UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas

PROYECTO FINAL DE CARRERA

DESARROLLO DE UN SISTEMA INTELIGENTE PARA LA TOMA DE DECISIONES DE INVERSIÓN EN EL MERCADO FINANCIERO UTILIZANDO HERRAMIENTAS DE MACHINE LEARNING

Alumno: Bartezaghi, Catriel

Director: Robledo, Miguel Angel

Santa Fe, Diciembre de 2023

Resumen 4
1. Introduccion 5
1.1. Historia y evolución del análisis en los mercados financieros 5
1.2. Transición de métodos tradicionales a técnicas más avanzadas 6
2. Marco conceptual 8
2.1. Contexto General 8
2.2. Arquitecturas 10
2.3. Variables Macroeconómicas 17
2.4. Análisis Fundamental 20
2.5. Cambio logarítmico 22
2.6. Datos 24
2.7. Variable Objetivo 29
2.8. Herramientas Tecnológicas 30
3. Selección de Datos 34
3.1. Yahoo Finance 35
3.2. The Federal Reserve Economic Data (FRED) 38
3.3. The Securities and Exchange Commission's (SEC) 41
3.4. The Bureau of Economic Analysis (BEA) 48
4. Dataset 58
4.1. Temporalidad de los datos 58
4.2. Datos de entrada 60
4.3. Variable Objetivo 62
4.4. Armado de Dataset 64
4.5. Escalado de Datos 65
4.6. Validacion 66
4.7. Métricas 68
5. Entorno de trabajo 70
5.1. Colab 70
5.2. Entorno Personal 71
6. Aprendizaje 75
6.1. Enfoque de Datos Integrados 75
6.2. Enfoque con Modelo Ensamblado 82
7. Descripción y Funcionamiento del Sistema 87
7.1. Módulo de Configuración 87
7.2. Módulo de Adquisición de Datos 88
7.3. Módulo de Procesamiento de Datos 89
7.4. Módulo de Red Neuronal 90
7.5. Módulo de Simulación y Resultados 91
8. Resultados 95
8.1. Yahoo Finance - GDP 95
8.2. Yahoo Finance - GDP - FRED 98

8.3. Yahoo Finance - GDP - FRED - SEC..... 103
9. Conclusiones..... 107
9.1. Evaluación de Modelos y Tecnologías Utilizadas..... 107
9.2. Limitaciones..... 108
9.3. Logro de objetivos..... 108
9.4. Trabajos futuros..... 109
Referencias..... 112

Quiero agradecer a Miguel, por su generosa disposición y tutoría, así como por asumir la dirección de este proyecto que capturó profundamente mi interés.

A mis amigos y compañeros de carrera, cuya ayuda y amistad durante estos años han sido invaluables, así como a mi familia y a todas las personas que me brindaron su apoyo y aliento constante.

Finalmente, dedico un agradecimiento especial a mis padres, quienes han sido una fuente inagotable de apoyo y cuyo sacrificio me ha permitido llegar hasta aquí.

Catriel

Resumen

El objetivo de este proyecto consiste en el desarrollo de un sistema inteligente enfocado en predecir cambios en los precios de activos financieros. La meta es proporcionar a los inversores herramientas avanzadas para identificar patrones y tendencias en el mercado financiero, mejorando así sus decisiones de inversión. Para lograrlo, se realiza un análisis detallado de datos relevantes que influyen en el precio de los activos, empleando técnicas de machine learning. El sistema propuesto busca combinar el análisis técnico y fundamental, utilizando algoritmos de aprendizaje automático para procesar y analizar grandes cantidades de datos, y así generar predicciones precisas de los movimientos del mercado.

1. Introduccion

1.1. Historia y evolución del análisis en los mercados financieros.

La práctica de análisis en los mercados financieros tiene sus raíces en la aparición de los primeros mercados de valores en el siglo XVII. La Bolsa de Amsterdam, fundada en 1602, es un ejemplo temprano, representando la primera corporación que ofreció acciones al público. En este contexto, los inversores comenzaron a darse cuenta de la necesidad de métodos sistemáticos para evaluar el valor y el potencial de las acciones, lo que llevó al desarrollo inicial del análisis financiero.

A principios del siglo XX, Charles H. Dow, cofundador del Wall Street Journal, formuló una serie de principios que más tarde se conocieron como la Teoría de Dow [1]. Esta teoría sentó las bases del análisis técnico moderno al sugerir que los precios de los activos no son aleatorios y que se mueven en tendencias identificables. Posteriormente, figuras como William P. Hamilton, Robert Rhea, y más tarde, John Magee y Robert Edwards, contribuyeron significativamente a la evolución y popularización del análisis técnico. El análisis técnico está fundamentado en la idea de que los precios de los activos financieros se mueven en tendencias identificables y predecibles basadas en la historia de precios y volúmenes. Su objetivo principal es identificar patrones y tendencias en el movimiento de los precios, con el fin de predecir alteraciones en la dinámica del mercado.

Por otro lado, el análisis fundamental ganó prominencia en la década de 1930, especialmente con el trabajo de Benjamin Graham y David Dodd denominado "Security Analysis" [2] publicado en 1934. Este libro sentó las bases para evaluar las acciones basándose en el estudio de los estados financieros de una empresa, sus perspectivas de crecimiento, calidad de la gestión, y el entorno económico general. Esta serie de datos permite calcular el valor intrínseco de la empresa lo que supone un indicador del rendimiento futuro que se espera del activo.

