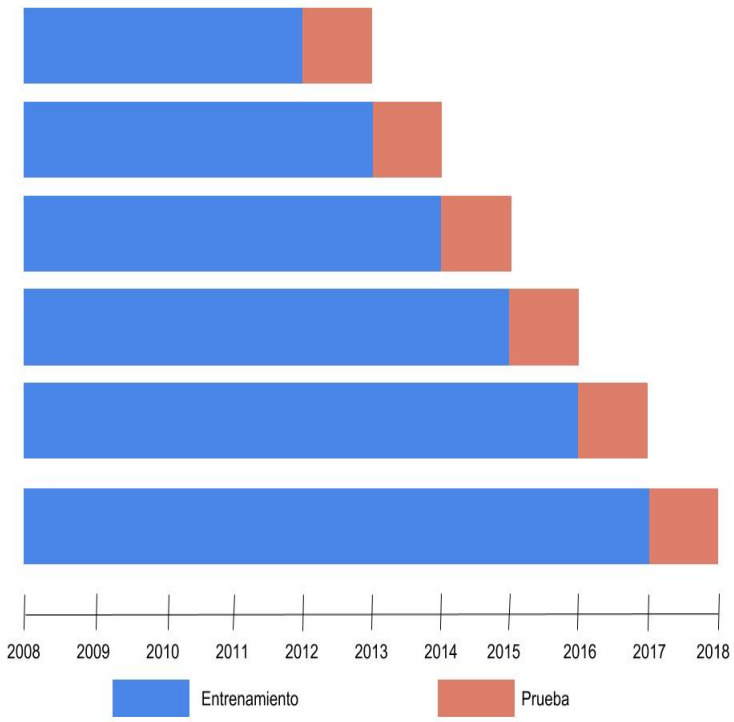

Figura 3.1: Metodología WalkForward

Si se utiliza un stop ó target muy altos, baja el número de observaciones que cumplan con la condición de la variable dependiente, por lo tanto se estaría en presencia de un problema de data desequilibrada que debe tener un tratamiento distinto. En base a esto, se establece el target en 2 % y el stop en 2,5 %. Por otro lado la elección del horizonte de tiempo influye también en el número de observaciones identificadas como 'buy', es de notar que mientras más grande el horizonte mayor número de observaciones 'buy' se tendrá. Sin embargo existe un punto en donde deja de crecer este número, en la práctica se establece en 20. En la figura 3.2 se ilustra el número de observaciones identificadas como 'buy' para distintas combinaciones de hiperparámetros.

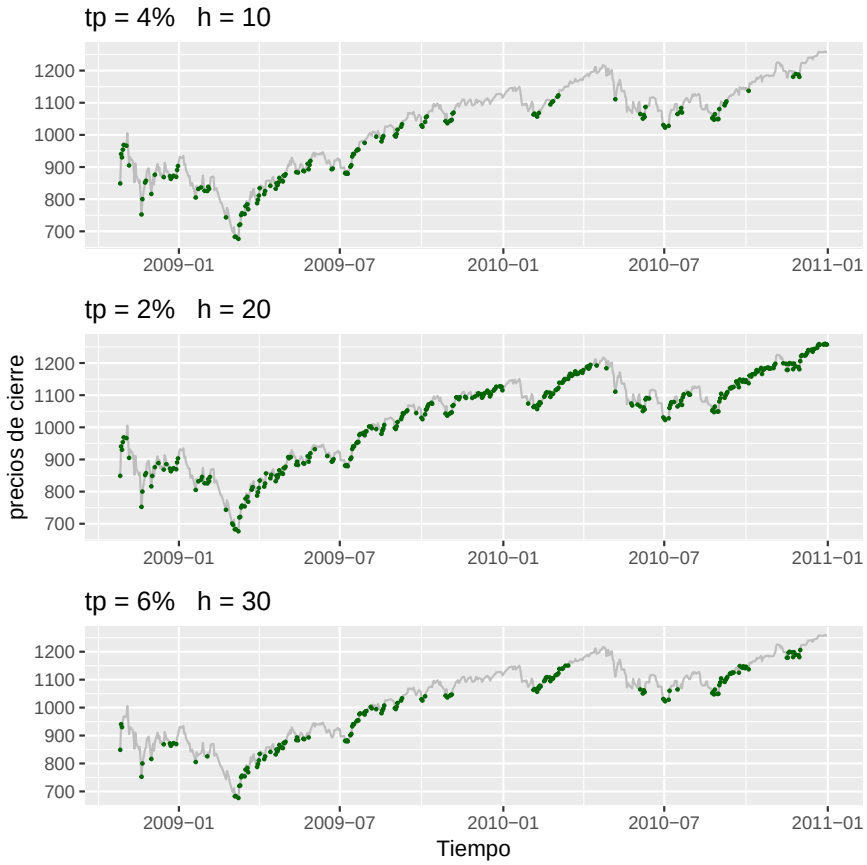

Figura 3.2: Observaciones que presentan ocurrencias con sl = 2.5 %, en el primer conjunto de datos de entrenamiento (26/10/2008 - 31/12/2010) para el índice S&P500, para diferentes valores de tp y h

3.5.2. Reducción de la dimensión con ACP

La técnica que utiliza el análisis de componentes principales (PCA) para reducir el número de variables predictoras es conocido como Principal Component Regression (PCR). PCR es utilizado para extraer la información más importante de una matriz de datos multivariante y expresar ésta información en nuevas variables llamadas componentes principales Éstas son una combinación lineal de las variables originales. Aunque el número de componentes principañes puede ser igual al número de variables, la idea es utilizar un grupo reducido de componentes que maximicen la variación.

Por su parte el modelo propuesto utiliza las interacciones entre las variables predictoras, esto aumenta el número de variables de 10 a 45, las cuales además en muchos casos están correlacionadas. Al utilizar PCR se reduce el número de variables a dos componentes que contienen alrededor del 50 % de la variación, es importante recordar que esta reducción se realiza en cada período de entrenamiento.

3.5.3. Modelado de los Retornos

Tal como se establece en la sección 3.1.3 la variable aleatoria retorno del trade puede obtener 2 posibles valores, si el trade es exitoso toma el valor c con probabilidad p, en caso contrario toma el valor -d con probabilidad 1-p. Para cada activo se puede asumir p como la probabilidad positiva -precisión- obtenida en el modelo. Por su parte c y -d son fijados en 200 y -250 respectivamente, esto viene dado por haber establecido los porcentajes de salida en 0.02 de ganancia y 0.025 de pérdida y asumir un capital a riesgo en cada trade de 10.000 USD

Asumiendo que la ocurrencia de los trades es una variable aleatoria i.i.d, es posible aplicar el teorema central del límite. Con un número de muestra suficientemente grande, la suma de estas variables se aproxima una distribución normal 0,1.

$$\frac{\sum x_i - nE(x)}{\sqrt{n}\sigma_x} \sim N(0,1)$$

siendo

$$E(x) = pc - (1-p)d \qquad y \qquad \sigma_x^2 = (pc^2 - (1-p)d^2) - (pc - (1-p)d)^2$$

Es posible modelar los retornos producidos por la estrategia y calcular el valor a riesgo -VaR- y pérdida esperada -ES- dado un número de operaciones. Tanto el VaR como el ES son medidas comúnmente utilizadas para representar el riesgo de pérdida en un período de tiempo determinado, en este caso, se establecerá en vez de tiempo, un número de trades cerrados por la estrategia.

Dado que el retorno de la estrategia se distribuye N(0, 1), se definen el VaR y ES cómo

$$VaR_\alpha = \sigma\Phi^{-1}(\alpha) - \mu$$

$$ES_\alpha = \sigma\frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha} - \mu$$

En la presente investigación se establecen $n = 300$ y $\sigma = 0.95$. Es decir que el VaR puede interpretarse como: "Existe una probabilidad del 5 % que la estrategia genere una pérdida igual ó menor que VaR_α, luego de 300 trades realizados".

Mientras que el ES se interpreta como: "En caso de que la estrategia genere una pérdida mayor que VaR_α luego de 300 trades, se espera que ésta pérdida sea de ES_α"

Capítulo 4

Análisis de Resultados

4.1. Análisis Exploratorio de los datos

En el presente capítulo se realiza la descripción de los resultados obtenidos luego de la aplicación del método propuesto para la simulación de la estrategia. De igual modo, se presentan los coeficientes arrojados por el MLG así como un analisis de los componentes principales para el índice S&P500

4.1.1. Análisis de los Índices

En la figura 4.1 se observa el precio de cierre de cada índice durante el período de estudio. Se puede apreciar que en general los cinco mercados tienen tendencia a la alta. En general para el S&P y NASDAQ se aprecia una menor variabilidad que en los demás índices.

