CARRERA: ESPECIALIZACIÓN EN CIENCIA DE DATOS

TRABAJO FINAL INTEGRADOR

TRADING ALGORÍTMICO: UN ESTUDIO COMPARATIVO ENTRE TÉCNICAS DE CLASIFICACIÓN Y REGRESIÓN EN EL ÁMBITO DE LAS FINANZAS

Nombre y Apellido del Alumno/a: Francisco Lonardi

Título de grado o posgrado (último): Ingeniero Industrial

Tutor:

Gustavo Arjones

Lugar y Fecha: San Isidro, 10 de diciembre de 2020

1. Abstract

En el presente trabajo se desarrolla una técnica de trading algorítmico basada en diferentes tipos de modelos de Machine Learning que utilizan como *features* variables propias del análisis técnico de activos. Utilizando una muestra de 10 acciones tomadas del índice NASDAQ, las estrategias desarrolladas permiten no sólo comparar la performance de modelos de clasificación con modelos de regresión sino también demostrar que éste es un enfoque superador en términos de riesgo-retorno a las técnicas tradicionales de trading.

2. Introducción

El uso de técnicas de Machine Learning en el ámbito del Trading Algorítmico es una práctica cada vez más frecuente tanto en artículos académicos como en el uso profesional. Este trabajo presenta un enfoque novedoso que permitirá comparar las dos metodologías de Aprendizaje Supervisado, regresión y clasificación, a la luz de un problema común, definir una estrategia de trading de acciones que sea capaz de generar retornos a la inversión.

3. Estado del Arte

3.1. Trading Algorítmico

El trading de acciones se refiere a la actividad de compra-venta de títulos empresariales en el mercado financiero. Actualmente, los modelos de trading de acciones se rigen a través del análisis técnico de los precios y volúmenes históricos o al análisis fundamental de las empresas emisoras de las acciones [1]. En el primer caso, se utilizan factores económicos y financieros para determinar el valor intrínseco de las acciones. En el segundo, se utilizan técnicas estadísticas descriptivas para determinar las tendencias futuras de los precios.

Dentro del mundo del trading algorítmico, existen dos corrientes claramente marcadas: quienes buscan sistematizar y automatizar el análisis técnico de acciones, y quienes utilizan la información disponible para alimentar modelos de Machine Learning para desarrollar sistemas expertos que sean capaces de obtener ganancias a través decisiones no predeterminadas según reglas.

Los datos disponibles públicamente para cada ticker (identificador único de una acción en un mercado dado) con frecuencia diaria son: precio de apertura (Open), precio máximo (High), precio mínimo (Low), precio de cierre (Close), precio de cierre ajustado (Adj. Close) y volumen (Volume). El precio de cierre ajustado modifica el precio de cierre de una acción para reflejar con precisión el

valor de esa acción después de contabilizar cualquier acción corporativa como splits o pago de dividendos.

El uso de Machine Learning para trading algorítmico presenta principalmente dos líneas de trabajo: la predicción de los precios futuros o la de la dirección de los precios futuros (si subirán o bajarán). La predicción de precios futuros utiliza modelos de regresión, y la predicción de la dirección de los precios futuros utiliza modelos de clasificación. En la literatura científica, el segundo enfoque es el preponderante mientras que en los blogs especializados se utiliza con mayor frecuencia el primero.

3.2. Análisis Técnico

El análisis técnico comprende un conjunto de definiciones teóricas basadas en la comparación de indicadores que determinan si es buen momento para comprar o vender. En la actividad práctica, los indicadores de mayor relevancia son: Moving Average (MA), Weighted Moving Average (WMA), Bollinger Bands (BB), Moving Average Convergence/Divergence (MACD), Relative Strength Index (RSI), Stochastic Oscillator (Stoch), Williams' %R (WPR) [1]–[3], On Balance Volume (OBV). Estos indicadores, en el contexto de la utilización de Machine Learning, son utilizados como features adicionales para los modelos. Estos cálculos estadísticos en el contexto del análisis técnico generalmente son utilizados con el precio de cierre del ticker al final del día como variable base.

3.2.1. Moving Average (MA)

Es un indicador simple formado al calcular el precio promedio de una acción durante un número específico de períodos. Su cálculo se repite para cada día en un gráfico y luego se une para formar una curva suave, lo que facilita la identificación de la dirección de la tendencia. La fórmula de cálculo se muestra en (1).

$$MA(p,n) = \frac{\sum_{i=0}^{n} p_i}{n} \quad (1)$$

3.2.2. Weighted Moving Average (WMA)

Es una familia de indicadores que se computan como un promedio móvil, pero asignando diferentes pesos a cada uno de los datos de la serie. En general, se asigna un mayor peso a los datos más recientes como se ve en la fórmula (2). En particular, un promedio móvil exponencial (EMA) es un

tipo de promedio móvil pesado que otorga un mayor peso e importancia a los puntos de datos más recientes con un crecimiento exponencial. Su mecanismo de cálculo se muestra en (3).

$$WMA(p,n) = \frac{p_1 \, x \, n + \, p_2 \, x \, (n-1) + \cdots + \, p_n}{\frac{n \, x \, (n+1)}{2}} \quad (2)$$

$$EMA(p,n) = p_i \, x \, \frac{2}{n+1} + MA_{i-1} \, x \, (1 - \frac{2}{n+1}) \quad (3)$$

3.2.3. Bollinger Bands (BB)

Las Bollinger Bands (BB) fueron diseñadas para comparar la volatilidad y niveles de precios relativos durante un período de tiempo. Se compone de tres bandas que abarcan la mayoría del rango de acción de un precio. Las bandas son: una MA en el medio (4); una banda superior (5), calculada como una MA más cierta cantidad (m) de desviaciones estándar; y una banda inferior (6), calculada como un SMA menos más cierta cantidad de desviaciones estándar. Además de identificar los niveles de precios relativos y la volatilidad, las BB también se pueden combinar con el precio para generar señales y pronosticar movimientos significativos. En general se utilizan 20 períodos.

$$BBM(p,n) = SMA(p,n) \quad (4)$$

$$BBU(p,n) = SMA\left(\frac{High + Close + Low}{3}, n\right) + m \, x \, \sigma\left(\frac{High + Close + Low}{3}\right) \quad (5)$$

$$BBD(p,n) = SMA\left(\frac{High + Close + Low}{3}, n\right) - m \, x \, \sigma\left(\frac{High + Close + Low}{3}\right) \quad (6)$$