Durante la segunda mitad del siglo XX, la Teoría del Mercado Eficiente (EMT) [3] propuesta por Eugene Fama en la década de 1960 desafió algunas de las premisas subyacentes tanto al análisis técnico como al fundamental. Según la EMT, toda la información disponible se refleja en los precios de los activos, lo que implica que no es posible superar consistentemente el rendimiento del mercado a través del análisis tradicional. Esto llevó a un debate continuo y al desarrollo de enfoques híbridos que intentan combinar diferentes métodos de análisis.

1.2. Transición de métodos tradicionales a técnicas más avanzadas

Hacia el final del siglo XX, el análisis de mercados financieros experimentó refinamientos significativos en ambos enfoques, técnico y fundamental. En el ámbito del análisis técnico, las innovaciones incluyeron el desarrollo de sistemas de trading computarizados y algoritmos avanzados que permitieron análisis más rápidos y precisos de los patrones de mercado. Esto no solo aumentó la eficiencia del análisis técnico sino que también abrió la puerta a estrategias de trading de alta frecuencia, que aprovechaban las pequeñas fluctuaciones de precios en plazos muy cortos.

Por su parte, el análisis fundamental también evolucionó, integrando enfoques más holísticos para evaluar empresas. Esto incluyó un mayor énfasis en el análisis cualitativo, como la calidad de la gestión y la estrategia empresarial, además de los tradicionales indicadores cuantitativos. La proliferación de la información disponible, gracias a la globalización y el auge de Internet, amplió significativamente el alcance y la profundidad del análisis fundamental.

A medida que ambos métodos maduraban, se hacía cada vez más evidente que, a pesar de sus fortalezas, existían limitaciones inherentes en los enfoques tradicionales. Estas limitaciones, junto con los desafíos planteados por la Teoría de la Eficiencia del Mercado, provocaron un creciente interés en desarrollar técnicas más sofisticadas y avanzadas. Los inversores y analistas comenzaron a buscar métodos que pudieran integrar la vasta cantidad de datos disponibles y proporcionar análisis más precisos y predictivos.

Este cambio de paradigma llevó a la exploración de técnicas innovadoras, como la inteligencia artificial y el aprendizaje automático, que prometían superar algunas de las limitaciones de los métodos tradicionales. La transición de los métodos tradicionales a estas técnicas más avanzadas marcó el comienzo de una nueva era en el análisis de mercados financieros, caracterizada por una mayor velocidad, eficiencia y capacidad para manejar y analizar grandes conjuntos de datos.

2. Marco conceptual

2.1. Contexto General

La inteligencia artificial (IA) ha revolucionado la manera en que se operan los mercados financieros. Gracias a su capacidad para procesar y analizar grandes volúmenes de datos a una velocidad y precisión inigualables, la IA ha permitido a los inversores y analistas obtener perspectivas más profundas y realizar predicciones más precisas. Los sistemas basados en IA pueden identificar patrones ocultos en los datos del mercado, anticipar cambios en las tendencias de inversión y optimizar estrategias para maximizar rendimientos y minimizar riesgos. Este avance tecnológico no solo ha mejorado la eficiencia y efectividad de las operaciones en el mercado, sino que también ha abierto nuevas vías para la investigación y desarrollo en el ámbito financiero.

Esta transformación impulsada por la IA en los mercados financieros no es un fenómeno aislado, sino parte de una tendencia más amplia de digitalización y automatización en el mundo financiero. La integración de tecnologías avanzadas en las operaciones financieras ha dado lugar a lo que se conoce como FinTech, un campo en el que la innovación tecnológica está redefiniendo los servicios y productos financieros. En este entorno, la IA actúa como un motor clave, potenciando desde el trading algorítmico hasta la gestión de riesgos y el asesoramiento financiero personalizado.

Además, la creciente disponibilidad y accesibilidad de los datos financieros, impulsada por la era del big data, ha jugado un papel crucial en la facilitación de estos avances. Con cantidades masivas de datos generados cada segundo, la IA y el machine learning ofrecen herramientas indispensables para extraer valor de este vasto océano de información. Estas tecnologías no solo permiten analizar datos históricos y en tiempo real de manera eficiente, sino que también proporcionan la capacidad de prever tendencias futuras y reaccionar a eventos del mercado casi en tiempo real.

Dentro del marco conceptual del machine learning aplicado a los mercados financieros, el forecasting [4] o pronóstico financiero emerge como una de las aplicaciones más relevantes. El forecasting consiste en hacer predicciones basadas en datos pasados y presentes. En el ámbito financiero, algunas variables de interés son los precios futuros o movimientos del mercado, los cuales son fundamentales para la toma de decisiones estratégicas en inversiones y operaciones financieras. Esta práctica se apoya en la premisa de que los patrones históricos y las tendencias de los datos pasados pueden ofrecer insights sobre futuros comportamientos del mercado.

Sin embargo, se debe destacar los desafíos que presenta, el uso de machine learning en el forecasting. Los modelos dependen de la calidad y la integridad de los datos disponibles, y su interpretación requiere un equilibrio entre el conocimiento técnico y la comprensión del mercado. Además, cuestiones como el sobreajuste y la generalización de los modelos son aspectos que deben manejarse con cuidado.

Dentro de este marco conceptual, el presente proyecto se enfoca en explorar la capacidad del machine learning para mejorar la precisión y eficiencia en las predicciones financieras. Para ello se desarrollará un sistema de aprendizaje automático que, utilizando datos históricos de precios y volúmenes, variables macroeconómicas e indicadores fundamentales, sea capaz de analizar y predecir la evolución futura de activos financieros, con el objetivo de ayudar al inversor en la toma de decisiones. En este contexto, por “evolución futura” se entiende la predicción de cambios en el valor de los activos, es decir, si el precio tenderá a subir o bajar en un determinado horizonte temporal. El objetivo es proporcionar al inversor información útil que le permita tomar decisiones de compra o venta en base a esas proyecciones.

