



UNIVERSIDAD CENTRAL DE VENEZUELA

FACULTAD DE CIENCIAS ECONÓMICAS Y SOCIALES  
ESCUELA DE ESTADÍSTICA Y CIENCIAS ACTUARIALES

EVALUACIÓN DE LA EFICACIA DE UNA  
ESTRATEGIA DE TRADING BASADA EN MODELO  
LINEAL GENERALIZADO

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

*Dedicado a mis padres y a mi hermana, por ser los pilares de mi vida y ser mi motivo de  
haber cumplido esta meta.*

# Agradecimientos

---

Agradecimientos

**Gracias**

# Índice general

---

<b>Índice de figuras</b>	<b>4</b>
<b>Índice de tablas</b>	<b>5</b>
<b>Introduccion</b>	<b>6</b>
<b>1. El Problema</b>	<b>7</b>
1.1. Justificación . . . . .	7
1.2. Planteamiento del Problema . . . . .	8
1.3. Objetivo General . . . . .	9
1.4. Objetivos Específicos . . . . .	9
<b>2. Marco Teórico</b>	<b>11</b>
2.1. Antecedentes . . . . .	11
2.1.1. Trading de Cryptomonedas basado en Aprendizaje Automatico . . . . .	11
2.1.2. Un enfoque de aprendizaje automático para el comercio automatizado . .	11
2.1.3. Modelos predictivos para el mercado FOREX . . . . .	11
2.1.4. Un Análisis de estrategias de trading técnico . . . . .	12
2.1.5. Modelos Ocultos de Markov Aplicados al Reconocimiento de Patrones del Análisis Técnico Bursátil . . . . .	12
2.1.6. Diseño e implementación de un sistema automatizado para operar en el mercado de divisas usando reglas de asociación . . . . .	12

2.2. Bases Teóricas . . . . .	12
2.2.1. Hipótesis del Mercado Eficiente . . . . .	12
2.2.2. Análisis Técnico . . . . .	13
2.2.3. Introducción al aprendizaje automático . . . . .	14
2.2.4. Regresión Logística . . . . .	15
2.2.5. Análisis de Componentes Principales como método de reducción de variables	17
2.2.6. Matriz de Confusión . . . . .	18
2.3. Bases Legales . . . . .	19
<b>3. Marco Metódico</b>	<b>20</b>
3.1. Análisis Exploratorio de los datos . . . . .	20
3.1.1. Datos OHLC y Fuente de los datos . . . . .	20
3.1.2. Series de Índices . . . . .	20
3.2. Entrenamiento del Modelo . . . . .	21
3.2.1. Variable dependiente . . . . .	21
3.2.2. Indicadores Técnicos como variables predictoras . . . . .	22
3.2.3. Validación Cruzada en Series de Tiempo . . . . .	24
3.2.4. WalkForward Backtesting . . . . .	25
3.2.5. Reducción de la dimensión con Análisis de Componentes Principales . . . .	26
<b>4. Análisis de Resultados</b>	<b>32</b>
4.1. Coeficientes del modelo . . . . .	32
4.2. Resultados de la simulación . . . . .	34
<b>Conclusiones y Recomendaciones</b>	<b>38</b>
<b>Lista de Referencias</b>	<b>39</b>

# Índice de figuras

---

2.1. Matriz de Confusión . . . . .	19
3.1. Precios de Cierre de los índices en el período de estudio (26/10/2008 - 18/01/2019)	21
3.2. Correlación entre indicadores originales calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012) . . . . .	23
3.3. Correlación entre indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012) . . . . .	24
3.4. Metodología WalkForward . . . . .	25
3.5. 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 $h$ . . . . .	26
3.6. Eigenvalores y Porcentaje de contribución para los 10 componentes más importantes obtenidos por la matriz de datos . . . . .	28
3.7. Contribución de cada variable para PC1, PC2 y el total de la contribución en ambos componentes . . . . .	29
3.8. Calidad de representación medida por $Cos^2$ de cada variable en PC1 y PC2 . . . .	30
3.9. Gráfico de Correlación entre PC1 y PC2 . . . . .	31
4.1. Clasificación de la simulación . . . . .	35
4.2. Retorno acumulado para cada índice . . . . .	36

## Índice de tablas

---

4.1. Resumen del modelo para cada período de entrenamiento utilizando S&P500 . .	32
4.2. Resumen de resultados de aplicar el modelo en la data de prueba para los 5 índices	35
4.3. Resultados del test de Wald–Wolfowitz (Test de Racha) . . . . .	37

# Introducción

---

Las economías de países latinoamericanos son reconocidas históricamente por depender en gran magnitud del comercio de sus materias primas, lo cual hace que el comercio con dichos commodities resulte de gran impacto para el gasto fiscal y para la balanza de pagos, esto deja como consecuencia, la necesidad de los actores económicos de estudiar el riesgo a profundidad para poder evitar resultados que reduzcan sus retornos positivos.



# El Problema

---

## 1.1. Justificación

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 compras y ventas 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 “Quantitative Funds” -Fondos Cuantitativos-. Según un informe realizado por Credit Suisse Group, estos fondos son los de mayor crecimiento 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. Es así como 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”.

La mayor ventaja del trading basado en algoritmos es 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 como 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 sobreoptimizar el sistema.

## 1.2. Planteamiento del Problema

Con este panorama es evidente la utilidad que supone la utilización 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 computacionales 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.

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.