3.2.4. Moving Average Convergence/Divergence (MACD)

El MACD es un indicador de impulso de tendencia que muestra la relación entre dos promedios móviles exponenciales (EMA) de precios y se calcula como la diferencia entre ellos (7). El valor más comúnmente utilizado es la comparación entre las medias de 12 y 26 días. Para obtener conclusiones de este indicador, se traza una EMA de nueve días del MACD llamada "línea de señal"

que se utiliza como disparador de señales de compra y de venta. Cuando el MACD cruza por encima de su línea de señal es recomendable comprar y cuando el MACD cruza por debajo de la línea de señal es recomendable vender.

$$MACD(p) = SMA(p, 12) - SMA(p, 26) \quad (7)$$

3.2.5. Relative Strength Index (RSI)

El RSI es un indicador de impulso que mide la magnitud de los cambios recientes en los precios para evaluar las condiciones de sobrecompra o sobreventa en el precio de una acción. Compara la magnitud de las ganancias recientes con la magnitud de las pérdidas recientes y genera un número que oscila entre 0 y 100. El cálculo está expresado en la fórmula (8). Se muestra como un oscilador (un gráfico lineal que se mueve entre dos extremos). Se interpreta en general que, cuando el RSI supera el valor 30, se considera una señal de que los precios tenderán a subir. Por el contrario, cuando cae por debajo de 70 se considera una señal de que es probable que los precios bajen.

$$RSI(n) = 100 - \left(\frac{100}{1 + \frac{EMA[Ganancias(n)]}{EMA[Pérdidas(n)]}} \right) \quad (8)$$

3.2.6. Stochastic Oscillator (Stoch)

El oscilador estocástico (Stoch) compara la ubicación del actual precio de cierre de una acción en relación con su rango de precios durante un cierto número de períodos. Este indicador está compuesto por dos líneas, la rápida (%K) y la lenta (%D), que se calculan de acuerdo con las siguientes ecuaciones (9) y (10).

$$\%K = 100\, x\, \frac{CierreActual - MínimoLow(n)}{MáximoHigh(n) - MínimoLow(n)} \quad (9)$$

$$\%D = MA(\%K)\ [3\ períodos] \quad (10)$$

Se considera que los niveles de cierre que están consistentemente cerca del tope del rango indican acumulación (presión de compra) y aquellos cerca del fondo del rango indican distribución (presión de venta)

3.2.7. Williams' %R (WPR)

Williams' %R es un tipo de indicador de impulso que se mueve entre 0 y -100 y mide los niveles de sobrecompra y sobreventa. Este valor puede usarse para encontrar puntos de entrada y salida en el mercado. El indicador es muy similar al oscilador estocástico y se usa de la misma manera. Compara el precio de cierre de una acción con el rango alto-bajo durante un período específico, generalmente 14 días o períodos. En general se considera que un valor por encima de -20 es una señal de venta y por debajo de -80 es una señal de compra. Su fórmula de cálculo se muestra en la ecuación (11).

$$Williams'\%R = \frac{MáximoHigh(n) - CierreActual}{MáximoHigh(n) - MínimoLow(n)} \quad (11)$$

3.2.8. On Balance Volume (OBV)

On Balance Volume es un tipo de indicador de impulso que mide el flujo volumétrico positivo y negativo de un determinado activo. Este indicador fue desarrollado en base a la idea que el volumen era la fuerza impulsora detrás de los mercados, de manera de poder proyectar cuándo ocurrirían movimientos importantes. Cuando el volumen transaccionado aumenta o disminuye drásticamente sin ningún cambio significativo en el precio, en algún momento el precio se dispara hacia arriba o hacia abajo. Durante un período de tiempo, el volumen comienza a impulsar el precio hacia arriba y lo contrario comienza a suceder cuando las instituciones comienzan a vender su posición y los inversores minoristas comienzan nuevamente a acumular sus posiciones.

El algoritmo de cálculo del OBV es el siguiente:

- Si el cierre de hoy es mayor que el cierre de ayer, entonces el volumen de hoy se agrega al OBV de ayer y se considera un volumen elevado.
- Si el cierre de hoy es menor que el cierre de ayer, entonces el volumen de hoy se resta del OBV de ayer y se considera un volumen bajo.
- Y si el cierre de hoy es igual al cierre de ayer, el OBV de hoy es igual al OBV de ayer.

3.3. Algoritmos de Machine Learning

Machine Learning es una aplicación de la Inteligencia Artificial con sus raíces en la estadística y las ciencias de la computación que busca a través del uso de datos darles la capacidad a las computadoras de aprender a detectar y replicar patrones sin ser explícitamente programadas [4].

Las técnicas de Aprendizaje Supervisado se subdividen según el tipo de problema a resolver: regresión o clasificación. En el primer caso, la variable respuesta o label toma valores cuantitativos y continuos. En cambio, en el segundo, la variable respuesta toma valores cualitativos o discretos. Existen numerosos algoritmos capaces de afrontar estos problemas, algunos de los cuales tendrán mención en el presente trabajo.

3.3.1. Linear Regression

Es una familia de técnicas utilizadas principalmente para abordar problemas de regresión que supone una relación lineal entre las variables dependientes y la variable a predecir [5].

Dentro de esta familia se encuentran también la Ridge Regression y la LASSO Regression. Ambas se basan en el concepto de regularización para reducir el efecto de overfitting en los modelos. La primera se caracteriza por utilizar un término de penalización cuadrático en su función de costos y la segunda utiliza un término lineal en valores absolutos [5]. Este último modelo además de ayudar a reducir el efecto de overfitting es muy bueno para la selección de variables relevantes.

3.3.2. K-Nearest Neighbors (KNN)

Es un algoritmo que supone comportamientos similares en las proximidades valiéndose generalmente del concepto de Distancia Euclidiana. En el caso de un problema de clasificación, devuelve la moda de los valores obtenidos en los k vecinos más cercanos al valor a predecir. En el caso de un problema de regresión, obtiene la media [1].

3.3.3. Logistic Regression

Es un algoritmo que utiliza los principios de una Linear Regression y le aplica a la combinación lineal de las variables dependientes una transformación no lineal utilizando la función logística [6]. Ésta, al estar acotada entre 0 y 1, permite que este algoritmo sea válido para problemas de clasificación.

3.3.4. Naive Bayes

Es una técnica de clasificación basada en el Teorema de Bayes que se basa en el supuesto de independencia entre las variables dependientes [7]. Este tipo de modelo supone que la presencia de una característica particular en una clase no está relacionada con la presencia de ninguna otra característica y que todas las características (o variables del modelo) tienen el mismo efecto en el resultado.