Como se profundizará más adelante, en este proyecto se realizará la predicción de la variación diaria del precio de los activos. Al definir un horizonte de un día, se busca aprovechar al máximo la disponibilidad de datos recientes y por lo tanto mejorar la precisión de las proyecciones.

2.2. Arquitecturas

Si bien la inteligencia artificial (IA) abarca una amplia gama de tecnologías y metodologías, como el procesamiento del lenguaje natural, la visión por computadora, los sistemas expertos, y el aprendizaje reforzado, en este caso nos centraremos en analizar las tecnologías principales del Deep Learning, que son particularmente relevantes debido a su amplio crecimiento y uso en los últimos años.

El Deep Learning [5], una rama avanzada del machine learning, se distingue por su capacidad para construir y entrenar redes neuronales profundas, que imitan la forma en que el cerebro humano procesa información. Esta aproximación se ha convertido en una herramienta poderosa y versátil para abordar problemas complejos de predicción y clasificación en diversos campos, incluyendo el financiero.

A continuación se desarrollan algunas de las arquitecturas que se estudiaron y consideraron para el desarrollo de este proyecto.

Redes Neuronales de Perceptrón Multicapa (MLP)

Las Redes Neuronales de Perceptrón Multicapa (MLP) son una de las arquitecturas más básicas y tradicionales en el campo del Deep Learning. Compuestas por varias capas, las MLP conectan cada neurona en una capa con todas las neuronas en la capa anterior y siguiente, formando una estructura densamente enlazada. Esta configuración permite a las MLP capturar y modelar relaciones complejas en los datos, siendo eficaces en tareas de regresión y clasificación.

Sin embargo, una limitación clave de las MLP para el forecasting, es su incapacidad para manejar dependencias temporales. Al no tener memoria de entradas anteriores, las MLP no pueden procesar eficientemente secuencias temporales, lo cual es crucial en la predicción de series temporales [6] donde el contexto y el orden temporal son esenciales para realizar predicciones precisas. Además, su susceptibilidad al overfitting y la dificultad para capturar las

complejidades inherentes a los datos financieros hacen que las MLP no sean siempre la opción más adecuada para el análisis predictivo en este campo. [7]

Redes Neuronales Convolucionales (CNN)

Las Redes Neuronales Convolucionales (CNN) representan una clase avanzada de modelos en el campo del Deep Learning, destacándose principalmente en el procesamiento y análisis de datos visuales. Su arquitectura está especialmente diseñada para detectar y manipular patrones espaciales en los datos, lo que las hace idóneas para tareas como el reconocimiento de imágenes y la visión por computadora. En una CNN, las capas de convolución actúan como filtros que pueden captar características específicas en los datos, como bordes o texturas en imágenes. Estas características son luego agrupadas por las capas de pooling, que reducen la dimensionalidad de los datos, manteniendo sólo los aspectos más relevantes. La combinación de estas capas permite a las CNN construir una comprensión jerárquica de los datos, aprendiendo desde patrones simples hasta representaciones complejas. [8]

Un estudio sobre modelos de predicción de acciones usando redes CNN [9], encontró que la configuración óptima con 5 o 10 núcleos de convolución alcanzó un error cuadrático medio [10] de 4.5459e-04, lo cual representa una alta precisión en la predicción.

Sin embargo, al igual que en las redes MLP, las CNN no suelen ser la opción más adecuada en el contexto del forecasting en mercados financieros. Su enfoque en patrones espaciales las hace menos eficientes en capturar y procesar dependencias temporales, que son fundamentales en la predicción de series temporales financieras. Mientras que en el análisis de imágenes, las CNN son excelentes en identificar elementos visuales y sus relaciones espaciales, esta habilidad no se traduce directamente a la comprensión de secuencias temporales y patrones en datos financieros, donde el orden y la continuidad temporal son críticos.

Redes Neuronales Recurrentes (RNN)

Las Redes Neuronales Recurrentes (RNN) son una clase de redes neuronales diseñadas para manejar datos secuenciales, como series temporales. A diferencia de las redes tradicionales, las RNN tienen la capacidad única de mantener una "memoria" de las entradas anteriores en su estructura interna, lo que les permite procesar no solo la entrada actual sino también incorporar el contexto proporcionado por las entradas previas. Esta característica las convierte en candidatas ideales para tareas donde el orden y la relación temporal entre los datos son cruciales.

La estructura de una RNN incluye nodos o unidades que forman un bucle, permitiendo que la información fluya de una etapa de la red a la siguiente. Esta recurrencia facilita que la red capture dependencias temporales, es decir, cómo los eventos en un momento dado están influenciados por eventos en momentos anteriores. Cada unidad en una RNN utiliza información del estado anterior y la entrada actual para calcular su salida, lo que efectivamente crea una forma de "memoria" sobre lo que la red ha procesado hasta el momento.

Estas características la convierten en una opción ideal para realizar predicciones en los mercados financieros, ya que los datos de este ámbito suelen tener una naturaleza temporal con patrones recurrentes. Las RNN están diseñadas específicamente para analizar y anticipar este tipo de comportamientos.