Las curvas del S&P y NASDAQ son muy parecidas ya que ambas representan cotizaciones de las mayores empresas americanas. Se aprecia un crecimiento sostenido desde el año 2009, con algunos períodos mas estables, a principios del 2016 -por la polémica victoria de Donald Trump como presidente- y en 2018 provocado por el miedo a un incremento en las tasas de interés por parte de la Reserva Federal y las tensas negociaciones con China.

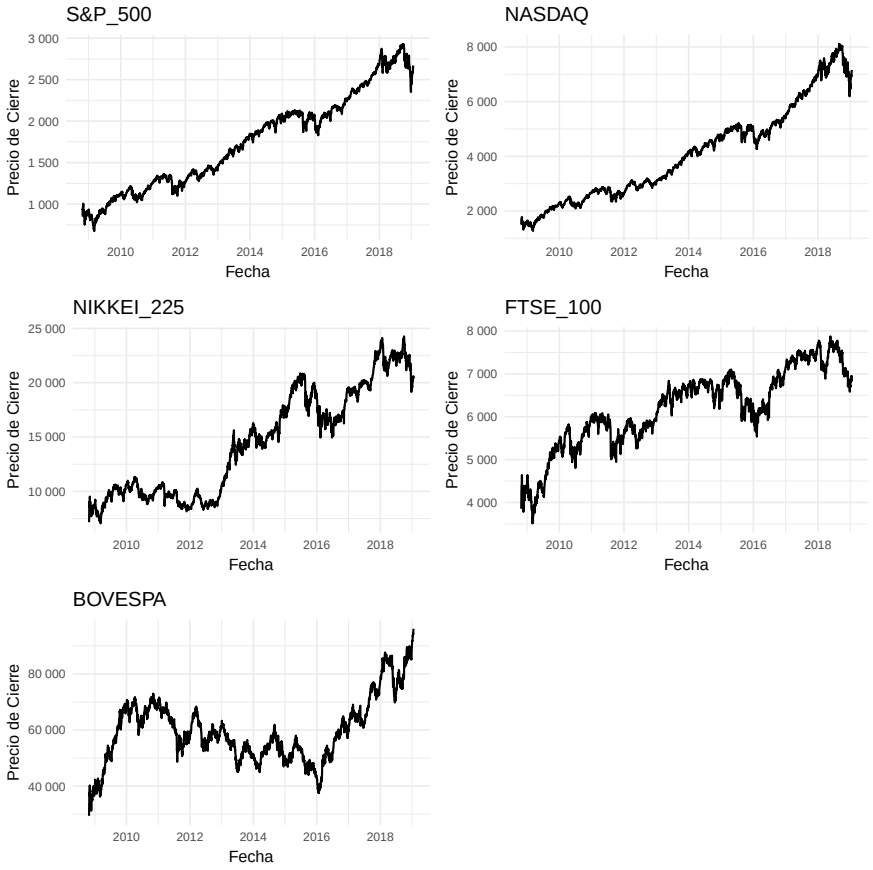

Figura 4.1: Precios de Cierre de los índices en el período de estudio (26/10/2008 - 18/01/2019)

Por su parte el NIKKEI es el índice bursátil de la bolsa de Tokio, resume las cotizaciones de 225 grandes empresas de Japón de distintos sectores industriales. Al ver la gráfica de sus cotizaciones se observa un rápido crecimiento entre 2009 y 2010, sin embargo con la ocurrencia del terremoto registrado en el norte del país a principios del 2011 el índice se vio afectado y llegó a su nivel más bajo de 8160 puntos el 25 de noviembre desde el 10 de marzo de 2009. La recuperación ocurrió en 2013 debido a los cambios implementados por el país en materia de política fiscal y monetaria. El Gobierno impulsó el gasto público y el Banco Central de Japón inyectó dinero en la economía a gran escala.

El FTSE 100 es un índice bursátil calculado con las cotizaciones de las 100 empresas más grandes de Reino Unido, en su gráfica se observa una tendencia alcista aunque con mayor ruido o variabilidad que los otros índices Se puede apreciar una caída en las cotizaciones el año 2015 debido a resultados negativos mostrados por la actividad manufacturera en Asia, especialmente en China. Esto afecta los mercados europeos dada que este es el principal consumidor del mercado asiático.

Por último el índice BOVESPA representa las 50 empresas de mayor capitalización de la Bolsa de Valores de Sao Paulo. Se aprecia que a diferencia de los anteriores, éste mantiene

una tendencia a la baja hasta 2016 debido a la recesión económica sufrida por Brasil mostrando una contracción del PIB en un 3.8 % para 2015. Aunado a factores ambientales como el virus del Zika y el escándalo de corrupción de Petrobras. A partir de 2016 con la inflación gradualmente bajo control y tasas de interés en decrecimiento, la confianza de los inversores llevó a un rápido crecimiento del índice en los últimos tres años.

4.1.2. Análisis de las Variables Predictoras

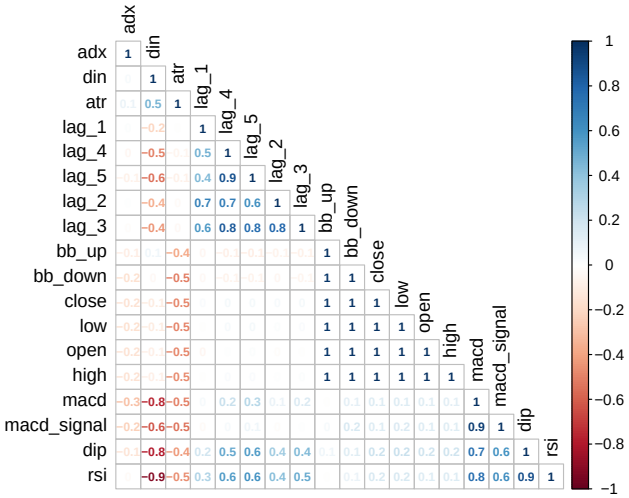

Figura 4.2: Correlación entre indicadores originales calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

En la figura 4.2 se observa la correlación de las variables originales para el primer período de entrenamiento del índice S&P500. Como se puede observar existe alta correlación entre las distintas variables del precio (Apertura, Cierre, Máximo y Mínimo) por lo que se decidió trabajar solo con los precios de cierre dado que ésta es la misma utilizada para determinar la variable dependiente. Así mismo se observa alta correlación entre los rezagos de los rendimientos, para esto se decidió trabajar solo con los rezagos de 1, 3 y 5 períodos. Por su parte se descarta la variable dip -elemento utilizado en el indicador ADX- por su fuerte correlación con el RSI. Se determina lo mismo para la banda inferior del indicador de Bollinger. En la figura 4.3 se muestra las correlaciones de las variables definitivas

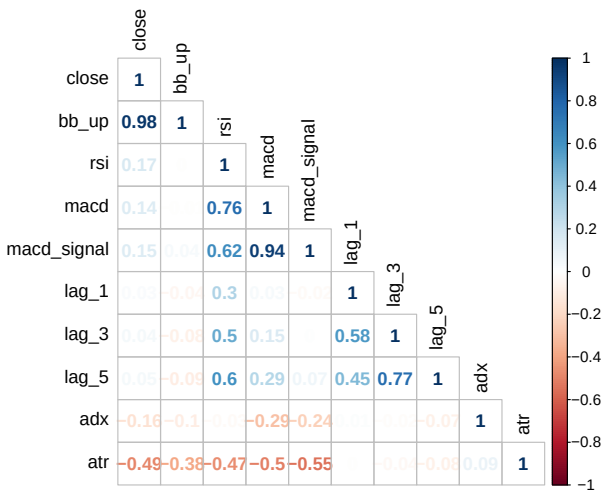

Figura 4.3: Correlación entre indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

A continuación se presentan una serie de gráficos para reflejar lo anteriormente expuesto en cuanto a la relación entre los indicadores.