3.3.5. SVM

Es una familia de algoritmos que pueden ser utilizados tanto para problemas de regresión como de clasificación. Su objetivo en el contexto de clasificación es encontrar un hiperplano en un espacio N-dimensional que clasifique correctamente los puntos de datos. Para separar las dos clases de puntos de datos, hay muchos hiperplanos posibles que podrían elegirse, pero esta técnica busca encontrar un plano que tenga el margen máximo, es decir, la distancia máxima entre puntos de datos de ambas clases. [4]

3.3.6. Decision Trees

Es una familia de modelos que intentan resolver problemas de clasificación y regresión mediante el uso de la representación del árbol. Cada nodo interno del árbol corresponde a un atributo, y cada nodo hoja corresponde a una etiqueta de clase o valor promedio de la variable independiente (regresión). Usando un árbol de decisión, es posible visualizar las decisiones, lo cual lo hace fácil de entender.

Un Random Forest es un método de ensamble que se utiliza en problemas de clasificación y regresión y surge a través de la combinación de árboles de decisión que crecen en subespacios de variables y subconjuntos de datos seleccionados al azar [8].

XGBoost es un algoritmo computacionalmente eficiente de Machine Learning que utiliza el framework de Gradient Boosting, que se basa en una cadena de modelos (árboles de decisión) anidados en el que cada clasificador altera sus criterios de evaluación mediante el algoritmo de gradiente en descenso basándose en los resultados de su predecesor. Esto aumenta la eficiencia del proceso de clasificación (o regresión) mediante la implementación de un proceso de evaluación más dinámico.

3.3.7. Multilayer Perceptron (NN)

Es una arquitectura de red neuronal artificial formada por múltiples capas, de tal manera que tiene capacidad para resolver problemas que no son linealmente separables. Está compuesto por múltiples capas que pueden o no estar totalmente conectadas. Es una extensión del perceptrón simple, que se compone de una sola capa que computa la combinación lineal entre los pesos y los valores de las variables de entrada y les aplica una función de activación. Este algoritmo puede ser utilizado para resolver problemas de regresión o clasificación.

3.4. Machine Learning aplicado al Trading

Los algoritmos de Machine Learning son utilizados en el ámbito del trading debido a la complejidad del problema inherente, y los comportamientos no lineales y dinámicos de los mercados financieros [2]. Además, estos mercados son extremadamente sensibles a los factores políticos, las condiciones microeconómicas y macroeconómicas, y las expectativas e inseguridades de los inversores [9]. Existe bibliografía referente al uso de técnicas de Aprendizaje Supervisado y Aprendizaje No Supervisado para el análisis de mercados de acciones. En el estudio [10] se resumen las diferentes técnicas utilizadas a lo largo de los últimos años, comenzando por Kalman Filters y finalizando por diferentes tipos redes neuronales recurrentes (RNN), recopilando resultados de diferentes trabajos realizados en el campo cada uno de ellos haciendo uso de información de diferentes bolsas del mundo.

Los modelos más utilizados en la literatura son Redes Neuronales (NNs), Support Vector Machines (SVMs), Multiple Kernel Learning (MKL), Random Forests (RFs) y K-Means [8]. Adicionalmente, la literatura muestra interés respecto al uso de modelos de ensamble, que combinan simultáneamente diferentes tipos de algoritmos, y sus beneficios respecto a la resolución de este tipo de problemas complejos [2]. En esta línea se encuentran los algoritmos de Random Forest y Multiple Kernel Learning ya mencionados, y también otros como AdaBoost. Los hallazgos respecto a la comparación de todas estas técnicas indican que Random Forest es el algoritmo que mejor se desempeña en este ámbito [2].

Debido a la naturaleza dinámica del movimiento de los precios de las acciones en los mercados financieros, existen líneas de investigación que prefieren la simplicidad de los modelos respecto a la performance de los mismos permitiendo así analizar una gran cantidad de tickers cada día [1].

Las técnicas de Machine Learning aplicadas a los pronósticos de acciones han demostrado tener buena performance (accuracy>80%) [8]. Sin embargo, no está claro para la mayoría de estos trabajos si las técnicas propuestas serían capaces de generar ganancias en la práctica. La performance técnica de un modelo de Machine Learning aplicado a trading algorítmico no necesariamente implica que el modelo sea capaz de generar ganancias. Los resultados del artículo

[1] muestran casos en los que se observa que, a un mismo nivel de precisión, las ganancias pueden resultar positivas o negativas.

4. Definición del Problema

Actualmente no existe en la bibliografía una comparación entre modelos de clasificación y regresión en el ámbito del trading algorítmico de acciones.

5. Justificación del estudio

El presente estudio busca anteponer un objetivo único de negocio para poder comparar diferentes estrategias de trading basadas en el uso combinado del análisis técnico y algoritmos de Machine Learning. En particular, se intentará generalizar los resultados obtenidos para comprender si el uso de técnicas de Clasificación es superador al uso de técnicas de Regresión en el ámbito planteado.

6. Alcances del trabajo y limitaciones

El presente trabajo tiene como alcance comparar estrategias orientadas al trading de acciones individuales, y deja fuera el análisis de carteras. Sin embargo, el objetivo final a largo plazo que persigue esta línea de investigación es la creación de un fondo común de inversiones parametrizado específicamente para la aversión al riesgo de cada inversor y que base sus decisiones de inversión en técnicas de Machine Learning. Es por eso por lo que se busca en el marco del trabajo generar un mecanismo utilizando Análisis Técnico y Machine Learning, que resulte confiable y parametrizable para hacer trading de acciones maximizando el retorno sobre la inversión y minimizando su riesgo.

Las limitaciones del estudio son principalmente de recursos humanos. Se estima que se podrá finalizar en el trascurso de 25 semanas con una dedicación de 4 horas semanales en promedio.

7. Hipótesis

En trading algorítmico, técnicamente es posible realizar abordajes de regresión o clasificación, pero las métricas inherentes de performance de cada uno son diferentes y no permiten la comparación entre sí. Imponiendo un objetivo de negocio, en este caso el retorno a la inversión obtenido a lo

largo de un período de tiempo, en vez de una métrica técnica se torna posible comparar ambos enfoques logrando no solo evaluar en este caso si uno se comporta mejor que el otro sino también sentar un precedente a la hora de generar un caso práctico de comparación entre ellos.