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 las investigaciones similares a la presente, se toman algunos métodos y se intenta definir cuáles de ellos arroja mayor precisión. 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  $h$ . Es decir que en el intervalo de tiempo  $h$  debe ocurrir primero la toma de ganancia -take profit- que la pérdida -stop loss-. 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.

### 1.3. Objetivo General

Evaluar la eficacia de una estrategia de trading basada en técnicas de aprendizaje estadístico

### 1.4. Objetivos Específicos

- Definir los parámetros de la estrategia, así como los indicadores técnicos a utilizar como variables predictoras
- Utilizar Análisis de Componentes Principales como técnica de reducción de la dimensión de variables predictoras
- Aplicar Regresión Logística con las componentes arrojadas por el ACP
- Desarrollar la metodología Walkforward para tomar en cuenta el dinamismo del mercado en la estrategia

- Evaluar la predicción del modelo como estrategia en los índices seleccionados

## 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 diversos 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. Prueban los algoritmos en seis pares basados en Tether: Bitcoin, Ethereum, Binance Coin, NEO, Litecoin y Bitcoin Cash.

#### 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 serie de precios de 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 activo 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 como 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 Africa. El estudio demuestra como 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.

#### **2.1.6. Diseño e implementación de un sistema automatizado para operar en el mercado de divisas usando reglas de asociación**

### **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 a los 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 cegada de los individuos, formuladas por la manera en como estos responden ante nueva información. Algunas de las hipótesis que explican este razonamiento cegado son:

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 como estos comportamientos pueden resultar en ineficiencias del mercado.

Agunos académicos como Hong y Stein's (1999) categorizan a los inversores en *Informados* y *Noinformados*. 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 competires, 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.

### 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 extrer patrones de conducta. A la primera técnica se le llama Análisis Fundamental y el segundo Análisis Técnico. El Análisis Fundamental está mas 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 mas 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, la cual según murphy se repite. Ésta afirmación tiene también una estrecha relación en los ciclos económicos.

### 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  $X_1, X_2, \dots, X_j$ . Se asume que existe una relación entre  $Y$  y  $X = X_1, X_2, \dots, X_j$ , la cual puede ser escrita de forma general como

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

donde  $f$  es una función desconocida de  $X$  y  $\epsilon$  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 esencia, el aprendizaje automático refiere a un conjunto de enfoques para estimar  $f$

### Métodos Paramétricos vs No Paramétricos

La mayoría de los metodos 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 mas 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 numero de observaciones -muchas más que las necesarias por los métodos paramétricos- para obtener una estimación precisa. Además en general la iterpretación 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 acertivamente 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 asociacion buscan descubrir patrones inherentes que describan el comportamiento de los datos u observaciones, por ejemplo, un grupo de clientes que compren un producto  $r$  conjunto con otro producto  $s$ .

### **Regresión vs Clasificación**

Las variables pueden ser divididas entre cuantitativas ó cualitativas -también llamadas categóricas-. Las cuantitativas toman valores numericos 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.

### **Compensación entre sesgo y varianza**

#### **2.2.4. 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:

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

En donde  $X_j$  representa las variables predictoras y los coeficientes  $\beta_j$  cuantifican la asociación entre la variable predictora  $X_j$  y la variable respuesta  $Y$ . Por lo que se interpreta a  $\beta_j$  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:

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

Despejando se obtiene

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

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

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

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  $\beta$ . Sin embargo dado que la relación entre  $p(X)$  y  $X$  no es una línea recta,  $\beta$  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  $\beta$  en relación a  $Y$ .

### Máxima Verosimilitud

Los coeficientes  $\beta$  son desconocidos, por lo que deben estimarse en la data de entrenamiento. Para esto se utiliza el método de *Mxima Verosimilitud*, el cual consiste en estimar los coeficientes para los cuales la probabilidad de predicción para cada individuo, utilizando (formula 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*

### 2.2.5. Análisis de Componentes Principales como método de reducción de variables

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

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

En situaciones donde  $p$  es relativamente grande con relación a  $n$ , seleccionar un valor de  $M \ll p$  puede reducir considerablemente la varianzade 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

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

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

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

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

que contiene la mayor varianza. El primer vector de loadings de componente principal resuelve el problema de optimización

$$\max_{\phi_{11}, \dots, \phi_{p1}} \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1}x_{ij} \right)^2 \quad \text{sujeeto a } \sum_{j=1}^p \phi_{j1}^2 = 1$$

El problema de maximización en (formula 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  $X_1, \dots, X_p$  que tiene la máxima varianza de todas las combinaciones lineales que no están correlacionadas con  $Z_1$ . Así los scores del segundo componente principal toman la forma

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

donde  $\phi_2$  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  $\phi_2$  a ser ortogonal a la dirección de  $\phi_1$ .

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

### 2.2.6. Matriz 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.

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

**Figura 2.1:** Matriz de Confusión

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 *Precisin*.