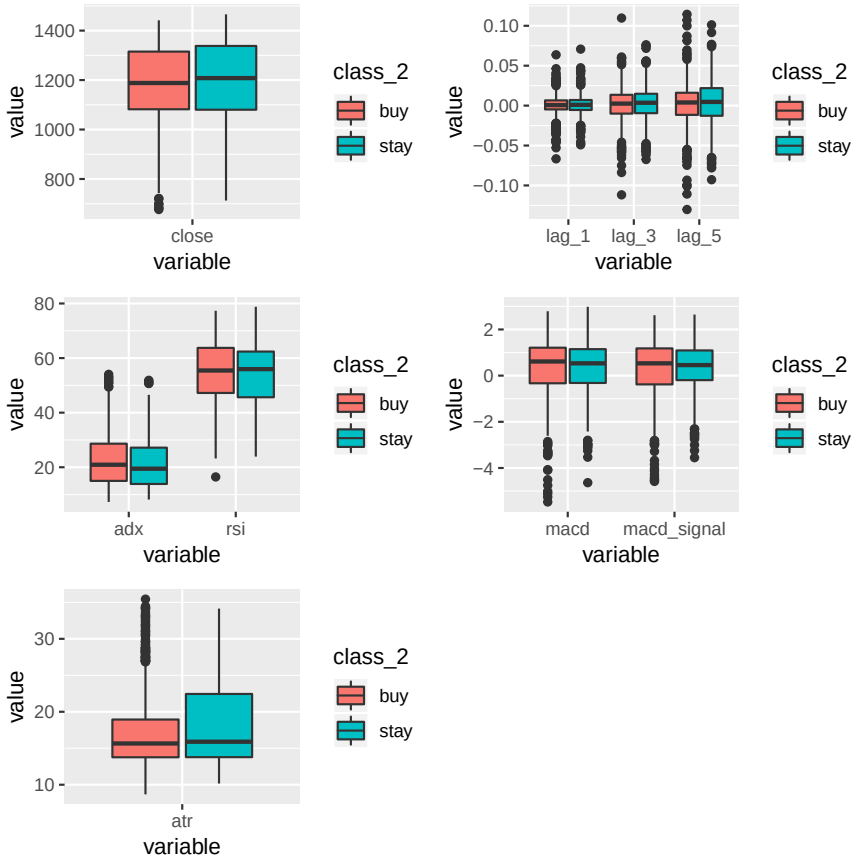

Figura 4.4: Gráfico de cajas de indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

En la figura 4.4 se observan gráficos de cajas para cada una de las variables predictoras, diferenciando entre los registros identificados como 'buy' y 'stay'. En el caso de la variable close, se aprecia que la caja está ligeramente mas abajo para los registros marcados como 'buy', señalando que efectivamente las entradas ocurren a precios bajos. En el caso los rezagos -lags- se observa que mientras mayor sea el número de períodos para su cálculo, más dispersos serán los valores y mayor diferencia existirá entre las clases.

Para la variable ADX la caja de los registros 'buy' está ligeramente más arriba que los 'stay' lo que indicaría que se debería comprar en tendencia alcista. Igualmente ocurre con la variable RSI lo que indicaría que se compra cuando se está cerca de un cambio de tendencia ó por el contrario la consolidación de la misma. Para la variable RSI se observa como la caja de los registros identificados como 'buy' esta más abajo y es más estrecha, por lo cual se inferiría que se debe comprar en períodos de baja volatilidad.

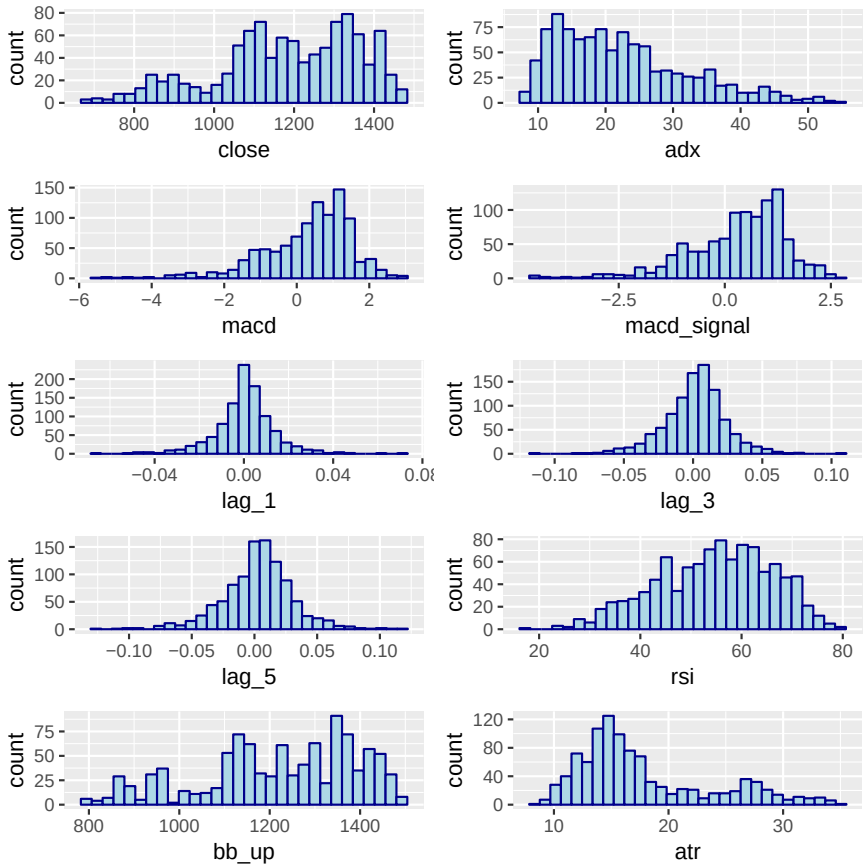

Figura 4.5: Histogramas de frecuencia de indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

En la figura 4.5 se observa histogramas de frecuencias para cada indicador. Se observa como la distribución de los rezagos es similar entre ellos, mostrando datos simétricos con picos en los valores cercanos a 0. En cuanto a la variable ATR se observa que los datos son asimétricos hacia la derecha con algunos casos atípicos cercanos a 30, mostrando que el índice en general es de baja volatilidad. Para el caso de la banda de bollinger, sigue una distribución muy parecida a la del valor de cierre de precio, una distribución asimétrica hacia la izquierda con algunos picos en la cola, lo que referencia una tendencia alcista. Por otro lado el ADX muestra una distribución asimétrica hacia la derecha con valores entre 10 y 20 lo cual indicaría que esta tendencia alcista del índice es constante pero lenta.

La distibucion de la variable RSI es asimétrica hacia la izquierda con la mayor frecuencia entre los valores cercanos a 60, lo cual reforzaría lo indicado por los otros indicadores en cuanto a la tendencia del índice

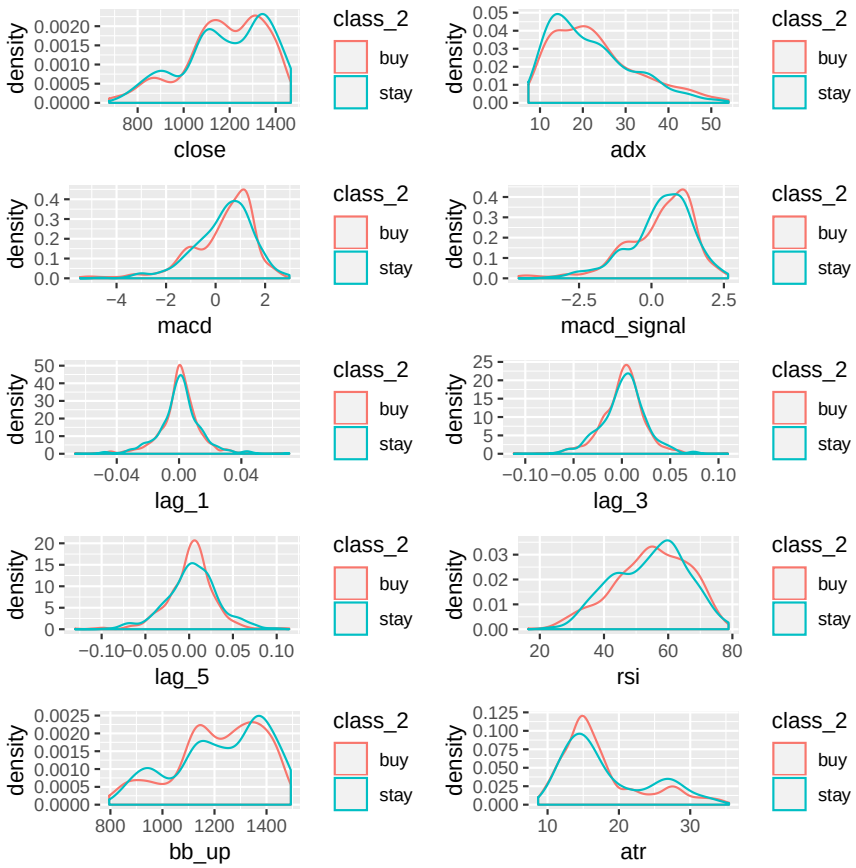

Figura 4.6: Gráfico de densidad de indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

En la figura 4.6 se observan gráficos de densidad para cada variable diferenciando entre clases, buscando algún comportamiento diferenciador. Sin embargo se muestra como para todos los indicadores la densidad es muy similar entre clases.