"La utilización de modelos de clasificación genera un mayor retorno a la inversión que los modelos de regresión"

7.1. Definición de Variables

1. *Retorno de la Inversión:* Es el valor de los activos del inversor luego de aplicada una cierta estrategia de trading durante un período de tiempo definido dividido el valor de los activos al comienzo del período.
2. *Precio de una acción:* Es el valor económico por el cual se transacciona una determinada acción.
3. *Porcentaje de cambio de precio futuro:* Resulta de la división de la diferencia entre el precio de una acción en un dado período y el precio en el período anterior, y el precio en el período anterior.
4. *Monto de una operación de trading:* Resulta de la multiplicación del precio de una acción por la cantidad de compra o venta según corresponda.
5. *Comisión de compra-venta de una acción:* Es un porcentaje fijo del monto de una operación que se le debe pagar a un bróker al realizar una operación de compra o venta de una acción.
6. *Límite mínimo para la compra de una acción:* Es el porcentaje mínimo de cambio del precio futuro a partir del cual un inversor tomará la decisión de compra.
7. *Límite mínimo para la venta de una acción:* Es el porcentaje mínimo de cambio del precio futuro a partir del cual un inversor tomará la decisión de venta.
8. *Ventana temporal de la simulación:* Es el intervalo de tiempo en el cual se ejecutará la simulación de una estrategia de trading.

8. Objetivos

8.1. Objetivo General

Utilizando la información diaria de precios de un conjunto de acciones que componen el índice NASDAQ, se buscará desarrollar una herramienta que permita comparar los retornos de la inversión de diversas estrategias de trading basadas en técnicas de regresión y clasificación, e incorporando en los modelos variables propias del análisis técnico.

8.2. Objetivos Específicos

1. Construir la base de datos que se utilizará para el desarrollo del trabajo y definir los valores de los parámetros y supuestos clave de este.
    a. Seleccionar una muestra representativa de entre las acciones que componen el índice NASDAQ para comenzar el estudio.
    b. Obtener la información de precios de cada acción y construir los indicadores propios del análisis técnico para incorporar como variables a los modelos de Machine Learning.
    c. Listar los supuestos del estudio y establecer el valor de los parámetros. Los parámetros serán: comisión de compra-venta de activos, límites mínimos para la compra y venta, y tamaño de las ventanas temporales de simulación.
2. Construir modelos de regresión para predecir el precio futuro de cada una de las acciones y realizar un ajuste de hiperparámetros para encontrar el mejor conjunto en términos del error cuadrático medio.
3. Construir modelos de regresión para predecir el porcentaje de cambio de precio futuro de cada una de las acciones y realizar un ajuste de hiperparámetros para encontrar el mejor conjunto en términos del error cuadrático medio.
4. Construir modelos de clasificación para predecir si el precio futuro de cada una de las acciones estará por encima o por debajo de los límites de compra y venta, y realizar un ajuste de hiperparámetros para encontrar el mejor conjunto en términos de la precisión.
5. Definir una estrategia base de trading como benchmark, y definir todo el set de estrategias a evaluar utilizando los modelos construidos.
6. Simular la utilización de cada una de las estrategias definidas en ventanas móviles para evaluar su retorno a la inversión en cada una de las acciones del estudio.
7. Comparar los resultados de cada una de las estrategias utilizadas y obtener conclusiones.

9. Metodología

9.1. Técnicas de Análisis Técnico

A través de la construcción de indicadores de Análisis Técnico se generarán variables adicionales para ser incorporadas en los modelos a evaluar. De los resultados obtenidos y su interpretación surgirán otras potenciales variables indicando las decisiones de movimientos (compra, retención o venta) derivadas directamente de ellas según la teoría clásica. Los principales indicadores que serán considerados en el estudio son los listados a continuación:

1. Moving Average (MA)
2. Weighted Moving Average (WMA)
3. Bollinger Bands (BB)
4. Moving Average Convergence/Divergence (MACD)

5. Relative Strength Index (RSI)
6. Stochastic Oscillator (Stoch)
7. Williams' %R (WPR)
8. On Balance Volume (OBV)

9.2. Técnicas de Machine Learning

A continuación, se detallan las técnicas de Machine Learning a ser utilizadas para los modelos de regresión y clasificación.

9.2.1. Regresión

1. Linear Regression
2. LASSO Regression
3. Random Forest Regression
4. XGBoost Regression

9.2.2. Clasificación

1. Logistic Regression
2. Random Forest
3. XGBoost Classification

9.3. Herramientas

El trabajo se realizará utilizando Jupyter Notebooks en lenguaje Python 3.7 corriendo en un entorno virtual a través de Google Colab. En caso de ser necesarios más recursos, se utilizarán máquinas virtuales de Google Cloud para este propósito. La librería base para correr los modelos de Machine Learning será scikit-learn.

10. Resultados

10.1. Proceso

Antes de describir los resultados alcanzados, es pertinente comentar de manera breve el proceso llevado a cabo de manera que se entienda el enfoque adoptado, los modelos utilizados y todas las decisiones tomadas a lo largo del camino.

El enfoque llevado a cabo propone una resolución sistemática y ordenada del problema desde la toma de datos hasta la verificación de los objetivos.

Para realizar el trabajo, se seleccionaron 10 acciones pertenecientes al índice NASDAQ de manera de tener una mezcla heterogénea de industrias y también de performance histórica para poder luego extraer conclusiones que puedan ser extrapolables a diferentes contextos. De las acciones utilizadas, 2 tuvieron resultados negativos en el período de simulación, 2 tuvieron resultados neutros, 3 tuvieron resultados positivos y las otras 3 tuvieron resultados positivos por encima del 50% anual.

Luego de levantar los datos históricos de la selección de acciones utilizando la API pública de Yahoo Finance y guardarlos en un dataframe multi-indexado, se procedió a realizar la Ingeniería de Variables. Este paso ocurre principalmente en dos etapas. La primera genera cálculos básicos entre las variables existentes como la amplitud del precio de una acción a lo largo de una jornada o la diferencia entre el precio de apertura y el de cierre, entre otras. La segunda involucra el cálculo de los indicadores de análisis técnico descriptos en las secciones anteriores a través de funciones ad-hoc generadas por encima de la librería talib, específicamente diseñada para estos fines.

Una vez finalizada la Ingeniería de Variables, se procedió a realizar un breve análisis descriptivo de las acciones utilizando resúmenes estadísticos y técnicas como un dendograma para entender el grado de similitud de los tickers elegidos a través de un agrupamiento jerárquico.

Luego comenzó la etapa de estimación, donde el objetivo principal fue utilizar los datos disponibles de cada uno de los activos para entrenar diferentes modelos de Machine Learning con el fin de obtener predicciones que luego sean susceptibles a un proceso de simulación de cartera.