A pesar de sus ventajas, las RNN también tienen limitaciones, como la dificultad para aprender dependencias a muy largo plazo, un fenómeno conocido como el problema de la desaparición del gradiente. Para abordar esto, se han desarrollado variantes como Long Short-Term Memory (LSTM) y Gated Recurrent Units (GRU), que incorporan mecanismos para controlar y mantener la información relevante a lo largo de periodos extensos. [11]

Redes Transformer

Las redes Transformer, introducidas en el campo del procesamiento del lenguaje natural, han revolucionado el enfoque en el aprendizaje profundo gracias a su

capacidad única para manejar secuencias de datos. A diferencia de las RNN y las LSTM, las redes Transformer no procesan los datos secuencialmente, sino que utilizan mecanismos de atención para procesar toda la secuencia de datos a la vez. Esto permite que capturen dependencias a largo plazo y relaciones complejas en los datos con mayor eficiencia.

El componente central de las redes Transformer es el mecanismo de atención, que les permite enfocarse en diferentes partes de la secuencia de datos para calcular una representación más rica y contextual. Este mecanismo de atención, fundamental en las redes Transformer, otorga la capacidad de ponderar y enfocarse en elementos específicos de la secuencia, permitiendo la identificación de patrones relevantes en conjuntos de datos extensos. [12]

Los avances significativos atribuidos a las redes Transformer, especialmente en el desarrollo de modelos como GPT (Generative Pretrained Transformer), han marcado un hito en el campo del procesamiento del lenguaje natural (PLN). GPT y sus sucesores han demostrado una capacidad excepcional para generar texto coherente y contextualmente relevante, superando los desafíos presentes en los enfoques más antiguos.

Pero aunque los avances de los modelos basados en Transformers, han sido revolucionarios en el procesamiento del lenguaje natural, ofreciendo nuevas capacidades en generación de texto y comprensión contextual, su eficacia no se extiende universalmente a todas las aplicaciones de aprendizaje automático. Específicamente, en tareas como el pronóstico de series temporales a largo plazo, los estudios recientes, como el mencionado en el trabajo "Are Transformers Effective for Time Series Forecasting?" [13], han mostrado que los modelos Transformer no son capaces de superar a modelos simples como los modelos lineales de una sola capa en tareas de forecasting. Esto se debe a que el mecanismo de auto-atención de los Transformers, aunque eficiente en captar correlaciones semánticas, pierde información temporal debido a su naturaleza invariante a la permutación, lo cual es crucial en series temporales.

Elección de arquitectura

Tras una evaluación exhaustiva de diversas arquitecturas de redes neuronales, se ha determinado que las Redes Neuronales Recurrentes (RNN), y en particular las Redes de Memoria a Corto y Largo Plazo (LSTM), son las más adecuadas para la tarea de predecir los precios de las acciones. Esta elección se basa en la capacidad superior de las RNN para manejar secuencias de datos, una característica fundamental en el análisis de series temporales financieras.

Un estudio comparativo, denominado "Stock price forecast with deep learning" [14], comparó el desempeño de las arquitecturas de redes neuronales completamente conectadas, convolucionales y recurrentes en la predicción del valor del índice S&P 500 para el día siguiente basándose en sus valores previos. La conclusión en el estudio citado fue que una red neuronal recurrente de una sola capa con el optimizador RMSprop proporcionaba los mejores resultados en términos de Error Absoluto Medio (MAE) en la predicción del valor del índice S&P 500 para el día siguiente, basándose en sus valores previos.

Adicionalmente, los modelos LSTM, una variante especializada de las RNN, han sido reconocidos por su habilidad para manejar dependencias a largo plazo, según lo descrito en "Stock price prediction using the RNN model" [15]. Esta característica es crucial para predecir precios de acciones, ya que permite a las LSTM retener y procesar información relevante a lo largo de períodos extensos, facilitando un análisis más profundo y preciso de las tendencias del mercado. Esta habilidad para superar los desafíos de dependencia a largo plazo en las series temporales hace que las LSTM sean particularmente aptas para el análisis de mercados financieros, donde los patrones de datos y las influencias pasadas juegan un papel importante en la formación de las tendencias futuras.

Redes LSTM

Las redes neuronales recurrentes (RNN) son modelos diseñados para procesar datos secuenciales, permitiendo que la información persista a lo largo del tiempo. Sin

embargo, las RNN tradicionales pueden enfrentar dificultades al aprender dependencias a largo plazo debido al problema del desvanecimiento o explosión del gradiente. Para superar este desafío, se han desarrollado arquitecturas avanzadas como las Redes de Memoria a Largo Corto Plazo (LSTM, por sus siglas en inglés).