4.1.3. Resultados del ACP

A continuación se analizan los resultados de los componentes arrojados por el modelo en el primer período de entrenamiento (2009-2012) utilizando el índice S&P500, en esta sección se referirá a ésta como 'matriz de datos'. Ahora bien, si se quisiera analizar cada una de las componentes arrojadas en cada muestra de entrenamiento se necesitaría repetir este análisis 30 veces lo cual no es práctico para los fines de la investigación. En este sentido se elaboró una aplicación con el paquete Shiny de R, para visualizar de manera interactiva las gráficas que ayudan a entender los componentes. La aplicación puede ser visitada con el siguiente enlace https://rodserr.shinyapps.io/trading-ML/.

Los eigenvalores miden la cantidad de variación retenida por cada componente. Los

eigenvalores son mayores para los primeros componentes, dado que el primer componente busca maximizar la cantidad de variación de la matriz de datos, por lo que cada vez es menor la cantidad de variación retenida por cada componente.

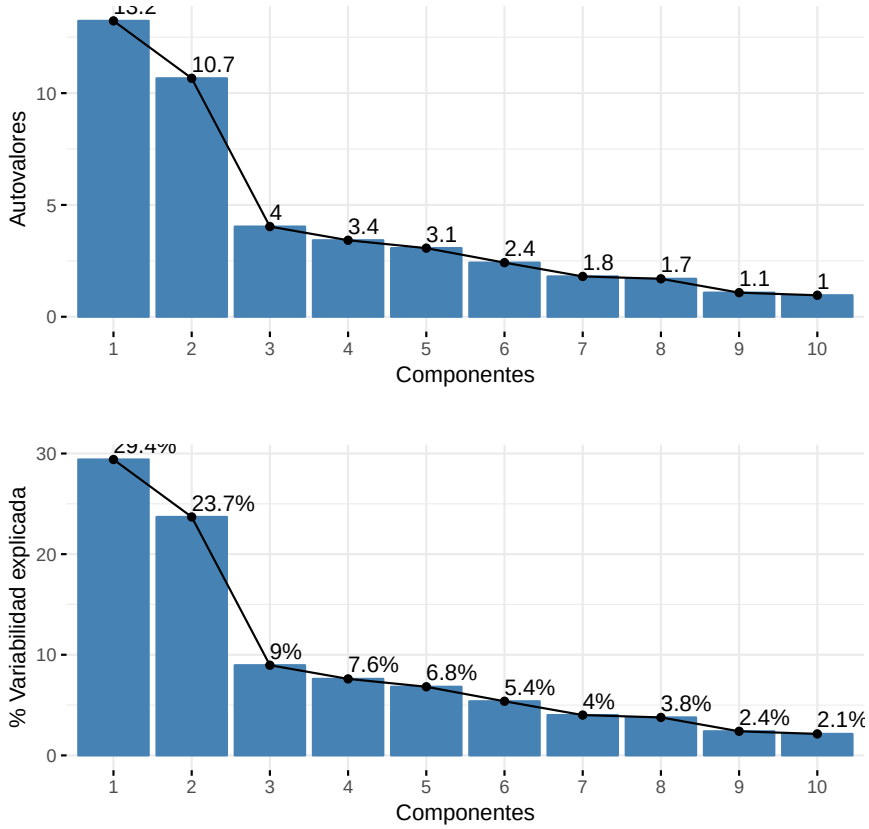

Figura 4.7: Eigenvalores y Porcentaje de contribución para los 10 componentes más importantes obtenidos por la matríz de datos

La proporción de variación explicada por cada eigenvalor viene dada de dividir cada eigenvalor por su sumatoria, en este caso 45 -el número de variables originales-. Un eigenvalor mayor que 1 indica que el componente tiene mayor variación que la contenida en una de las variables originales. En la figura 4.7 se puede observar que el 85 % de la variación está contenida en los primeros 7 componentes. Igualmente se aprecia que el eigenvalor de los 10 PCs es mayor que 1.

El número de componentes a utilizar se establece en función a estos dos gráficos. Este comportamiento anteriormente descrito se replica en los demás períodos de entrenamiento y demás pares. Se decide utilizar los dos primeros componentes como variables explicatorias en el MLG ya que la diferencia entre el 2do y 3er componentes es significativa, por lo que se espera que la variabilidad explicada en los dos primeros componentes -aproximadamente 50 %- sea suficiente. Esto también se basa en el hecho de que los indicadores son transformaciones del

precio del activo por lo que existe correlación entre ellos. Se asume entonces que no es necesaria toda la información provista por los indicadores.

La contribución de las variables representan la variabilidad contenida en un componente. Las variables correlacionadas con el componente principal 1 (PC1) y PC2 son las más importantes en explicar la variabilidad en la matriz de datos. Aquellas que no se correlacionan con ninguna componente son desechadas por su baja contribución. En la figura 4.8 se observa la contribución de las primeras 30 variables en PC1, PC2 y la contribución obtenida en ambas.

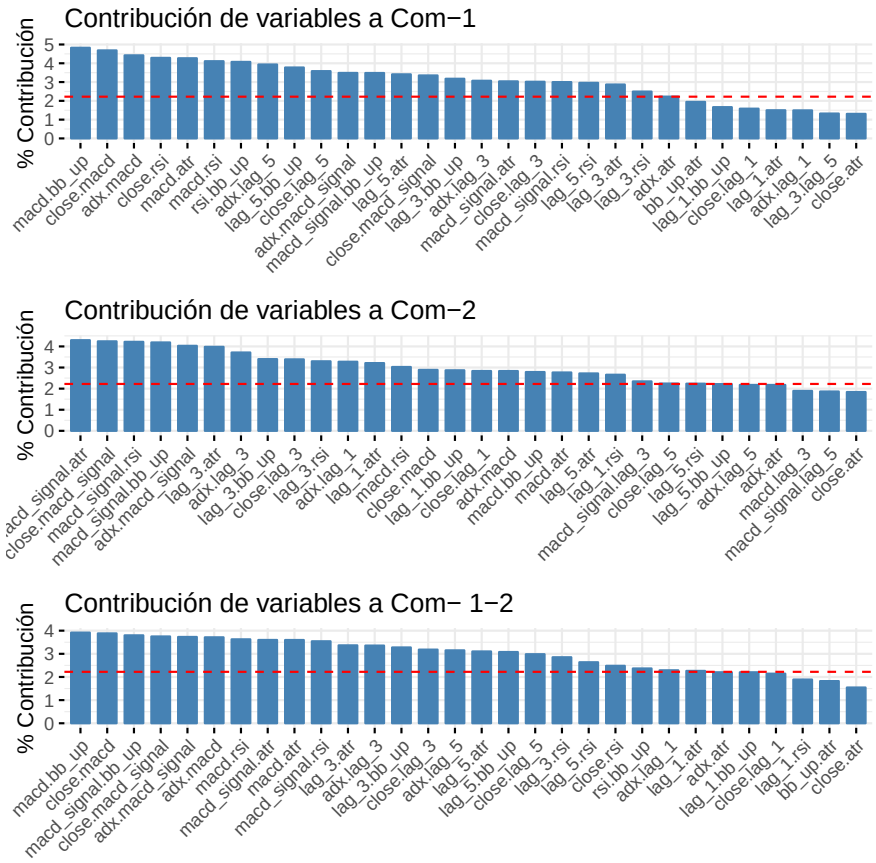

Figura 4.8: Contribución de cada variable para PC1, PC2 y el total de la contribución en ambos componentes

La línea roja indica el promedio esperado de contribución si las variables fueran uniformes, es decir $\frac{1}{NdeVariables} = \frac{1}{45} = 2,2\,\%$. Una variable sobre este umbral se considera importante en la contribución al componente. Se aprecia cómo las interacciones que predominan en ambos componentes están relacionadas con el indicador MACD.