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

## 2.3. Bases Legales

## Marco Metódico

---

### 3.1. Análisis Exploratorio de los datos

#### 3.1.1. Datos OHLC y Fuente de los datos

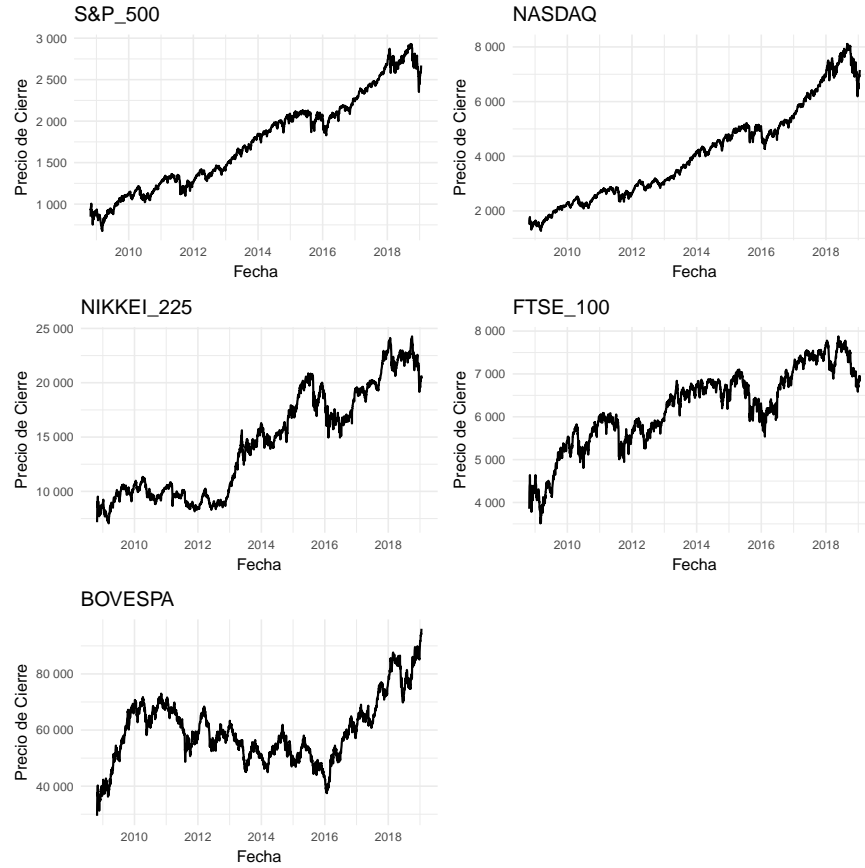
La estructura de los datos utilizados en el trabajo es de tipo OHLC por sus siglas en inglés Open, High, Low, Close. La misma, resume en 4 registros el comportamiento del precio del activo (Apertura, Cierre, Mínimo y Máximo) en un intervalo de tiempo. En el caso de la presente investigación, de un día. 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](http://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.

#### 3.1.2. Series de Índices

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.



**Figura 3.1:** Precios de Cierre de los índices en el período de estudio (26/10/2008 - 18/01/2019)

## 3.2. Entrenamiento del Modelo

### 3.2.1. 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 materializará 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.

### 3.2.2. Indicadores Técnicos como 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 como 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.

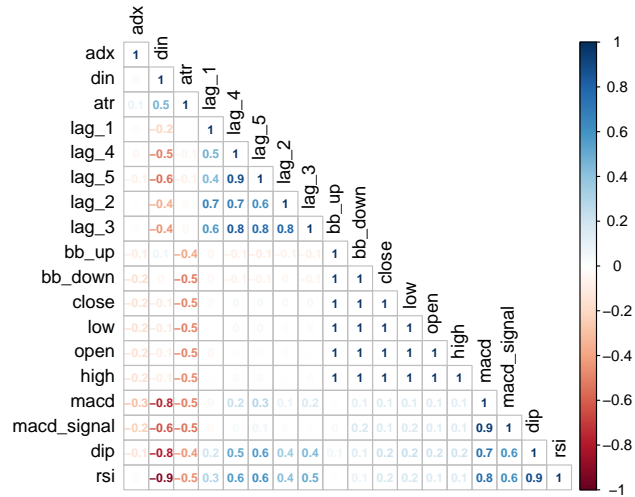
Lo que busca la investigación es utilizar la relación entre estos indicadores como variables independientes que ayuden al modelo a predecir oportunidades de entradas. En este sentido se asume la existencia de una dinámica local del mercado que puede ser predecida con ayuda de estos indicadores.

A continuación se presentan 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 MA(Moving Average) 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.

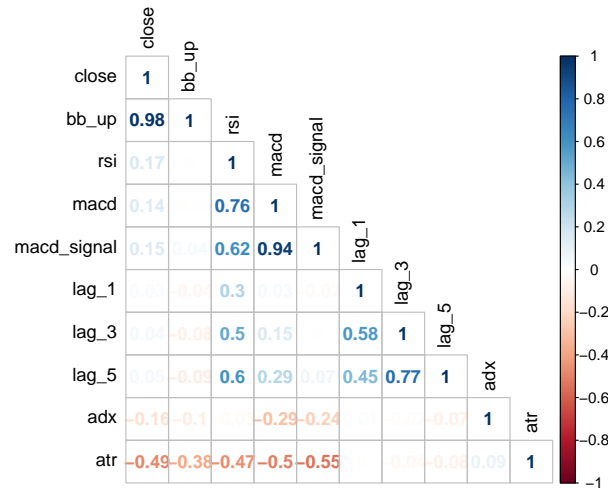




**Figura 3.2:** Correlación entre indicadores originales calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

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 – se muestra las correlaciones de las variables definitivas



**Figura 3.3:** Correlación entre indicadores definitivos calculados con los precios del S&P500 en el primer período de entrenamiento(01/01/2009 - 31/12/2012)

Ahora bien, la idea base de la investigación era utilizar los valores de cada indicador como variables predictoras. 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 detallará 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 sí solos, sino, las relaciones entre cada uno de ellos. Esto se abordó agregando al modelo las interacciones entre todos los indicadores, 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.