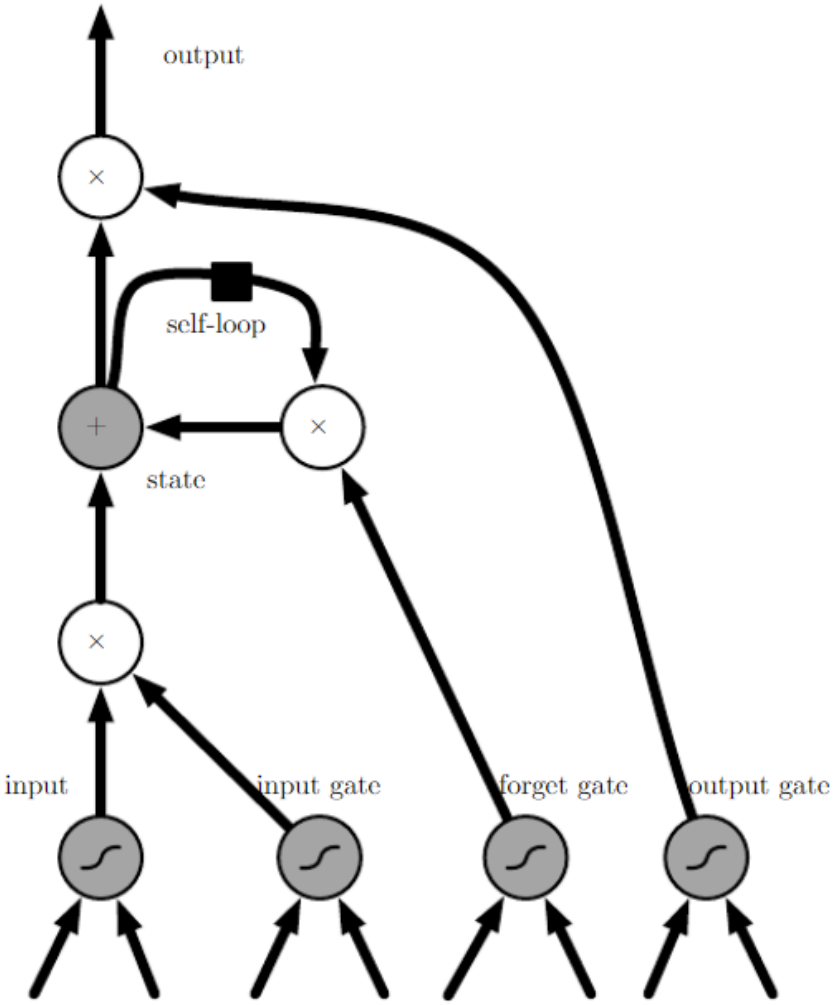

Fig. 1 Diagrama en bloque de LSTM

De Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Capítulo 10: Sequence Modeling: Recurrent and Recursive Nets. MIT Press. [1]

Las LSTM introducen una celda de memoria que puede mantener información durante largos períodos. Esta celda está controlada por tres mecanismos clave llamados puertas, que regulan el flujo de información dentro y fuera de la celda:

[1] De Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Capítulo 10: Sequence Modeling: Recurrent and Recursive Nets. MIT Press. https://www.deeplearningbook.org/

1. **Puerta de olvido**: Decide qué información debe descartarse de la celda. Se calcula utilizando una función sigmoide que toma como entrada el estado oculto anterior **h** y la entrada actual **x**:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right)$$

Donde **b**, **U** y **W** denotan respectivamente los sesgos, los pesos de entrada y los pesos recurrentes en la celda LSTM

El estado de la celda interna es actualizado de la siguiente manera utilizando un peso de bucle propio **f**:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right)$$

2. **Puerta de entrada**: Determina qué información nueva se almacenará en la celda. Se calcula de manera similar a la puerta de olvido pero con sus propios parámetros:

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right)$$

3. **Puerta de salida:** Controla la salida de información de la celda hacia el siguiente estado oculto. Se calcula como:

$$h_i^{(t)} = \tanh\left(s_i^{(t)}\right) q_i^{(t)},$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right)$$

La cual tiene también sus propios parámetros **b**, **U** y **W.**

Aquí, **q** controla cuánto del estado de la celda **s** pasará a la salida.

Gracias a su estructura de puertas y memoria interna, las LSTM pueden aprender y mantener información relevante durante largos intervalos, lo que las hace efectivas para tareas que involucran dependencias a largo plazo en secuencias, como la predicción en series temporales [16].

Por estos fundamentos es que se decidió adoptar esta red para realizar predicciones en este trabajo.

2.3. Variables Macroeconómicas

Las variables macroeconómicas son indicadores clave que reflejan la salud y el desempeño de una economía en su conjunto. Su influencia en los mercados financieros es significativa, ya que afectan las expectativas de los inversores y, por ende, los precios de los activos. A continuación, se detallan las principales variables macroeconómicas y su impacto.

Producto Interno Bruto (PIB)

El PIB es una medida del valor total de los bienes y servicios producidos en una economía durante un período específico. Habitualmente se calcula por país y en periodos anuales. Se puede calcular como:

$$PIB = C+I+G+X-M$$

Donde las iniciales indican Consumo, Inversión, Gasto público, Exportaciones e Importaciones. [17]

Un crecimiento fuerte del PIB generalmente indica una economía saludable, lo que puede generar mayores ganancias corporativas y mayores retornos en el mercado de valores. Los inversores suelen considerar que un crecimiento positivo del PIB es una señal de condiciones económicas favorables, lo que puede aumentar la confianza de los inversores y hacer subir los precios de las acciones [18].

Inflación

La inflación es un índice que mide el aumento generalizado de los precios de bienes y servicios. Niveles moderados de inflación son considerados normales en una economía en crecimiento, pero un alto nivel afecta negativamente tanto a los consumidores como a las empresas.

En los mercados financieros, la inflación tiende a generar volatilidad y afecta de manera diferenciada a las acciones de valor y crecimiento.

Durante periodos de alta inflación, el gasto de los consumidores disminuye, lo que reduce los ingresos corporativos y presiona los márgenes de ganancia. Las acciones de valor suelen desempeñarse mejor en estos contextos, ya que representan empresas con flujos de caja más estables. En cambio, las acciones de crecimiento, que dependen de expectativas de ganancias futuras, sufren más debido al aumento de las tasas de interés que suele acompañar la inflación, lo que reduce su atractivo para los inversores.