La calidad de representación en el gráfico viene dada por el valor de Cos^2, el cual se refiere a la importancia que tiene la variable para interpretar el componente. Para una variable la suma de Cos^2 en todas las componentes equivale a 1. En la figura 4.9 se muestra los valores

de Cos^2 para las primeras 2 componentes

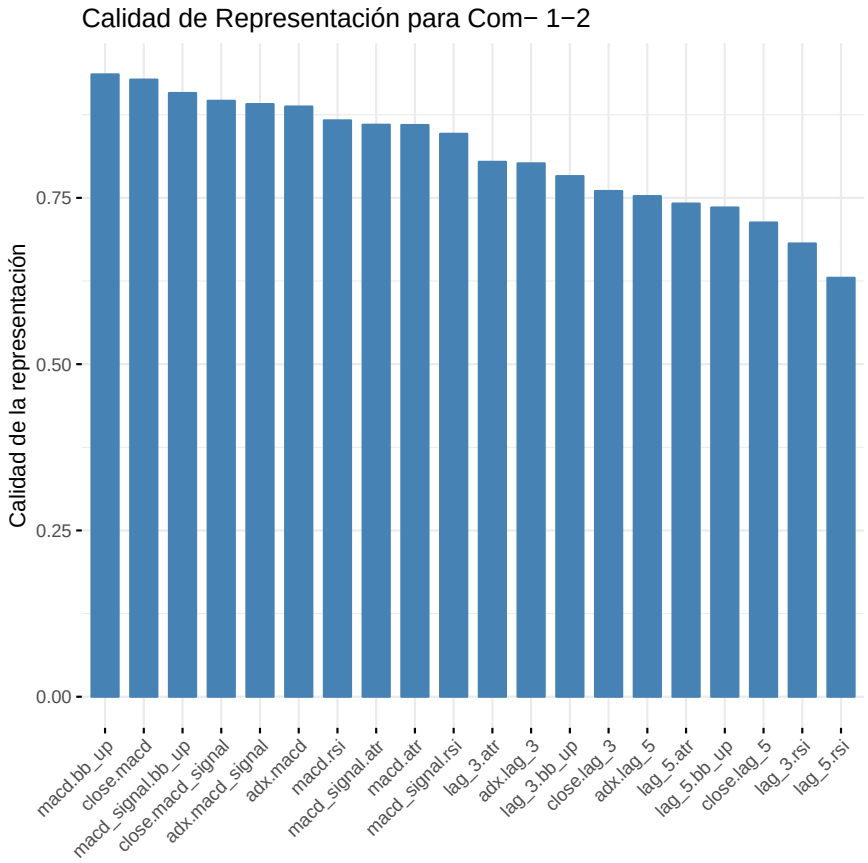

Figura 4.9: Calidad de representación medida por Cos^2 de cada variable en PC1 y PC2

El gráfico de correlación ó Factor map muestra la relación entre las variables. Las claves para su interpretación son:

- Las variables positivamente correlacionadas se encuentran agrupadas entre sí
- Las variables negativamente correlacionadas se posicionan en cuadrantes opuestos.
- La distancia entre las variables y el origen mide la calidad de representación de las variables en el gráfico. Mientras más alejado del origen, mejor representadas

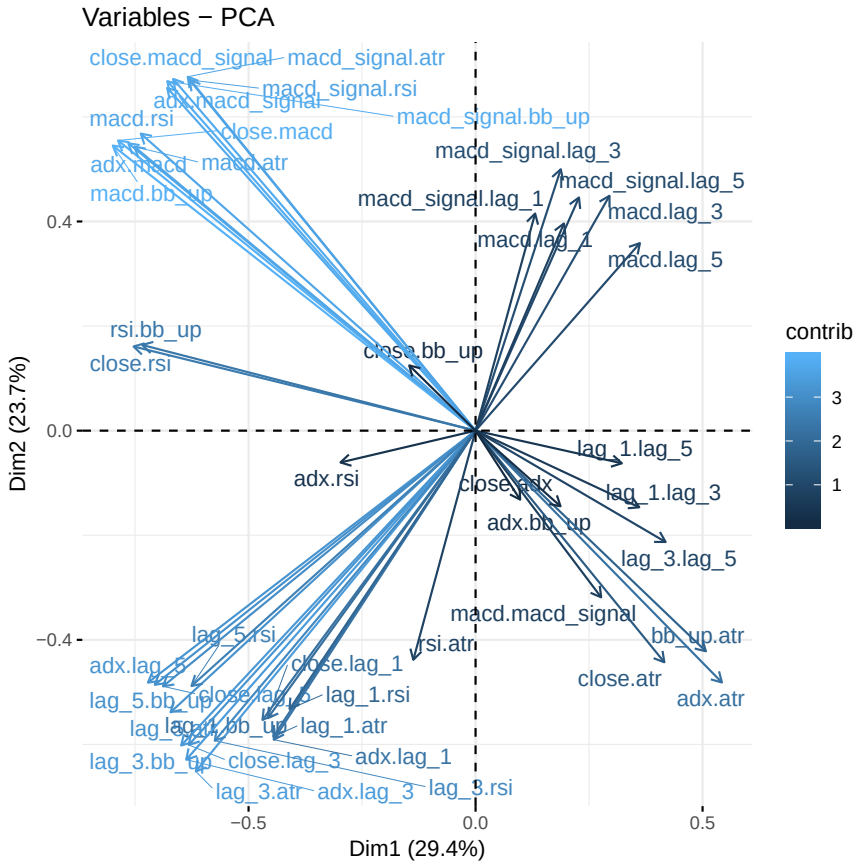

Figura 4.10: Gráfico de Correlación entre PC1 y PC2

En la figura 4.10 se observa el gráfico de correlación para PC1 y PC2, el color de cada variable viene dado por su contribución, mientras más oscuro menor es su contribución a los componentes. Se puede apreciar que las variables con mayor contribución están agrupadas por dos tipos de indicadores predominantes, en el cuadrante superior izquierdo aparecen variables constituidas por interacciones con los indicadores del MACD, mientras que en el cuadrante inferior izquierdo los indicadores predominantes son los rezagos. Se podría resumir que un cuadrante representa la información actual del activo mientras que el otro, la relación entre el precio actual y el precio en períodos anteriores. Por otro lado los grupos forman un ángulo de 90° por lo que no están correlacionados. En los cuadrantes positivos para el eje x predominan variables con una menor contribución a los componentes.

Figura 4.11: Gráfico Dispersión entre los componentes. Los puntos rojos representan las observaciones marcadas como 'buys', los triángulos azules los 'stays'

En la figura 4.11 se muestra el gráfico de dispersión por clases. Se observa que no hay una región que agrupe una sola clase sino que las observaciones están dispersas en todos los cuadrantes, por lo que no es posible hacer un agrupamiento eficientemente.

4.2. Coeficientes del modelo

En el presente capítulo se realiza la descripción de los resultados obtenidos después de la aplicación del método propuesto para la estrategia. De igual modo, se presentan los resultados arrojados por las pruebas de Backtesting simulando las entradas y salidas. Los nombres de los componentes fueron sustituidos por etiquetas que intentan explicar la representación del componente en el modelo con ayuda de las gráficas mostradas en la sección 4.1.3.

```
> pca$x[,1:2] %>% cor()
```

```
             PC1          PC2
PC1 1.000000e+00 1.479632e-16
PC2 1.479632e-16 1.000000e+00
```

```
> list_model[[1]][[1]]$finalModel %>% car::durbinWatsonTest()
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.6562454     0.6855534       0
 Alternative hypothesis: rho != 0
```

```
> list_model[[1]][[6]]$finalModel %>% car::durbinWatsonTest()
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.7068565      0.585374       0
 Alternative hypothesis: rho != 0
```

```
> list_model[[1]][[1]]$finalModel %>% lmtest::dwtest()
```

```
        Durbin-Watson test

data:  .
DW = 0.68143, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

```
> list_model[[1]][[6]]$finalModel %>% lmtest::dwtest()
```

```
        Durbin-Watson test

data:  .
DW = 0.57755, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

En la tabla 4.1 se describen los resultados de los parámetros arrojados por la regresión logística en los 6 períodos de entrenamiento para la serie del S&P500. Se muestra para cada variable su respectivo coeficiente, así como su error estándar. Se realiza también el contraste Wald - Chi-Cuadrado para verificar la significancia de cada variable en el modelo. Por lo que para cada variable se expone estadístico z que viene dado por $\beta_i/SE(\beta_i)$ y su respectivo p-valor.