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

### 3.2.3. 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

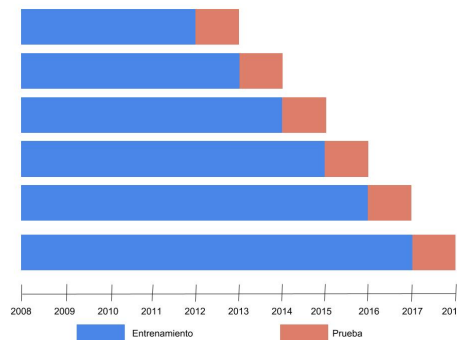
veces, y el resultado final es el promedio arrojado por cada una de las  $k$  validaciones.

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.

### 3.2.4. 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 testeaba 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, dado el dinamismo de los mercados bursátiles la estrategia no puede permanecer estática en el tiempo.

Para contrarrestar esta situación se optó por el método de *backtestingWalk forward*, 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 incluido 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 mas reciente con respecto al cual será implementado. En la presente figura se ilustra la metodología implementada.

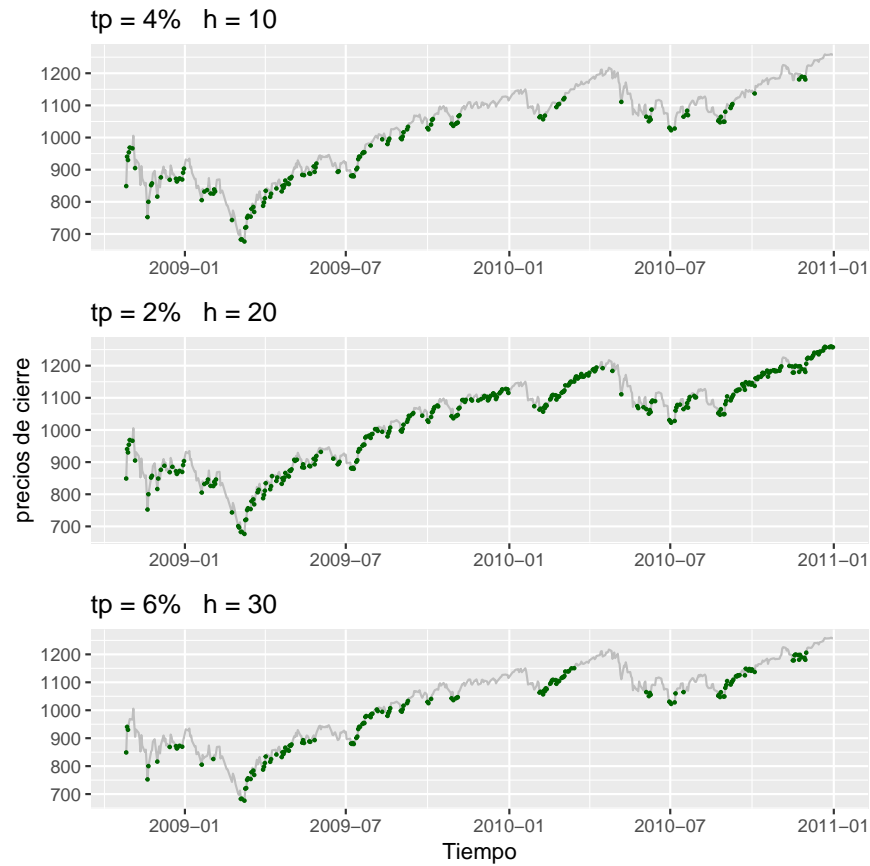


**Figura 3.4:** Metodología WalkForward

Otra de las características de la metodología que se modificó fue la elección de los parámetros óptimos. Previamente se utilizaba la data de validación para buscar la combinación de parámetros óptima. Ahora bien en la metodología de Walkforward se utilizan los mismo parámetros. A opinión del investigador al buscar los mejores parámetros se estaría incurriendo en un posible cesgo de sobreoptimización. El hecho de que en un año determinado unas configuraciones óptimas den los mejores resultados no asegura que se replique en el siguiente año.

La selección de los parámetros debería ser un estudio previo de la serie financiera a testear. Evidentemente al seleccionar un período de tiempo mas corto se obtendrán menos

observaciones que cumplan con el patrón por lo tanto se estaría en presencia de un problema de data imbalanceada que debe tener un tratamiento distinto. Por otro lado el utilizar un horizonte mayor no representa gran cambio en el número de ocurrencias, pero sí en el caso de que la transacción quede abierta -hay recordar que  $h$  representa una condición de salida para la estrategia de no ocurrir el target ni el stop loss-. La estrategia implementada en la investigación toma un horizonte de 20 períodos ya que este valor representa un umbral para la ocurrencia del objetivo.



**Figura 3.5:** 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  $h$

### 3.2.5. Reducción de la dimensión con Análisis de Componentes Principales

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 principales

puede ser igual al número de variables, la idea es utilizar un grupo reducido de componentes que maximizen 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 en la mayoría de los casos a 7 componentes donde las dos primeras contienen al rededor del 25 % de la variación, es importante recordar que esta reducción se realiza en cada período de entrenamiento.