Históricamente, los mayores retornos reales en los mercados bursátiles han ocurrido cuando la inflación se mantiene entre el 2% y el 3%. Por encima o por debajo de este rango, la incertidumbre económica aumenta, afectando negativamente el rendimiento de las acciones y aumentando la volatilidad [19]. Esto resalta la relevancia del índice de inflación en la toma de decisiones financieras.

Tasas de Interés

Las tasas de interés, establecidas por los bancos centrales, son un instrumento clave de la política monetaria. Tasas de interés más bajas facilitan el crédito y pueden estimular la inversión y el consumo, mientras que tasas de interés más altas encarecen el financiamiento logrando el efecto opuesto.

Las tasas de interés y el mercado de valores suelen moverse en direcciones opuestas. Cuando un banco central incrementa las tasas de interés, eleva los costos de endeudamiento tanto para los consumidores como para las empresas. Las instituciones financieras, como los bancos, también aumentan las tasas que cobran a los prestatarios, lo que encarece el crédito para empresas y particulares. Este aumento en los costos de financiación reduce la capacidad de gasto de los consumidores, lo que a su vez afecta los ingresos y las ganancias corporativas. A medida que disminuyen los ingresos, las valoraciones de las acciones tienden a bajar, lo que lleva a una caída en los precios de las acciones.

Por otro lado, cuando la Reserva Federal reduce las tasas de interés, las condiciones financieras se flexibilizan. El costo de pedir dinero prestado disminuye, lo que incentiva tanto a los consumidores como a las empresas a gastar e invertir más. Para las empresas, esto significa menores costos de financiamiento para proyectos, adquisiciones y expansiones, lo que puede aumentar sus ganancias futuras. Este aumento de la actividad económica genera una mayor confianza en el mercado de valores, lo que impulsa los precios de las acciones

Las acciones con mayor expectativa de crecimiento futuro, son más sensibles a los cambios en las tasas de interés. Dado que los inversores valoran estas acciones en función de sus flujos de caja futuros descontados, un aumento en las tasas de interés reduce el valor presente de esos flujos, lo que disminuye el atractivo de estas acciones en un entorno de tasas elevadas [20]

Índices de Confianza

Los índices de confianza, como el Índice de Confianza del Consumidor (ICC) son índices económicos que miden el nivel de optimismo o pesimismo de los consumidores respecto a la situación económica actual y futura. Este índice refleja las percepciones y expectativas de los consumidores sobre la economía, el mercado laboral, los ingresos y su capacidad de ahorro.

Estos índices tienen un impacto importante en los mercados financieros. Por ejemplo, un ICC bajo podría indicar una ralentización económica, lo que podría llevar a una disminución de las tasas de interés y una posible devaluación del dólar. En cambio, un ICC alto puede estar asociado con tasas de interés más elevadas y un dólar más fuerte. [21]

2.4. Análisis Fundamental

El análisis fundamental es una metodología utilizada para evaluar el valor intrínseco de un activo, como una acción, mediante el estudio de los estados financieros de una empresa y otros datos financieros más amplios. Este enfoque tiene como objetivo determinar el verdadero valor de una empresa, basándose en su salud financiera, las condiciones del mercado y el contexto económico general. Los inversores utilizan el análisis fundamental para decidir si invertir o no en una empresa, en función de su valor presente y su proyección futura. [22]

Principios del Análisis Fundamental

El análisis fundamental se centra en el examen de diversos aspectos internos y externos de una empresa. Estos incluyen:

- Crecimiento de ingresos y rentabilidad: Se analiza la capacidad de la empresa para generar ingresos y maximizar beneficios.
- Ventajas competitivas: El posicionamiento de la empresa dentro de su sector y la demanda de sus productos o servicios.

- Equipo directivo: Se evalúa la efectividad del liderazgo y su capacidad para gestionar desafíos y aprovechar oportunidades.
- Factores macroeconómicos: Se consideran variables externas como el estado general de la economía, la inflación y las tasas de desempleo.

Tipos de Análisis Fundamental

El análisis fundamental se divide principalmente en dos enfoques:

- Análisis cuantitativo: Se refiere a los datos numéricos, como las métricas y los ratios financieros que se derivan de los estados financieros de la empresa. Estos ratios incluyen:
    - Precio/beneficio (P/E): Relación entre el precio de la acción y las ganancias por acción.
    - Retorno sobre el capital (ROE): Indicador de la rentabilidad del capital invertido.
    - Deuda/capital (D/E): Indicador del nivel de endeudamiento de la empresa.
- Análisis cualitativo: Se centra en aspectos más intangibles, como la calidad del equipo directivo, la fortaleza de la marca y la ventaja competitiva. Estos factores pueden incluir la satisfacción de los empleados, la lealtad de los clientes y la reputación de la empresa en el mercado.

Ambos enfoques, cuantitativo y cualitativo, son fundamentales para comprender la verdadera situación de una empresa y hacer predicciones sobre su desempeño futuro.

Importancia del Análisis Fundamental

El análisis fundamental permite a los inversores centrarse en los factores subyacentes que impulsan las operaciones y el rendimiento a largo plazo de una empresa, evitando el ruido de las fluctuaciones de precios a corto plazo. Su principal valor radica en:

- Evaluar el valor real de una empresa: Al examinar los datos financieros, los inversores pueden obtener una visión clara de la rentabilidad, liquidez y estabilidad financiera de la empresa.
- Identificar oportunidades: Ayuda a identificar empresas infravaloradas que tienen potencial para crecer, lo que permite a los inversores capitalizar en tendencias a largo plazo.
- Detectar riesgos: Al investigar la salud financiera y la posición en el mercado de una empresa, se pueden evitar inversiones en acciones sobrevaloradas o con mayores probabilidades de bajo rendimiento.