El test Wald - Chi-Cuadrado se puede definir de la siguiente manera:

- **H0**: $\beta_i = 0$
- **Hi**: $\beta_i \neq 0$

Tabla 4.1: Resumen del modelo para cada período de entrenamiento utilizando S&P500

Período de entrenamiento 2009 - 2012

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.36	0.06	-5.56	0.00
ATR.Pos-Macd.Neg	-0.01	0.02	-0.81	0.42
Macd.Pos-Rezago.Neg	-0.01	0.02	-0.71	0.48

Período de entrenamiento 2009 - 2013

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.41	0.06	-7.04	0.00
ATR.Pos-Macd.Neg	0.02	0.02	1.33	0.18
Macd.Pos-Rezago.Neg	-0.01	0.02	-0.46	0.65

Período de entrenamiento 2009 - 2014

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.28	0.05	-5.29	0.00
ATR.Pos-Macd.Neg	0.05	0.02	2.93	0.00
Macd.Pos-Rezago.Neg	0.02	0.02	1.39	0.16

Período de entrenamiento 2009 - 2015

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.17	0.05	-3.51	0.00
ATR.Pos-Macd.Neg	-0.06	0.02	-4.31	0.00
Macd.Pos-Rezago.Neg	0.02	0.02	1.21	0.23

Período de entrenamiento 2009 - 2016

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.13	0.04	-2.80	0.01
ATR.Pos-Macd.Neg	-0.06	0.01	-4.53	0.00
Macd.Pos-Rezago.Neg	0.03	0.01	1.76	0.08

Período de entrenamiento 2009 - 2017

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.11	0.04	-2.67	0.01
ATR.Pos-Macd.Neg	0.06	0.01	4.80	0.00
Precio.Pos-Rezago.Neg	0.03	0.01	2.18	0.03

Se observa que recién para el 3er período de entrenamiento se obtiene un p-valor menor a 0.05 lo cual indica que el coeficiente es significativo. A medida que la data de entrenamiento es más grande, los coeficientes son más significativos. Se infiere entonces que mientras más observaciones para entrenar el modelo, mayor será la asociación entre los componentes y la capacidad de predecir el retorno objetivo.

También se observa que el signo de los coeficientes varía en las distintas datas de entrenamiento, por ejemplo para el período 2009-2014 ambas variables influyen positivamente en la correcta predicción del retorno objetivo. En los períodos 2009-2015 y 2009-2016 la variable ATR.Pos-Macd.Neg influye negativamente en la variable dependiente caso contrario para la variable Macd.Pos-Rezago.Neg, que influye positivamente.

4.3. Resultados de la simulación

Tabla 4.2: Resumen de resultados de aplicar el modelo en la data de prueba para los 5 índices

	S&P_500	NASDAQ	NIKKEI_225	FTSE_100	BOVESPA
False Buys	62	72	116	58	141
True Buys	80	122	153	72	170
N° trades	142	194	269	130	311
Accuracy	56.34 %	62.89 %	56.88 %	55.38 %	54.66 %
Net Profit	5939.81	9101.82	3162.62	1869.01	-1250.00
Max Drawdown	1.07 %	1.29 %	3.49 %	2.20 %	5.69 %

En la tabla 4.2 se muestran los resultados de la simulación para cada uno de los índices. El número de trades cerrados es mayor en los índices BOVESPA y NIKKEI, lo que puede deberse a que estos mercados tuvieron una mayor volatilidad en el período de estudio. Por otra parte la predicción ronda entre 54 % al 62 %, la relación pérdida/ganancia de los parámetros utilizados es $2.5/2 = 1.25$, es decir que por cada trade negativo se necesita 1.25 trades positivos para mitigar la pérdida. En este sentido una precisión del 60 % asegura un margen de ganancia, sin embargo, el retorno acumulado obtenido es pobre comparado con inversiones pasivas del mismo índice. Además es bastante cercano al 50 % lo que supondría un comportamiento aleatorio del modelo.

En la figura 4.12 se observa los trades realizados por la simulación según el resultado de la operación, los trades verdes son aquellos clasificados como 'True buys' y resultaron en ganancia, los rojos, son clasificados como 'False buys' y resultaron en pérdidas y los azules son clasificados como 'False buys' pero cerraron el trade por límite de tiempo.

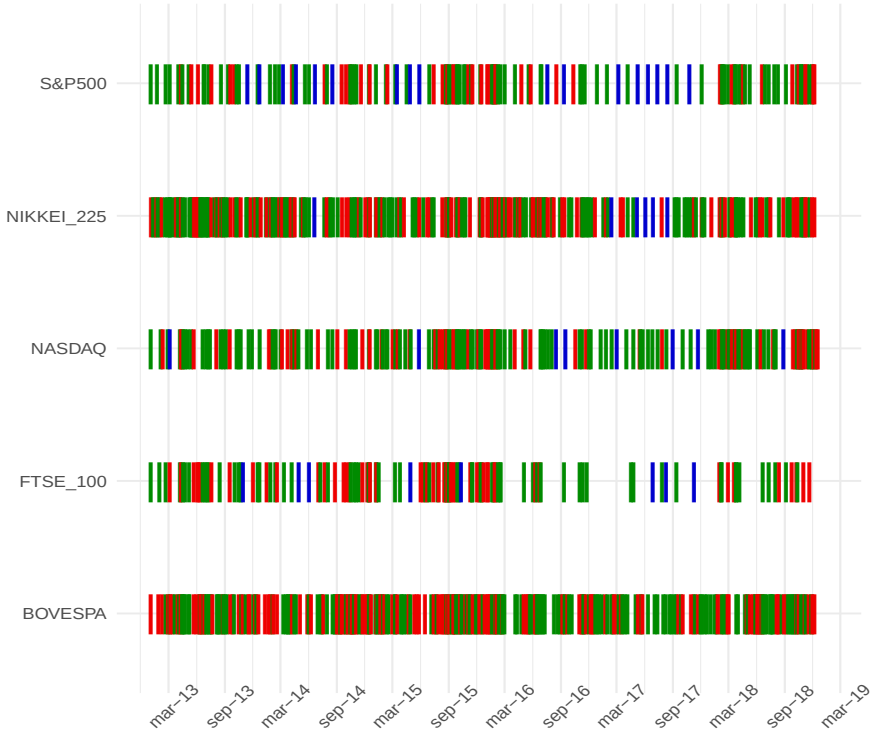

Figura 4.12: Clasificación de los trades

Se observa como en la simulación utilizando el índice BOVESPA, los trades positivos aumentan su frecuencia a partir del segundo semestre del 2016, esto puede deberse al hecho de tener mayor número de observaciones para entrenar el modelo. Igualmente se aprecia como para el S&P500 las operaciones se concentran en los primeros años de prueba, cerrando los demás años prácticamente sin operaciones.