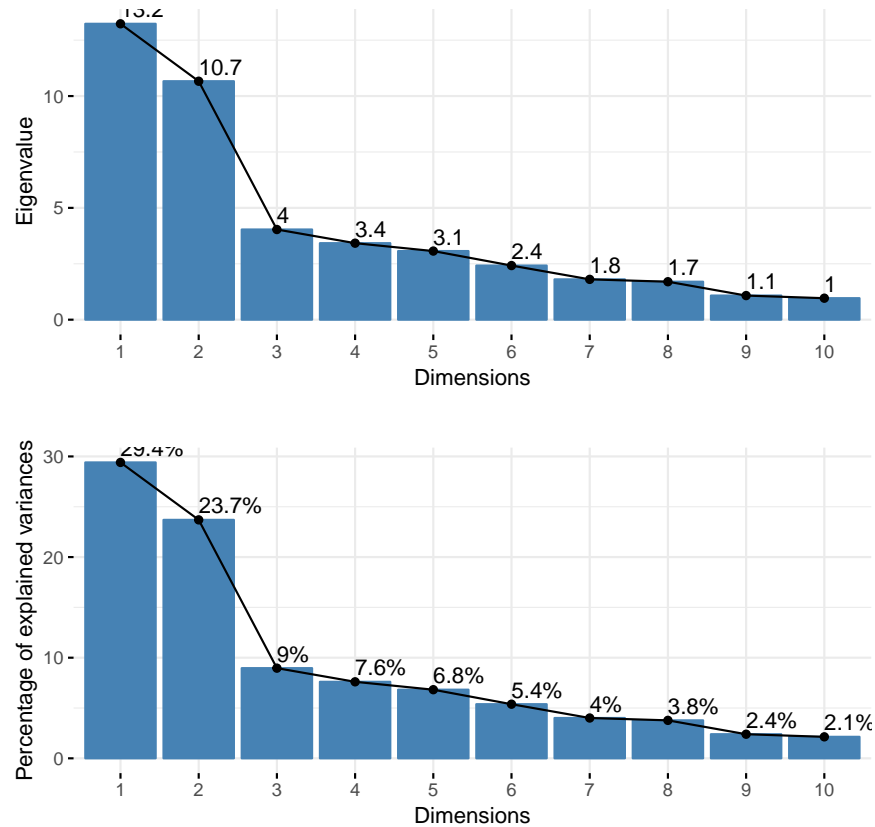
A continuación se analizan los resultados de los componentes arrojados por el modelo en el primer período de entrenamiento (2009-2012) utilizando el índice S&P500, en esta sección se referira a ésta como 'matriz de datos'.

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.

Los eigenvalores pueden usarse para establecer el número de componentes a utilizar. Para el modelo se estableció que en cada período de entrenamiento se utilizaría la cantidad de componentes necesarias para explicar el 85 % de la variación de la matriz de datos.

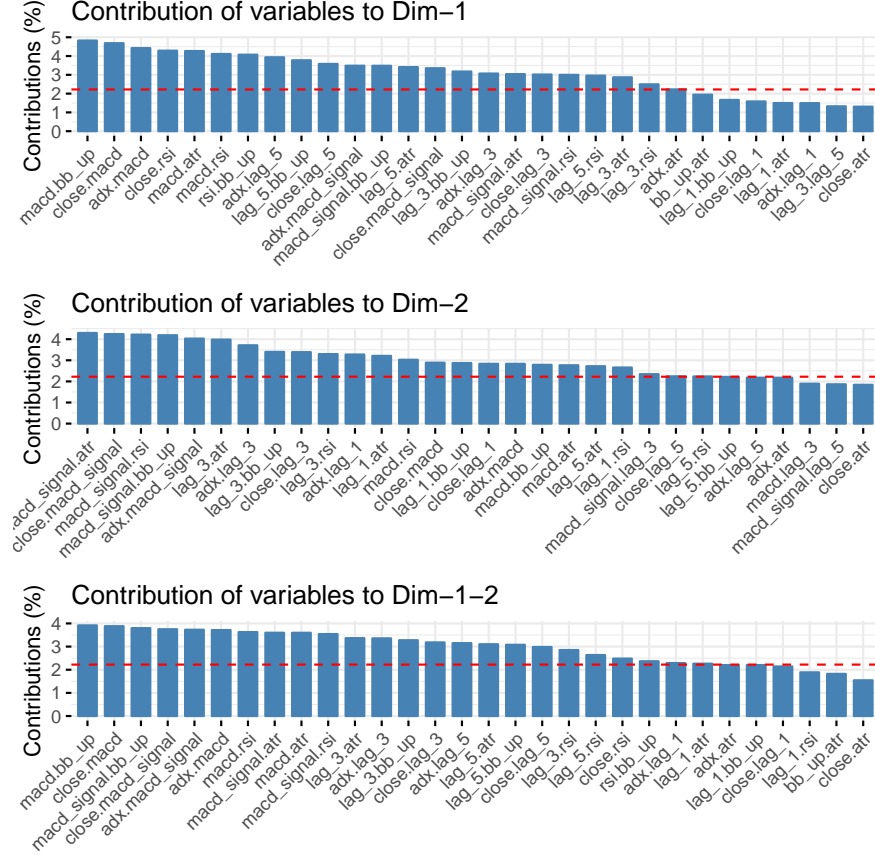
La proporción de variación explicada por cada eigenvalor viene dada de dividir cada eigenvalor por su sumatoria, en este caso 45 -el número de variables originales-.

Un eigenvalor mayor que 1 indica que el componente tiene mayor variación que la contenida en una de las variables originales. En la figura – Se puede observar que el 85 % de la variación esta contenida en los primeros 7 componentes. Igualmente se aprecia que el eigenvalor de los 10 PCs es mayor que 1.