En primer lugar, se definieron 3 variables objetivo diferentes para analizar independientemente: el precio al cierre del día, la variación porcentual del precio al cierre del día respecto al cierre del día anterior y la dirección de la variación del precio al cierre del día respecto al cierre del día anterior. Las primeras dos se resuelven con técnicas de Regresión y la última con técnicas de Clasificación Múltiple. Se utilizan Regresión Lineal, Random Forest y XGBoost como técnicas de Regresión, y Regresión Logística, Random Forest y XGBoost como técnicas de Clasificación Múltiple. Las clases definidas al generar la variable objetivo del modelo de clasificación múltiple son: Up (cuando el precio al cierre sube por encima del límite mínimo para la compra de una acción), Down (cuando el

precio al cierre cae por debajo del límite mínimo para la venta de una acción) y None (cuando el cambio porcentual del precio de cierre se sitúa entre los límites mínimos de compra y venta).

En segundo lugar, se definieron 3 técnicas de preprocesamiento de datos diferentes: la utilización de todos los datos disponibles, la utilización de Análisis de Componentes Principales (PCA) para la extracción de variables relevantes derivadas de las existentes y el uso de los 5 componentes principales como los únicos inputs a un modelo, y la utilización de una Regresión LASSO para la determinación de las variables más influyentes del modelo y el uso de esas variables como únicos inputs al modelo.

En esta etapa entonces se combinaron, para cada acción, 3 técnicas de preprocesamiento de datos con 3 tipos de modelos de Machine Learning para cada variable objetivo. El resultado es un set de 9 modelos para cada activo que deben luego ser utilizados para simular en el pasado y medir su efectividad.

De manera de poder seleccionar el mejor modelo posible, se confeccionaron métricas de evaluación de desempeño ad-hoc para este problema. Para medir los modelos de regresión, se diseñó una métrica que únicamente penaliza la sobreestimación de la variable objetivo cuando la predicción señala una caída del precio respecto a la jornada anterior y únicamente penaliza la subestimación de la variable objetivo cuando la predicción señala un aumento del precio. Para los modelos de clasificación múltiple, se utilizó la métrica Balanced Accuracy que promedia la precisión de cada una de las clases.

Adicionalmente, para obtener el mejor conjunto de hiperparámetros de cada modelo se utilizó un proceso de Optimización Bayesiana con una grilla definida para cada modelo. De esta manera, el modelo utilizado para simular posteriormente será el que tenga una combinación de hiperparámetros tal que minimice el error de predicción en cada acción. Dado que los datos se encuentran en series de tiempo, es importante que el proceso de validación cruzada de los modelos se haga con un esquema de particiones de datos para series de tiempo como el que realiza la función TimeSeriesSplit de la librería scikit-learn. Entonces, para cada acción en cada variable objetivo y preprocesamiento de datos definidos se obtiene un único modelo que resulta de la comparación entre los tres modelos de Machine Learning definidos con sus hiperparámetros previamente optimizados a través de las métricas ad-hoc definidas para el caso.

Finalmente se llegó a la etapa de backtesting, en la cual se utilizan los modelos previamente entrenados para realizar las estimaciones diariamente y simular su uso ante decisiones concretas de inversión utilizando únicamente una acción como alternativa. Las únicas operaciones permitidas en el marco de la simulación fueron la compra o la venta de la acción determinada al precio de apertura de mercado, pagando comisiones de 0.5% en cada una de las operaciones para asemejar más fielmente la realidad. El periodo ventana para todas las simulaciones del presente trabajo es desde el 02/01/2019 al 30/12/2019 inclusive, permitiendo comparar todas las estrategias utilizadas en un lapso fijo e igual, y el presupuesto para la inversión inicial es de USD 10.000 en todos los casos.

Para poder determinar si una estrategia es exitosa o no, se determina una estrategia base, que consiste en utilizar todo el presupuesto disponible para comprar la acción al principio del periodo y

luego venderla al final de este sin hacer ninguna otra operación intermedia. De esta manera se tiene un benchmark para comparar el resto de las estrategias dominadas a través de los modelos de Machine Learning creados anteriormente.

Además de la estrategia base, se simularon otras tres estrategias: basada en la estimación del precio al cierre, en la magnitud de su cambio porcentual o en la dirección de su cambio respecto al día anterior. En todos los casos, el modelo genera una estimación utilizando el precio al cierre del día lo cual prohíbe al operador de realizar la operación en ese mismo momento si se considera un escenario realista. Para sobrellevar este detalle, se genera la estimación luego del cierre, pero la decisión de compra o venta se genera a la apertura del día siguiente, tomando también en cuenta como información adicional el precio de la acción en ese preciso momento.

Además de medir el Retorno de la Inversión al final del período luego de cada simulación, se decidió incorporar la métrica conocida como Sharpe Ratio, que permite dimensionar el retorno a la vez que la volatilidad del valor de la cartera, logrando así determinar si una estrategia fue más ventajosa en términos de riesgo-retorno que otra en lugar de solamente medir el retorno.

En resumen, para cada acción se corrieron 10 estrategias en total: una base y una por cada modelo de los definidos en la etapa de estimación.

El repositorio con el código utilizado para el desarrollo del presente trabajo se encuentra en el siguiente enlace: https://github.com/franlonardi/algo_trading_itba.

10.2. Resultados

En la tabla 1 se pueden observar el cambio porcentual del precio al cierre entre el primer y el último día de la simulación de las acciones elegidas para llevar a cabo el experimento.

Ticker	Cambio en el Precio al Cierre (%)
M	-41%
AAL	-12%
CTSH	-1%
BA	3%
AMZN	20%
GOOG	28%
TSLA	34%
MSFT	58%
NVDA	71%
AAPL	87%

Tabla 1: Cambio porcentual del precio al cierre de las acciones elegidas para el estudio. Producción propia.

Del análisis exploratorio de los datos se desprende la tabla 2, que muestra el precio promedio al cierre de cada una de las acciones elegidas durante el período de la simulación, su amplitud diaria promedio y el cambio promedio de precio al cierre respecto del día anterior.

Ticker	Precio al cierre Promedio	Amplitud (%) Promedio	Cambio Diario (%) Promedio
AAL	$30.58	2.91%	-0.08%
AAPL	$50.81	1.74%	0.18%
AMZN	$1,788.96	1.66%	0.04%
BA	$358.91	2.12%	0.09%
CTSH	$63.99	1.72%	0.00%
GOOG	$1,187.80	1.56%	0.12%
M	$18.66	3.24%	-0.08%
MSFT	$128.41	1.50%	0.05%
NVDA	$173.79	2.93%	0.09%
TSLA	$54.59	3.35%	0.16%

Tabla 2: Métricas resumen de los tickers elegidos durante el período de simulación. Producción propia.