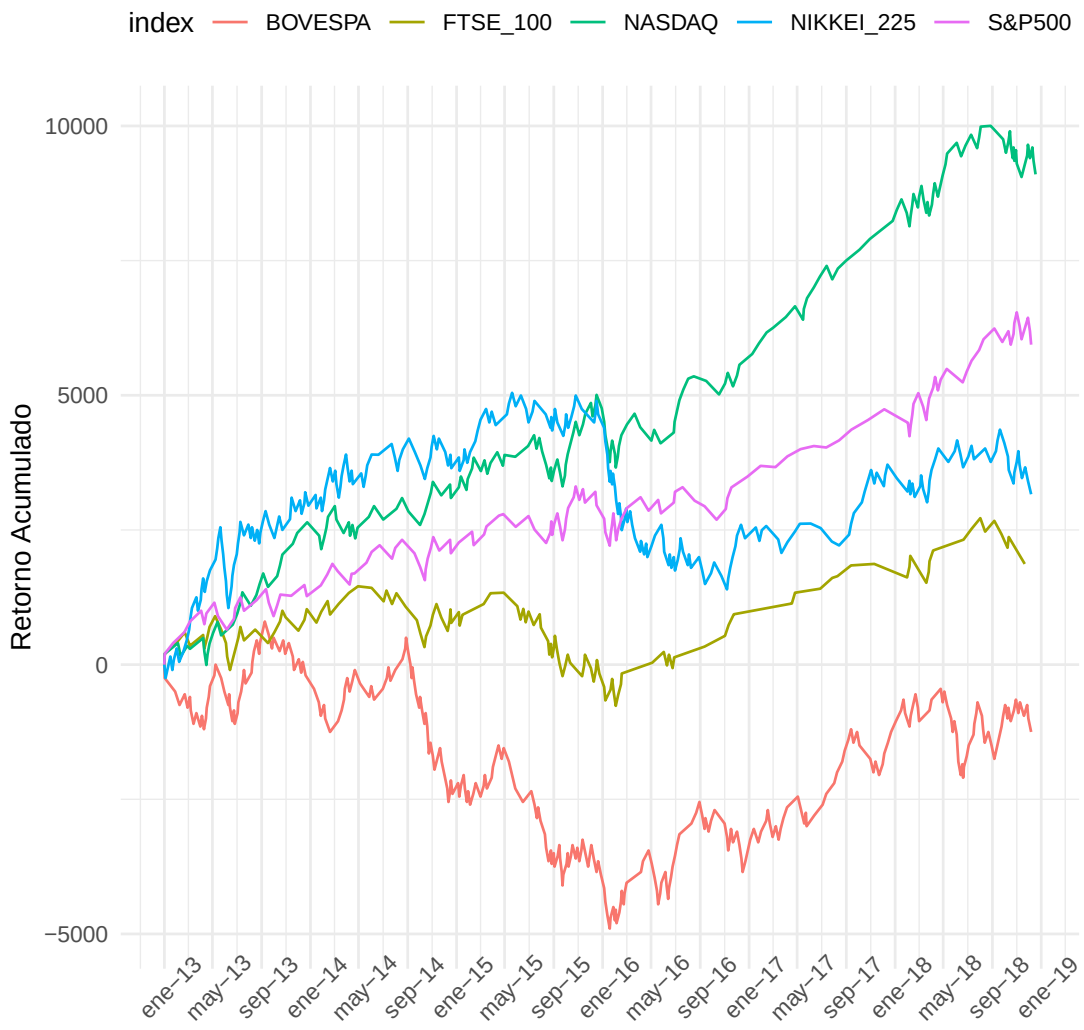

Figura 4.13: Retorno acumulado para cada índice

En la figura 4.13 se observa la curva de capital de las operaciones realizadas con cada uno de los índices. EL NASDAQ y S&P500 son los de mejor desempeño, mostrando ganancias consistentemente durante todo el período de entrenamiento. Por su lado el NIKKEI y FTSE tienen momentos tanto de ganancia como de pérdida, pero manteniendo un rendimiento positivo. Mientras que el BOVESPA pierde consistentemente durante la primera mitad del período y gana

en la segunda mitad, pero terminando con un rendimiento positivo.

4.3.1. Medidas de Riesgo

Al asumir que siempre se abre la posición con la misma cantidad de dinero, en este caso 10.000 USD, se sabe que los trades sólo pueden arrojar dos resultados -0.02 % de ganancia en caso que sea positivo ó 0.025 % de pérdida en caso contrario, descartando las liquidaciones por límite de tiempo- En este sentido se aplica el contraste Wald–Wolfowitz comúnmente llamado test de racha, para verificar la aleatoriedad de los resultados de los trades.

La prueba de Wald–Wolfowitz se puede definir de la siguiente manera:

- **H0**: La secuencia es producida de manera aleatoria.
- **Hi**: La secuencia no es producida de manera aleatoria.

Tabla 4.3: Resultados del test de Wald–Wolfowitz (Test de Racha)

	statistic	p.value
S&P_500	-0.28	0.78
NASDAQ	0.60	0.55
NIKKEI_225	1.27	0.20
FTSE_100	0.37	0.71
BOVESPA	-0.25	0.81

Frente a p-valores mayores a 0.05, y con un nivel de significación del 5 % no existen elementos suficientes para rechazar la hipótesis nula de aleatoriedad en la secuencia de los resultados de los trades, por lo que se puede concluir que los trades son independientes. Esta independencia permite asumir que la suma de las variables al tener una muestra suficientemente grande, se distribuye N(0, 1). Se calcula entonces el VaR y ES para cada una de las estrategias.

Tabla 4.4: VaR y ES para retornos de cada índice

	S&P_500	NASDAQ	NIKKEI_225	FTSE_100	BOVESPA
VaR	6,297.49	5,622.23	6,246.23	6,386.23	6,451.87
ES	7,912.80	7,195.66	7,859.20	8,005.20	8,073.22

Se puede observar en la tabla 4.4 como el VaR para cada estrategia varía en función de la precisión del modelo, como es de esperar. Para los índices con mayor precisión el VaR es menor y por lo tanto también lo es el ES. En este sentido se puede interpretar el VaR y ES del S&P500 como: 'Cuando la estrategia toma como activo el índice S&P500, existe una probabilidad del 5 % de que genere una pérdida igual ó mayor a 6,297.49 USD luego de 300 trades realizados. En caso de que ocurra una pérdida mayor, se espera que el déficit total sea de 7,912.80 USD'

Conclusiones y Recomendaciones

En esta investigación se ha planteado un marco de trabajo que permite el desarrollo y prueba para una estrategia de trading automatizado. Se busca demostrar que es posible obtener rendimientos con una estrategia basada en indicadores técnicos utilizados como variables predictoras en un modelo de aprendizaje automático. En los resultados no se contemplan las comisiones acarreadas por la operación de los activo. Se asume también que siempre se logra comprar la cantidad establecida en el precio de cierre de la vela, hecho que no siempre ocurre sobre todo en mercados de alta volatilidad y poca liquidez.

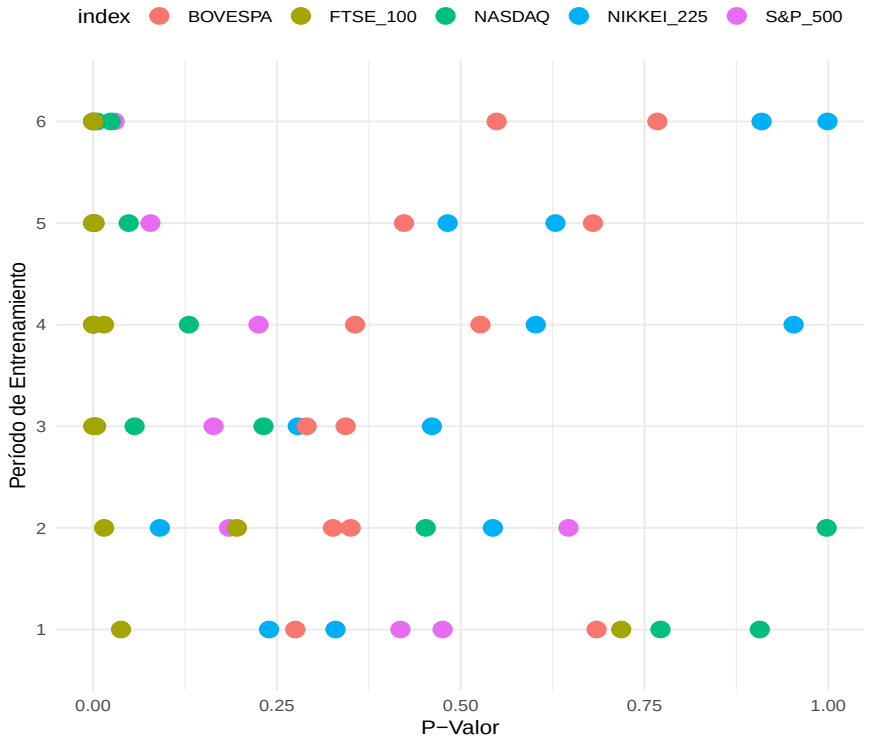

Figura 4.14: Gráfico de Disperción de los p-valores vs período de entrenamiento

Lo primero que se debe destacar es que en los primeros períodos de entrenamiento los coeficientes no son significativos según el test Wald-Chi-Cuadrado. Sin embargo al aumentar el tamaño de los datos de entrenamiento la significancia aumenta en la mayoría de los índices, esto se puede ver reflejado en la figura 4.14. Se observa que los p-valores de los coeficientes de los índices FTSE, NASDAQ y S&P disminuyen al utilizar mayor data de entrenamiento, caso contrario a los coeficientes de los índices NIKKEI y BOVESPA que pareciera aumentar.

Por otro lado al observar los gráficos de correlación entre los componentes, en la mayoría de los casos se observan dos grupos que aportan la mayor cantidad de información a los componentes; uno de los grupos está dominado por las interacciones de MACD, mientras que el otro por las interacciones de los rezagos. Normalmente estos grupos forman un ángulo de 90° pero poseen el mismo signo en cuanto al eje x -componente con mayor información-. Esto deja ver que aún utilizando un número considerable de indicadores, la información que proporcionan al modelo no es suficiente para obtener una precisión aceptable.

Si bien el retorno acumulado durante 6 años de prueba es poco atractivo, se sabe que en la práctica los fondos utilizan los futuros como activos de especulación, estos permiten un apalancamiento en el riesgo, lo cual se traduce en mover altos volúmenes de dinero con una posición menor que la que se necesitaría en el mercado de acciones. De cualquier manera los resultados de la precisión dejan ver la posibilidad de un amplio rango de mejora en el performance de la estrategia.