**Figura 3.6:** Eigenvalores y Porcentaje de contribución para los 10 componentes más importantes obtenidos por la matriz de datos

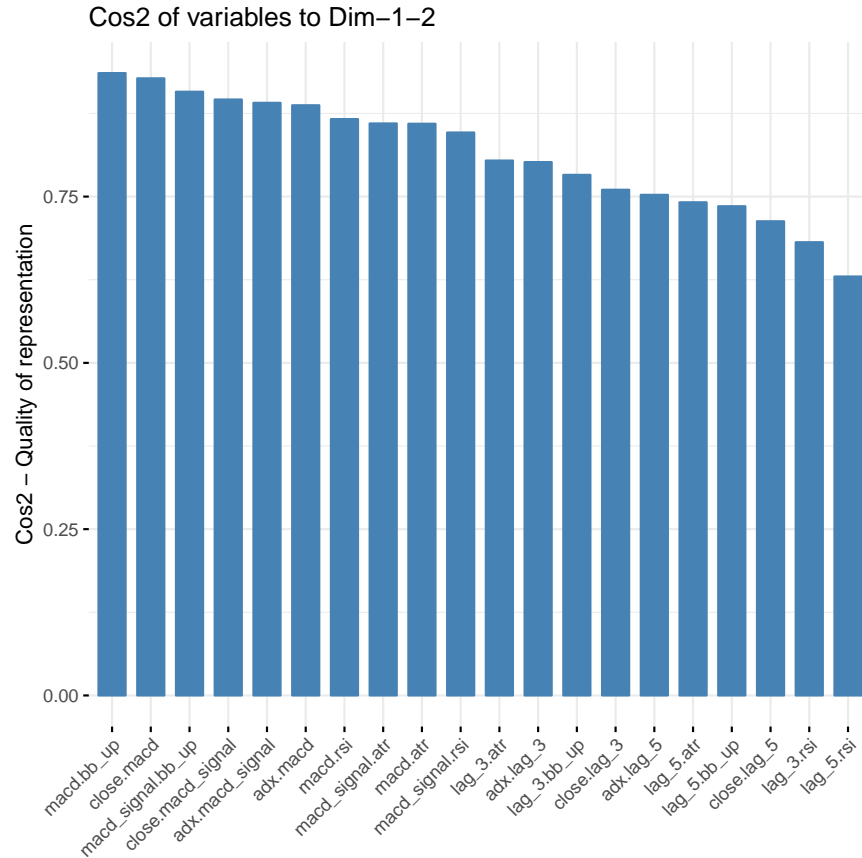
La contribución de las variables representan la variabilidad contenida en un componente. Las variables correlacionadas con el componente principal 1 (PC1) y PC2 son las más importantes en explicar la variabilidad en la matriz de datos. Aquellas que no se correlacionan con ninguna componente son desechadas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



**Figura 3.7:** 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  $\frac{1}{N_{deVariables}} = \frac{1}{45} = 2,2\%$ . Una variable sobre este umbral se considera importante en la contribución al componente. Se aprecia como 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 – se muestra los valores de  $Cos^2$  para las primeras 2 componentes



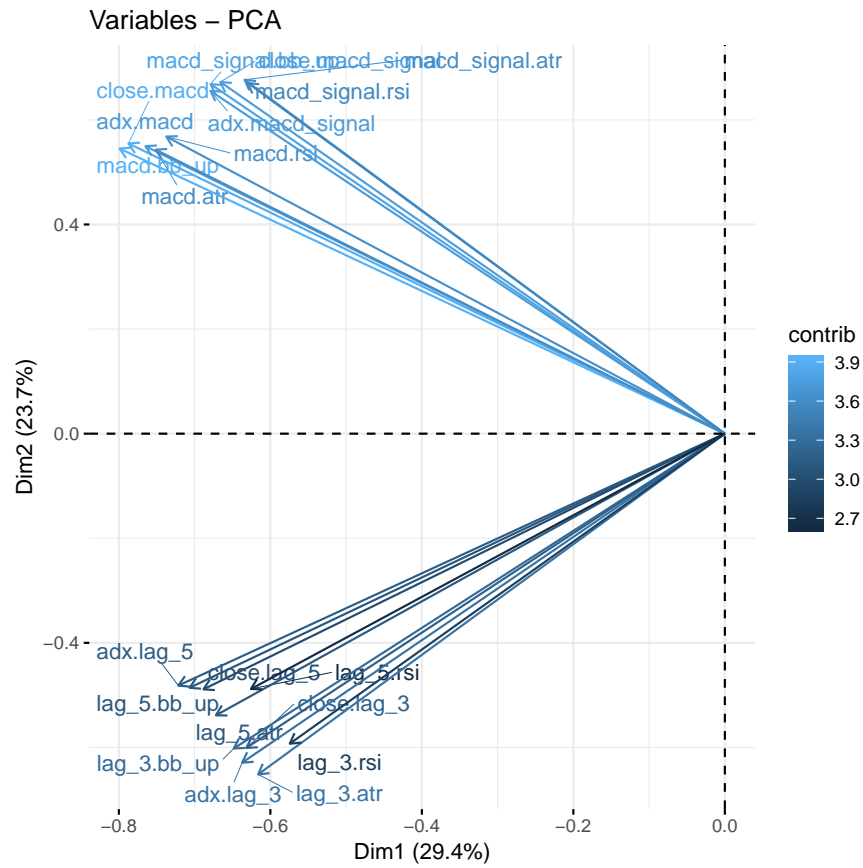
**Figura 3.8:** Calidad de representación medida por  $Cos^2$  de cada variable en PC1 y PC2

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

En la figura – 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. Por otro lado los grupos forman un ángulo de  $90^\circ$  por lo que no están correlacionados.