La figura 1 resume el grado de asociación de las acciones elegidas para el estudio a través de un dendograma jerárquico entrenado con todas las variables disponibles para cada activo.

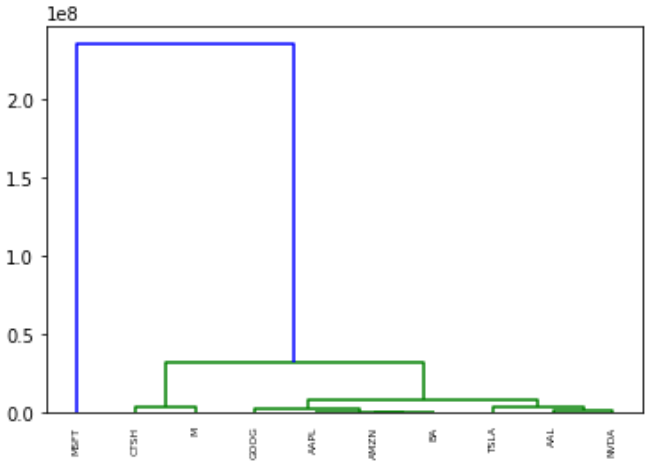

Figura 1: Representación gráfica del dendograma obtenido al agrupar jerárquicamente los activos en estudio. Producción propia.

Las figuras 2 y 3 presentan evidencia de cómo los modelos de regresión entrenados son capaces de ajustar a los precios de cierre en diferentes acciones durante el período de simulación.

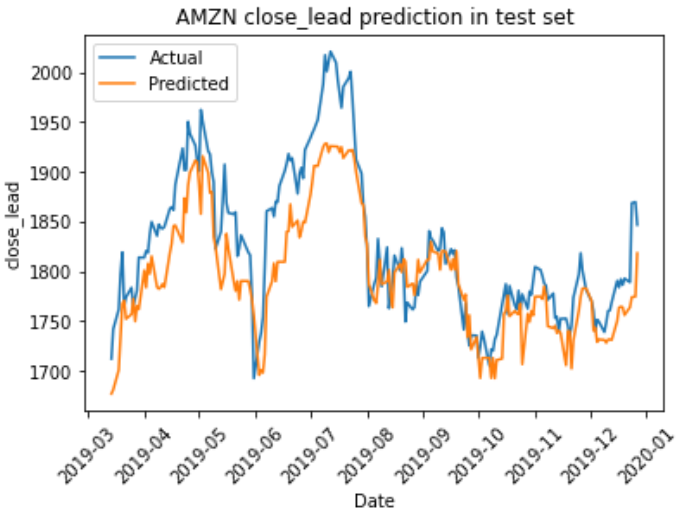

Figura 2: Gráfico temporal del precio al cierre real vs el estimado en el ticker AMZN. Producción propia.

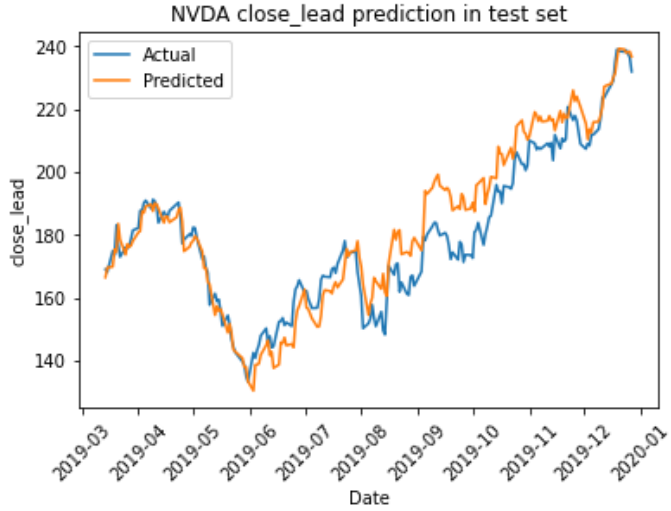

Figura 3: Gráfico temporal del precio al cierre real vs el estimado en el ticker NVDA. Producción propia.

Los resultados de performance en el set de validación cruzada de los modelos desarrollados en la etapa de estimación se presentan en las tablas 3 y 4. Las mismas permiten comparar a través de las métricas antes descriptas los modelos obtenidos en función de la técnica de preprocesamiento utilizada o de la variable respuesta elegida. Es importante notar que, si bien ambas refieren a

performance del modelo, no se pueden comparar directamente los resultados obtenidos en los modelos de regresión con los modelos de clasificación múltiple.

Ticker	Precio al Cierre	Cambio del precio al cierre	Dirección del cambio del precio al cierre
AAL	99.82%	92.09%	26.83%
AAPL	99.96%	96.81%	29.57%
AMZN	99.71%	95.43%	32.40%
BA	99.93%	97.73%	34.25%
CTSH	99.89%	97.10%	33.23%
GOOG	99.52%	97.38%	34.85%
M	99.87%	94.95%	23.28%
MSFT	99.97%	97.49%	34.48%
NVDA	99.93%	93.95%	25.85%
TSLA	99.59%	87.97%	31.37%
Promedio General	**99.82%**	**95.09%**	**30.61%**

Tabla 3: Performance promedio en el set de validación cruzada de los modelos de Machine Learning estimados para las diferentes variables objetivo en los tickers elegidos. Producción propia.

Ticker	Todas las variables	Variables LASSO	Variables PCA
AAL	72.77%	71.99%	73.97%
AAPL	74.67%	74.62%	77.06%
AMZN	76.17%	75.49%	75.87%
BA	77.33%	77.10%	77.48%
CTSH	76.75%	76.64%	76.82%
GOOG	77.05%	76.97%	77.72%
M	72.25%	72.08%	73.77%
MSFT	77.37%	77.36%	77.21%
NVDA	73.07%	72.77%	73.89%
TSLA	73.30%	72.41%	73.22%
Promedio General	**75.07%**	**74.74%**	**75.70%**

Tabla 4: Performance promedio en el set de validación cruzada de los modelos de Machine Learning estimados para los diferentes procesos de preprocesamiento en los tickers elegidos. Producción propia.

Los retornos de inversión porcentual promedio según cada estrategia para cada ticker se muestran en la tabla 5, de manera de poder comparar a nivel general estos enfoques diferentes bajo un mismo indicador que revela el nivel de ingresos generados a través de las estrategias.