Fuentes de Información para el Análisis Fundamental

Existen diversas fuentes donde los inversores pueden encontrar los datos fundamentales de una empresa, como:

- Informes financieros y presentaciones públicas.
- Bases de datos económicas y financieras como Bloomberg[2], FactSet[3] y Morningstar[4].
- Informes de investigaciones de corredores y firmas de inversión.

Cada una de estas fuentes ofrece una perspectiva complementaria, lo que permite al analista tener una visión completa y diversificada de la empresa que está evaluando.

Este enfoque proporciona una base sólida para tomar decisiones informadas sobre la compra, venta o retención de acciones en el mercado financiero.

2.5. Cambio logarítmico

El cambio logarítmico es una técnica utilizada en estadística y análisis de series temporales que convierte las variaciones absolutas en relativas mediante la aplicación del logaritmo natural a los datos. Esto es útil en modelos de Machine Learning por

[2] "Bloomberg." https://www.bloomberg.com/
[3] "FactSet." https://www.factset.com/
[4] "Morningstar." https://www.morningstar.com/

varias razones. Primero, estabiliza la varianza de las series temporales, lo cual es importante, ya que muchos modelos suponen homocedasticidad (varianza constante) [23]. Segundo, ayuda a reducir la heterocedasticidad, es decir, el impacto de los valores atípicos, permitiendo un análisis más suavizado de las tendencias [23]. Además, esta transformación facilita la comparación entre diferentes rangos de precios y, en ciertos casos, mejora la normalidad de los datos, lo que favorece el cumplimiento de los supuestos estadísticos de algunos algoritmos de Machine Learning [24].

La fórmula es la siguiente:

$$S(t) = ln\,[P(t) / P(t\text{-}1)]$$

Donde:

- S(t) representa el cambio logarítmico a tiempo t.
- ln es el logaritmo natural.
- P(t) precio del activo a tiempo t.
- P(t-1) precio del activo a tiempo t-1.

Es importante aclarar que se emplea el término "cambio logarítmico" para diferenciarlo de la "transformación logarítmica", la cual es ampliamente utilizada en finanzas para ajustar la escala de precios [25]. Mientras que la transformación logarítmica reescala los datos para un mejor análisis de variabilidad, el cambio logarítmico se enfoca en medir las variaciones porcentuales entre periodos consecutivos.

En este trabajo se utiliza el cambio logarítmico para predecir la dirección del movimiento del precio, pero es importante destacar que es posible recuperar el valor original del precio aplicando la transformación inversa:

$$P(t) = P(t-1) * exp(S(t))$$

Esta transformación permite estimar el precio futuro basado en el cambio logarítmico predicho. Sin embargo, en este estudio no se aplicó esta transformación inversa debido a que el interés principal radica en determinar la dirección del cambio de precio (alza o baja) y la magnitud de dicho cambio para la toma de decisiones en estrategias de trading. Al utilizar directamente el cambio logarítmico, se obtiene una medida más representativa y útil para decidir acciones de compra o venta.

2.6. Datos

La precisión de cualquier modelo de Machine Learning (ML) depende enormemente de la calidad y relevancia de los datos con los que se entrena. Datos precisos, actualizados y bien estructurados son cruciales para desarrollar modelos predictivos fiables y útiles en el mercado financiero.

En el ámbito del análisis financiero, diversos datos impactan en las decisiones de los inversores y, consecuentemente, en el valor de los activos. Estos datos incluyen información derivada de los análisis realizados por los operadores durante sus actividades de trading, así como aquellos que proporcionan una visión general de la situación financiera de las empresas y de la economía en su conjunto. Estos datos se pueden clasificar en distintas categorías.

Indicadores Técnicos

Los indicadores técnicos son variables de entrada ampliamente aplicadas en la mayoría de los estudios de predicción del mercado de valores. Estos indicadores son herramientas analíticas utilizadas para evaluar y predecir tendencias futuras en el precio de las acciones, basándose en su comportamiento histórico. Transforman datos de precios y volúmenes en indicadores de fácil lectura, lo que ayuda a identificar patrones y tendencias. Algunos de los más utilizados son:

- Medias móviles: Indican tendencias promediando los precios de un activo durante un período específico.

- RSI: Mide el impulso y la velocidad de los cambios de precios para identificar condiciones de sobrecompra o sobreventa.
- MACD (Convergencia/Divergencia de la Media Móvil): Evalúa la relación entre dos medias móviles para predecir cambios de tendencia.

Al investigar en la bibliografía actual, se constata que el uso de indicadores técnicos como patrones de entrada para modelos predictivos es bastante extendido. Esta tendencia se debe en gran medida a la disponibilidad y el fácil acceso a los datos de precios y volumen, que son la base de estos indicadores.

Sin embargo, se puede plantear un debate interesante sobre la efectividad real de proporcionar a los modelos predictivos valores de indicadores técnicos en comparación con el suministro de datos de precio y volumen sin procesar. Dado que los indicadores técnicos son transformaciones de los datos originales, podrían estar introduciendo un paso innecesario o incluso distorsionar la información fundamental. En esencia, estos indicadores son una reducción de dimensionalidad de los datos originales, lo que podría limitar la capacidad del modelo para capturar la complejidad y las sutilezas inherentes a los datos originales.

Por esta razón, para el módulo de análisis técnico se decidió utilizar datos de precios y volumen directamente, permitiendo al modelo de aprendizaje automático realizar interpretaciones a partir de la información original y completa.