**Figura 3.9:** Gráfico de Correlación entre PC1 y PC2

# Análisis de Resultados

---

En el presente capítulo se realiza la descripción de los resultados obtenidos despues 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.

## 4.1. Coeficientes del modelo

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.

**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)	-0.3599	0.0645	-5.58	0.0000
PC1	-0.0146	0.0181	-0.81	0.4183
PC2	-0.0134	0.0203	-0.66	0.5085
PC3	0.0050	0.0327	0.15	0.8779
PC4	0.0873	0.0353	2.47	0.0135
PC5	-0.0222	0.0373	-0.60	0.5514
PC6	-0.0158	0.0425	-0.37	0.7100
PC7	0.1057	0.0484	2.18	0.0289

Período de entrenamiento 2009 - 2013

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.4093	0.0580	-7.06	0.0000
PC1	0.0222	0.0165	1.35	0.1785
PC2	-0.0072	0.0182	-0.40	0.6925
PC3	0.0169	0.0290	0.58	0.5596
PC4	0.0537	0.0307	1.75	0.0800
PC5	0.0237	0.0336	0.71	0.4797
PC6	0.0393	0.0365	1.08	0.2817
PC7	-0.1222	0.0414	-2.95	0.0032

Período de entrenamiento 2009 - 2014

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.2806	0.0527	-5.33	0.0000
PC1	0.0464	0.0156	2.98	0.0029
PC2	0.0260	0.0173	1.50	0.1334
PC3	0.0755	0.0265	2.85	0.0043
PC4	-0.0440	0.0271	-1.63	0.1039
PC5	-0.0449	0.0314	-1.43	0.1527
PC6	-0.0774	0.0324	-2.39	0.0168
PC7	-0.1044	0.0384	-2.72	0.0065

Período de entrenamiento 2009 - 2015

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.1756	0.0490	-3.58	0.0003
PC1	-0.0670	0.0151	-4.43	0.0000
PC2	0.0259	0.0168	1.54	0.1239
PC3	-0.1284	0.0238	-5.39	0.0000
PC4	-0.0246	0.0270	-0.91	0.3622
PC5	-0.0647	0.0295	-2.19	0.0287
PC6	-0.1049	0.0309	-3.40	0.0007
PC7	-0.1040	0.0364	-2.86	0.0042

Período de entrenamiento 2009 - 2016

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.1306	0.0456	-2.86	0.0042
PC1	-0.0643	0.0139	-4.61	0.0000
PC2	0.0319	0.0157	2.03	0.0423
PC3	-0.1211	0.0217	-5.58	0.0000
PC4	-0.0209	0.0254	-0.82	0.4101
PC5	-0.0352	0.0274	-1.28	0.1988
PC6	-0.0993	0.0300	-3.31	0.0009
PC7	-0.0901	0.0342	-2.63	0.0084

Período de entrenamiento 2009 - 2017

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.1148	0.0427	-2.69	0.0072
PC1	0.0629	0.0130	4.85	0.0000
PC2	0.0334	0.0144	2.31	0.0208
PC3	0.0977	0.0199	4.90	0.0000
PC4	0.0053	0.0231	0.23	0.8202
PC5	0.0223	0.0255	0.87	0.3816
PC6	-0.0584	0.0275	-2.12	0.0340
PC7	-0.0725	0.0319	-2.28	0.0228

Se observa que para todos los períodos el ACP arroja componentes que recogen el 85 % de la variación. También se aprecia que a medida que aumentamos los años de entrenamiento el número de p-valores menores que 0.05 aumentan, insinuando que mientras más observaciones para entrenar el modelo, mayor será la asociación entre los componentes y la capacidad de predecir el retorno objetivo.

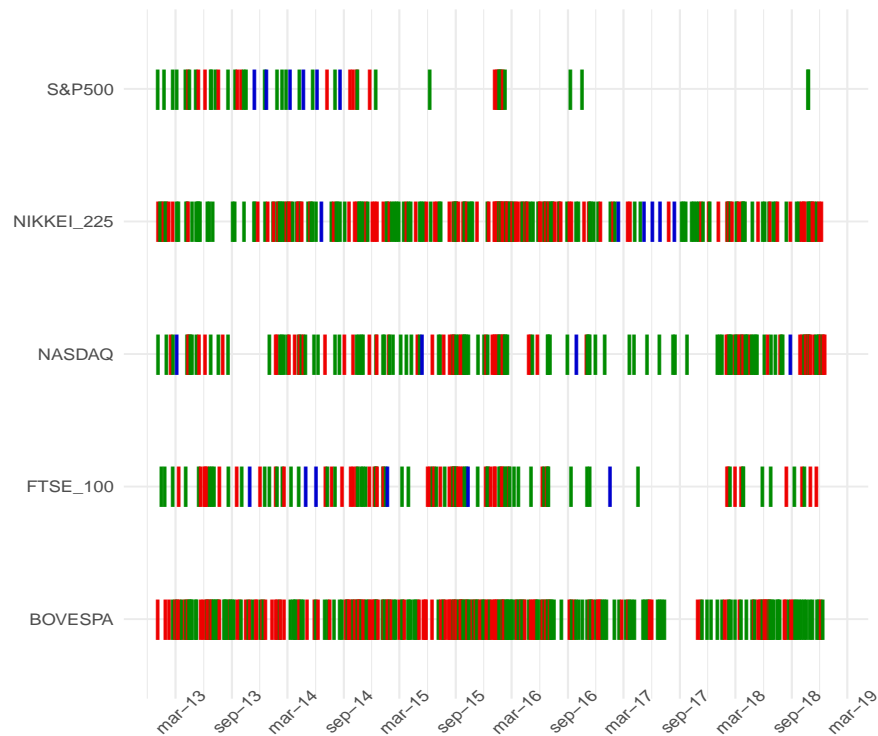
## 4.2. Resultados de la simulación

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 %, dado que 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.