Ticker	Compro primer día y vendo el ultimo	Predicción del precio al cierre	Predicción del % de cambio del precio al cierre	Predicción de la dirección del cambio del precio al cierre
AAL	-12.22	-9.52	0.00	0.00
AAPL	84.37	12.90	33.80	0.00
AMZN	18.82	5.53	7.18	-0.11
BA	3.20	2.29	21.75	0.00
CTSH	-1.23	0.00	0.00	0.55
GOOG	26.46	0.00	17.21	4.98
M	-42.38	-13.91	0.00	-8.89
MSFT	57.56	0.00	9.34	14.26
NVDA	72.40	53.46	26.48	18.54
TSLA	37.30	12.91	17.03	28.45
Promedio General	**24.43**	**6.37**	**13.28**	**5.78**

Tabla 5: Retornos a la Inversión porcentuales promedio de cada una de las estrategias aplicadas en los tickers elegidos. Producción propia.

La figura 4 muestra una perspectiva diferente de los resultados obtenidos al aplicar las estrategias de trading antes descriptas al resaltar el mayor valor de retorno a la inversión de entre los distintos modelos utilizados en una misma estrategia.

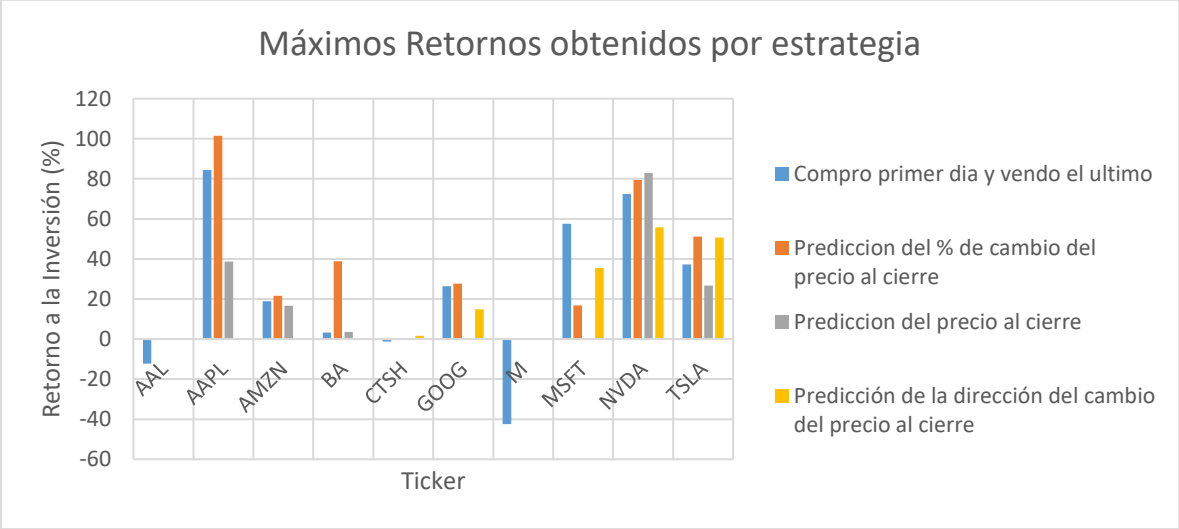

Figura 4: Retornos a la Inversión porcentuales máximos obtenidos de cada una de las estrategias aplicadas en los tickers elegidos. Producción propia.

Finalmente, la tabla 6 expone los máximos valores de Sharpe Ratio alcanzados por cada estrategia en cada acción y resalta a modo de mapa de calor en color verde a la estrategia más exitosa y en color rojo a la menos exitosa.

Ticker	Compro primer día y vendo el ultimo	Predicción del precio al cierre	Predicción del % de cambio del precio al cierre	Predicción de la dirección del cambio del precio al cierre	Maximo Sharpe Ratio
AAL	-5.45	0.00	0.00	0.00	**0.00**
AAPL	51.07	23.43	61.39	0.00	**61.39**
AMZN	13.98	12.34	16.01	0.00	**16.01**
BA	1.78	1.99	21.60	0.00	**21.60**
CTSH	-0.83	0.00	0.00	1.11	**1.11**
GOOG	18.27	0.00	19.10	10.30	**19.10**
M	-16.29	0.00	0.00	0.00	**0.00**
MSFT	46.24	0.00	13.45	28.49	**46.24**
NVDA	28.41	32.53	31.17	21.91	**32.53**
TSLA	12.16	8.70	16.65	16.50	**16.65**

Tabla 6: Sharpe Ratio máximos de cada una de las estrategias aplicadas en los tickers elegidos. Producción propia.

11. Discusión

Al analizar los resultados de la tabla 3, se puede observar que las estimaciones de los precios al cierre son las más precisas a nivel modelo que las estimaciones del cambio porcentual en el precio de cierre. A priori, es esperable que esto se traduzca en un mayor retorno durante la simulación. Si bien el cálculo del error es diferente en el caso de clasificación, se puede notar que la performance es notablemente más baja que en los otros dos casos. Estas observaciones permiten también sospechar que los modelos que predicen el precio tienden más fácilmente a caer en Overfitting.

En cuanto a los métodos de preprocesamiento, se desprende de los resultados expuestos en la tabla 4 que los 3 llevan a resultados muy similares en cuanto a la performance que producen en los modelos. Resulta ligeramente ventajoso tomar la estrategia de preprocesar las variables con la técnica de PCA que el resto de los métodos evaluados, pero sin evidencias concluyentes a nivel estadístico.

En todos los casos, es importante destacar que surgieron como mejor modelo todos los modelos probados inicialmente con una gran variedad de hiperparámetros luego de optimizados. Igualmente, la técnica más utilizada para las simulaciones debido a su superadora performance frente a las otras en la mayoría de los casos fue XGBoost.

Si se comparan los resultados observados en la tabla 3 con los expuestos en la tabla 5, se puede ver que, si bien los modelos con mayor performance son los que predicen el precio al cierre del activo, los modelos que generaron mayor retorno promedio son los que predicen la magnitud de su cambio porcentual respecto al día anterior. Esto puede deberse a una infinidad de razones, entre las cuales se encuentra la restricción presupuestaria inicial de la simulación. Al ser este ejercicio codificado con una lógica de todo o nada, el inversor simplemente invierte todo su efectivo en la compra de la acción cuando el modelo le dice de hacerlo (si es que tiene efectivo disponible) y vende toda su posición (si es que su dinero está invertido) en caso de que el modelo lo decida. Estas pequeñas salvedades son claves para entender cómo un modelo que parece tener mejor performance en lo técnico es capaz de generar menores ganancias. Si el modelo indica que es buen momento de comprar, pero el inversor no tiene efectivo para hacerlo, entonces esas ganancias no podrán ser debidamente atribuidas a esa decisión en particular cuando el modelo de estimación sí está genuinamente anotando un punto a su favor. Esto radica en la secuencialidad temporal de los hechos de la simulación y en el hecho de que el modelo no está diseñado para elegir el mejor momento de inversión sino para definir día a día si conviene invertir o desinvertir el dinero. Este hecho es un hallazgo muy interesante ya que permite separar la performance técnica de los modelos subyacentes de la ganancia posteriormente obtenida al simular las estrategias de inversión. Es más, en algún punto pone en duda el proceso de elegir los mejores modelos en base a su performance técnica antes de hacerlos pasar por la simulación en lugar de utilizar todos los modelos disponibles para simular y elegir el que mejor resultado económico tenga.