Este modelo puede ser mejorado de muchas maneras, incluyendo, por ejemplo la optimización de los parámetros target, stop y horizonte para el activo a operar, así como la elección del tamaño de los períodos de entrenamiento ó el número de componentes a utilizar como variables predictoras. Otras de las limitaciones presente es el de utilizar un modelo lineal, si bien se decide utilizar MLG como punto de partida es muy probable que algún modelo que no asuma linealidad como Bosques Aleatorios, SVM ó Redes Neuronales mejoren la predicción.

El modelo solo considera la prediccion del incremento del precio. Se utiliza la figura del stop loss como reduccion del riesgo. Un alternativa para obtener protección podría ser invertir el modelo, es decir predecir una disminución del precio, de esta manera se podría dejar de lado el stop loss y liquidar la posición cuando el modelo prediga una disminución en los precios.

Otra posible fuente de optimización podría ser la elección de los indicadores técnicos. Se podría profundizar en el aspecto técnico para su elección y el de las configuraciones de los mismos buscando mejorar la predicción del modelo. De cualquier manera esta investigación busca ser un punto de partida para futuras investigaciones.

Capítulo 5

Anexos

5.1. Coeficientes de los modelos

Tabla 5.1: Resumen del modelo para cada período de entrenamiento utilizando NASDAQ

Período de entrenamiento 2009 - 2012

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3127	0.0638	-4.90	0.0000
PC1	-0.0021	0.0179	-0.12	0.9068
PC2	0.0058	0.0201	0.29	0.7718

Período de entrenamiento 2009 - 2013

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3667	0.0574	-6.39	0.0000
PC1	0.0124	0.0164	0.75	0.4524
PC2	0.0001	0.0184	0.00	0.9978

Período de entrenamiento 2009 - 2014

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3027	0.0522	-5.80	0.0000
PC1	0.0291	0.0153	1.91	0.0565
PC2	-0.0203	0.0170	-1.20	0.2318

Período de entrenamiento 2009 - 2015

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2748	0.0483	-5.69	0.0000
PC1	0.0524	0.0148	3.55	0.0004
PC2	-0.0238	0.0158	-1.51	0.1303

Período de entrenamiento 2009 - 2016

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2641	0.0451	-5.85	0.0000
PC1	-0.0426	0.0136	-3.13	0.0018
PC2	0.0291	0.0148	1.97	0.0484

Período de entrenamiento 2009 - 2017

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3420	0.0428	-8.00	0.0000
PC1	0.0358	0.0129	2.78	0.0055
PC2	-0.0323	0.0143	-2.26	0.0240

Tabla 5.2: Resumen del modelo para cada período de entrenamiento utilizando NIKKEI 225

Período de entrenamiento 2009 - 2012

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2532	0.0645	-3.92	0.0001
PC1	-0.0169	0.0173	-0.97	0.3297
PC2	-0.0255	0.0217	-1.18	0.2393

Período de entrenamiento 2009 - 2013

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3273	0.0580	-5.64	0.0000
PC1	-0.0260	0.0154	-1.69	0.0908
PC2	-0.0118	0.0195	-0.61	0.5438

Período de entrenamiento 2009 - 2014

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3303	0.0529	-6.24	0.0000
PC1	-0.0154	0.0142	-1.08	0.2784
PC2	-0.0130	0.0177	-0.74	0.4609

Período de entrenamiento 2009 - 2015

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3038	0.0489	-6.21	0.0000
PC1	-0.0008	0.0135	-0.06	0.9529
PC2	-0.0087	0.0166	-0.52	0.6022

Período de entrenamiento 2009 - 2016

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2830	0.0457	-6.20	0.0000
PC1	0.0091	0.0129	0.70	0.4823
PC2	-0.0075	0.0154	-0.48	0.6290

Período de entrenamiento 2009 - 2017

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2731	0.0427	-6.39	0.0000
PC1	-0.0000	0.0120	-0.00	0.9989
PC2	0.0016	0.0145	0.11	0.9093

Tabla 5.3: Resumen del modelo para cada período de entrenamiento utilizando FTSE 100

Período de entrenamiento 2009 - 2012

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1520	0.0633	-2.40	0.0164
PC1	-0.0369	0.0178	-2.08	0.0379
PC2	-0.0072	0.0199	-0.36	0.7184

Período de entrenamiento 2009 - 2013

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1554	0.0567	-2.74	0.0061
PC1	0.0390	0.0160	2.43	0.0150
PC2	0.0234	0.0181	1.29	0.1956

Período de entrenamiento 2009 - 2014

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0390	0.0518	-0.75	0.4517
PC1	-0.0570	0.0149	-3.82	0.0001
PC2	-0.0481	0.0167	-2.88	0.0040

Período de entrenamiento 2009 - 2015

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0475	0.0481	0.99	0.3228
PC1	-0.0724	0.0140	-5.19	0.0000
PC2	-0.0369	0.0151	-2.44	0.0149

Período de entrenamiento 2009 - 2016

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0138	0.0449	-0.31	0.7592
PC1	-0.0648	0.0130	-4.98	0.0000
PC2	-0.0444	0.0143	-3.10	0.0019

Período de entrenamiento 2009 - 2017

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0372	0.0424	0.88	0.3809
PC1	-0.0730	0.0125	-5.86	0.0000
PC2	-0.0520	0.0137	-3.81	0.0001

Tabla 5.4: Resumen del modelo para cada período de entrenamiento utilizando BOVESPA

Período de entrenamiento 2009 - 2012

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2403	0.0641	-3.75	0.0002
PC1	-0.0190	0.0174	-1.09	0.2750
PC2	-0.0089	0.0219	-0.41	0.6849

Período de entrenamiento 2009 - 2013

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1491	0.0570	-2.61	0.0089
PC1	-0.0152	0.0154	-0.98	0.3262
PC2	-0.0183	0.0196	-0.93	0.3503

Período de entrenamiento 2009 - 2014

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1055	0.0517	-2.04	0.0414
PC1	-0.0134	0.0141	-0.95	0.3435
PC2	-0.0186	0.0176	-1.06	0.2905

Período de entrenamiento 2009 - 2015

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0741	0.0478	-1.55	0.1209
PC1	0.0120	0.0130	0.92	0.3564
PC2	-0.0103	0.0162	-0.63	0.5268

Período de entrenamiento 2009 - 2016

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1257	0.0447	-2.81	0.0049
PC1	0.0096	0.0120	0.80	0.4229
PC2	-0.0063	0.0154	-0.41	0.6798

Período de entrenamiento 2009 - 2017

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1851	0.0423	-4.38	0.0000
PC1	0.0068	0.0113	0.60	0.5489
PC2	-0.0043	0.0145	-0.30	0.7675

Lista de Referencias

- Lu Ning (2016). *A machine Learning Approach to Automated Trading*. USA, Coston College Computer Science.
- Huertas López Alejandro (2015). *Modelos Predictivos para el mercado FOREX*. España, Universidad de Murcia.
- Shagilla Kadida Ramadhani (2006). *Ab Analysis of Technical Trading Strategies*. Inglaterra, Leeds University Business School.
- Bach William y Nielsen Kasper (2018). *On Machine Learning Based Cryptocurrency Trading*. Dinamarca, Aalborg University.
- Araneda Hugo (2015). *Diseño e Implementación de un Sistema Automatizado Para Operar en el Mercado de Divisas Usando Reglas de Asociación*. Chile, Universidad de Chile.
- Fernández Cristián (2008). *Modelos Ocultos de Markov Aplicados al Reconocimiento de PAtrones del Análisis Técnico Búrsatil*. Argentina, Universidad Nacional de Córdoba.
- James Gareth, Witten Daniela, Hastie Trevor y Tibshirani Robert (2013). *An Introduction to Statistical Learning*. Springer.
- Abhijit Ghatak (2017). *Machine Learning with R*. Springer.
- Murphy, J.J. (1999). *Análisis Técnico de los Mercados Financieros*. New York Institute of Finance. New York: Gestión 2000.
- Danielsson, Jon (2011). *Financial Risk Forecasting*. Wiley.
- Pfaff, Bernhard (2016). *Financial Risk Modelling and Portfolio Optimization with R*. Wiley.
- González, Lisbeth (2002) *(*Análisis Financiero del Portafolio Mercantil de Inversión de MERINVEST. Período 1998-2000). Venezuela, Universidad Católica Andrés Bello
- Balestrini, Miriam (1997) *(*Como se Elabora el Proyecto de Investigación). BL Consultores y Asociados
- Rebolledo, Valeria (2018) *(*Evaluación de la Efectividad del Método GARCH-EVT-COPULAS para el Cálculo del VaR como Medida de Riesgo en Mercados de Commodities Latinoamericanos). Venezuela, Universidad Central de Venezuela