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

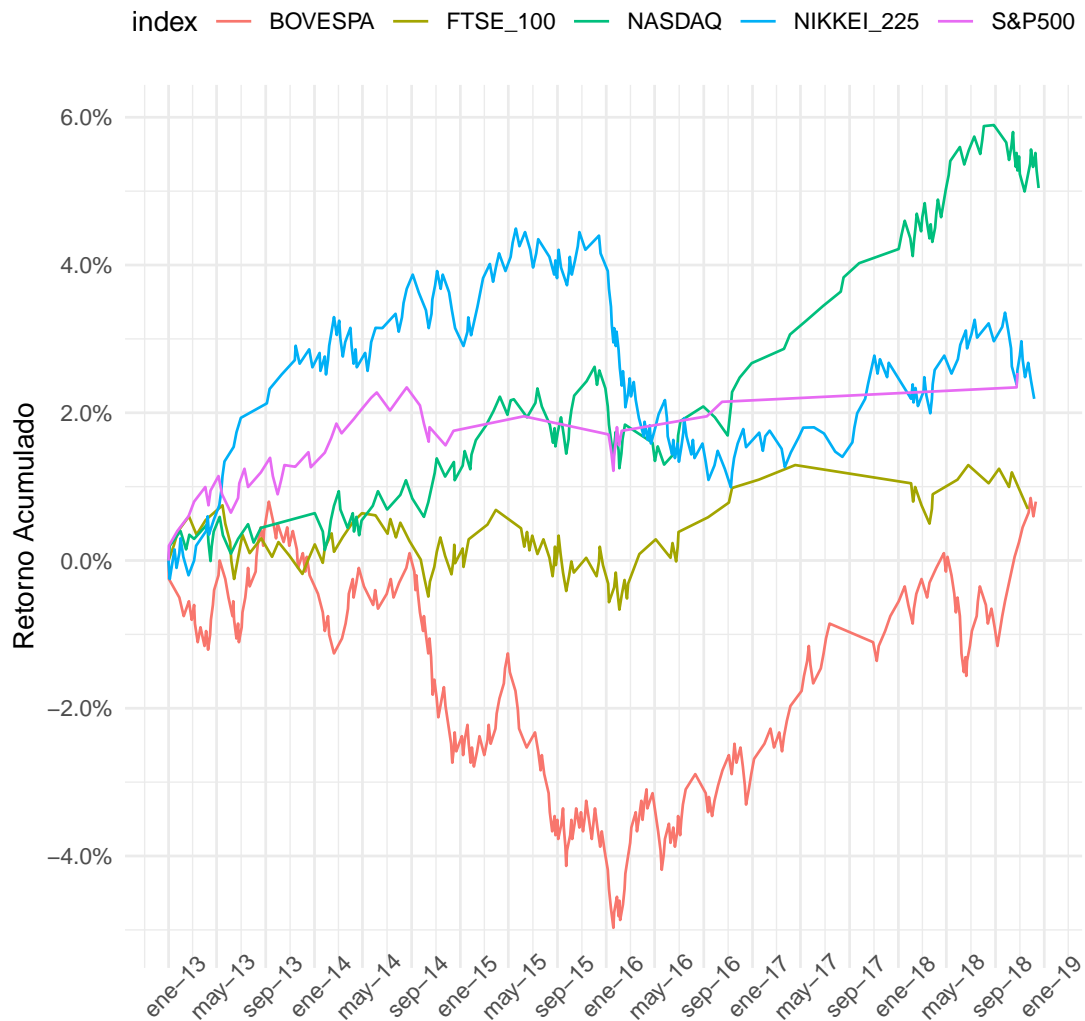
	S&P_500	NASDAQ	NIKKEI_225	FTSE_100	BOVESPA
False Buys	21	61	99	53	128
True Buys	32	98	127	62	164
N° trades	53	159	226	115	292
Accuracy	60.38 %	61.64 %	56.19 %	53.91 %	56.16 %
Accumulative Return	2.57 %	5.17 %	2.21 %	0.70 %	0.80 %
Max Drawdown	1.13 %	1.37 %	3.46 %	1.42 %	5.64 %

En la figura 4.1 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 perdidas y los azules son clasificados como 'False buys' pero cerraron el trade por límite de tiempo.

**Figura 4.1:** Clasificación de la simulación

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 4.2:** Retorno acumulado para cada índice

Al asumir que siempre se abre la posición con la misma cantidad de dinero, en este caso 1000 USD, se sabe que los trades solo 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 comunmente 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	-0.83	0.41
NASDAQ	0.16	0.87
NIKKEI225	1.19	0.23
FTSE_100	1.09	0.28
BOVESPA	-0.09	0.93

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.

## 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. En ningún momento se busca medir la posible ganancia en base a la simulación, sino, 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.

Si bien el retorno acumulado durante 6 años de prueba es poco atractivo, 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  $tp$ ,  $sl$  y  $h$  para el activo a operar, así como la elección del tamaño de los períodos de entrenamiento. 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 árboles de decisión o SVM mejoren la predicción.

El modelo solo considera la predicción del incremento del precio. Se utiliza la figura del stop loss como reducción 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.



## 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.