Si bien utilizar los retornos a la inversión promedio es un análisis válido para comparar las diferentes estrategias, es importante entender individualmente cuál estrategia es la que más retorno genera.

La figura 4 muestra los máximos retornos obtenidos con ese fin en particular. Se puede desprender de este gráfico que la estrategia que mayor ganancia genera consistentemente en la mayoría de las acciones es la de la utilización de la predicción del porcentaje de cambio del precio al cierre a través de técnicas de regresión. Sin embargo, es importante medir el riesgo-retorno y no únicamente la ganancia para poder determinar la estrategia más conveniente. La tabla 6 muestra, a través de los Sharpe Ratio máximos obtenidos por estrategia, que la estrategia que maximiza el retorno a la vez de minimizar el riesgo es la predicción del cambio porcentual del precio al cierre. De las 10 acciones utilizadas, 7 muestran que esta estrategia es más conveniente que las demás. En el caso del ticker CTSH, la estrategia más ventajosa resulta ser la predicción de la dirección del cambio de precio a través de técnicas de clasificación. En el caso de NVDA, la mejor estrategia resulta la predicción del precio al cierre, y en el caso de MSFT la estrategia base, que consiste en comprar el primer día y vender el último, es la superadora de todas las demás.

Si se analizan las acciones por grupos, lo que se desprende es que en los casos que presentan una pérdida neta a lo largo del período de simulación es posible evitar esa pérdida a través de las técnicas desarrolladas de trading algorítmico. Tanto en AAL como en M todas las estrategias basadas en Machine Learning protegen al inversor de la pérdida. En los casos donde el cambio de precio al cierre a lo largo del período fue despreciable (menor al 5%), los modelos son capaces de captar pequeñas ganancias a lo largo del mismo para obtener una ganancia neta positiva, como se ve en BA y CTSH. Respecto al resto de los tickers, cuyos precios aumentaron durante el período, se puede observar en la mayoría que a través de los métodos desarrollados se logran mayores ganancias y mayor Sharpe Ratio que en la estrategia base. Es importante notar también que en ningún caso las estrategias de trading algorítmico generaron pérdidas para el inversor.

12. Conclusiones

En resumen, el estudio realizado muestra evidencias concluyentes a favor de las técnicas de trading algorítmico, cuando comparadas a un inversor pasivo que deja su dinero en una acción sin mayores evaluaciones. La utilización de modelos de Machine Learning para definir estrategias de compra y venta de acciones individuales permite obtener mejores resultados en términos de riesgo-retorno que la estrategia base, a la vez que funciona como escudo ante la pérdida de capital en activos que presentan tendencias a la baja. Finalmente, este trabajo cumplió con el objetivo de comparar con un criterio común la performance de estrategias de trading basadas en técnicas de regresión y clasificación refutando la hipótesis inicial que sostenía que los modelos de clasificación darían mejores resultados que los de regresión. Quedó fuera del alcance de este trabajo la elaboración de un modelo que permita parametrizar y optimizar el margen de cobertura ante la volatilidad de la predicción, y la ampliación del modelo desarrollado hacia una orientación de cartera multi producto en lugar de un modelo de una única acción a la vez en la que sea posible reevaluar la totalidad de la posición diariamente y realizar ajustes en consecuencia. Estos puntos serán objeto de futuras investigaciones en el marco de la misma temática.

13. Referencias-Bibliografía

[1] L. A. Teixeira and A. L. I. De Oliveira, “A method for automatic stock trading combining technical analysis and nearest neighbor classification,” *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6885–6890, 2010, doi: 10.1016/j.eswa.2010.03.033.

[2] D. Van Den Poel, C. Chesterman, M. Koppen, and M. Ballings, “Equity price direction prediction for day trading: Ensemble classification using technical analysis indicators with interaction effects,” *2016 IEEE Congr. Evol. Comput. CEC 2016*, pp. 3455–3462, 2016, doi: 10.1109/CEC.2016.7744227.

[3] R. Dash and P. K. Dash, “A hybrid stock trading framework integrating technical analysis with machine learning techniques,” *J. Financ. Data Sci.*, 2016, doi: 10.1016/j.jfds.2016.03.002.

[4] B. M. Henrique, V. A. Sobreiro, and H. Kimura, “Literature review: Machine learning techniques applied to financial market prediction,” *Expert Syst. Appl.*, vol. 124, pp. 226–251, 2019, doi: 10.1016/j.eswa.2019.01.012.

[5] A. Sharma, D. Bhuriya, and U. Singh, “Survey of stock market prediction using machine learning approach,” *Proc. Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2017*, vol. 2017-Janua, pp. 506–509, 2017, doi: 10.1109/ICECA.2017.8212715.

[6] S. Borovkova and I. Tsiamas, “An ensemble of LSTM neural networks for high-frequency stock market classification,” *J. Forecast.*, vol. 38, no. 6, pp. 600–619, 2019, doi: 10.1002/for.2585.

[7] E. A. Gerlein, M. McGinnity, A. Belatreche, and S. Coleman, “Evaluating machine learning classification for financial trading: An empirical approach,” *Expert Syst. Appl.*, vol. 54, pp. 193–207, 2016, doi: 10.1016/j.eswa.2016.01.018.

[8] P. Vats and K. Samdani, “Study on machine learning techniques in financial markets,” *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, pp. 1–5, 2019, doi: 10.1109/ICSCAN.2019.8878741.

[9] F. D. Paiva, R. T. N. Cardoso, G. P. Hanaoka, and W. M. Duarte, “Decision-making for financial trading: A fusion approach of machine learning and portfolio selection,” *Expert Syst. Appl.*, vol. 115, pp. 635–655, 2019, doi: 10.1016/j.eswa.2018.08.003.

[10] D. Shah, H. Isah, and F. Zulkernine, “Stock market analysis: A review and taxonomy of prediction techniques,” *Int. J. Financ. Stud.*, vol. 7, no. 2, 2019, doi: 10.3390/ijfs7020026.