Variables Macroeconómicas

Las Variables Macroeconómicas son un componente esencial en el análisis y la predicción de los mercados de valores, recibiendo una atención considerable en numerosos estudios. Estas variables ofrecen una visión amplia de la economía, lo que ayuda a comprender las fuerzas subyacentes que afectan a los mercados financieros.

De acuerdo al trabajo de investigación centrado en revisar los trabajos actuales llamado “Machine learning techniques and data for stock market forecasting: A

literature review" [26], algunas de las variables macroeconómicas más frecuentes utilizadas para realizar predicciones son:

- Rendimiento Económico: Esta categoría engloba variables clave en la economía. Incluye la Producción Industrial, que mide el nivel de producción de fábricas, minas y servicios públicos. También considera índices de inflación como el Índice de Precios al Consumidor (CPI) y el Índice de Precios al Productor (PPI), que ofrecen perspectivas sobre la variación de precios desde el punto de vista del consumidor y del productor, respectivamente. Además, abarca medidas de la actividad económica total, tales como el Producto Interno Bruto (PIB) y el Producto Nacional Bruto (PNB), junto con sus variables relacionadas, que son esenciales para comprender la salud económica general de un país
- Tasa de Interés y Oferta Monetaria: Estas variables incluyen las tasas de interés clave, como la tasa de fondos federales, y medidas de la oferta monetaria, reflejando la política monetaria de un país y su impacto en la economía.
- Tasa de Cambio: Las tasas de cambio entre monedas importantes, como el USD/EUR, influyen en el comercio internacional y en la valoración de las inversiones en divisas, afectando así los mercados de valores.
- Commodities: El precio de las materias primas, como los metales preciosos y el petróleo crudo, puede ser un indicador importante de las tendencias económicas y tiene un impacto directo en varios sectores del mercado de valores.

La integración de estas variables macroeconómicas en modelos de predicción, especialmente con el uso de técnicas modernas de aprendizaje automático y análisis de datos, permite a analistas e inversores obtener una visión más precisa y fundamentada de las posibles direcciones del mercado.

Indicadores Fundamentales

Los Indicadores utilizados en el análisis fundamental del mercado de valores se enfocan en la información financiera específica de las empresas cotizadas en bolsa y están principalmente basados en los datos financieros que la empresa reporta.

Estos incluyen datos detallados de estados financieros como ganancias, pérdidas, y balances generales. La evaluación se extiende a ratios financieros clave como el precio-beneficio y el retorno sobre capital, que ayudan a entender la rentabilidad y eficiencia operativa de la empresa. Además, se considera el desempeño de la empresa en relación con su sector, el crecimiento de sus ingresos y beneficios, y su política de dividendos para evaluar su estabilidad y potencial de crecimiento. Aspectos como la calidad de la gestión, las prácticas de gobernanza corporativa, y la responsabilidad social también son relevantes. Todos estos factores juntos ofrecen una visión integral de la salud y el potencial futuro de una empresa en el mercado de valores.

De acuerdo al trabajo “Machine learning techniques and data for stock market forecasting: A literature review” [26] los indicadores fundamentales más utilizados en los estudios de predicción del mercado de valores, son los siguientes:

- Relación de Precios (Price Ratio): 18 menciones, representando el 30.5% del total de papers analizados.
- Ganancias (Earnings): 13 menciones, representando el 22.0%.
- Valor de Mercado (Market Value): 10 menciones, representando el 16.9%.
- Dividendos (Dividend): 8 menciones, representando el 13.6%.
- Valor Contable (Book Value): 5 menciones, representando el 8.5%.
- Número de Acciones y Accionistas (Number of Stocks & Shareholders): 5 menciones, representando también el 8.5%.

Estas categorías reflejan diferentes aspectos de la información financiera de las empresas que cotizan en bolsa y su relevancia se debe a su capacidad para ofrecer una visión integral de la salud y el desempeño económico de una empresa. Por ejemplo, la relación de precios proporciona una perspectiva sobre cómo el mercado valora las acciones de la empresa en relación con sus ingresos, mientras que las ganancias reflejan la rentabilidad y eficiencia operativa. El valor de mercado indica la capitalización de mercado total de la empresa, un indicador clave del tamaño y la estabilidad de la empresa en el mercado. Los dividendos ofrecen una vista del rendimiento directo que los inversores pueden esperar de sus inversiones en

acciones, mientras que el valor contable y el número de acciones y accionistas proporcionan información fundamental sobre la estructura de capital y la distribución de la propiedad de la empresa.

Otras variables

Además de las categorías mencionadas anteriormente, existen otras variables que suelen ser relevantes en el análisis financiero:

- Precios de Otros Índices Bursátiles: Los índices bursátiles de diferentes países o sectores proporcionan una visión general de las tendencias del mercado global y sectorial. Por ejemplo, un índice como el Dow Jones o el NASDAQ puede dar una idea del comportamiento del mercado estadounidense, mientras que otros índices internacionales pueden ofrecer perspectivas sobre mercados emergentes o específicos de un sector.
- Información de Noticias Financieras: Las noticias financieras son una fuente crucial de información para el análisis del mercado. Los eventos reportados en las noticias, como cambios regulatorios, fusiones y adquisiciones, o incluso rumores de mercado, pueden tener un impacto inmediato y significativo en los precios de las acciones.
- Comunicados e Informes: Estos incluyen comunicados de prensa oficiales de las empresas, informes de ganancias y otros anuncios significativos. Estos eventos pueden provocar fluctuaciones en el mercado, ya que los inversores reaccionan a la nueva información.
- Redes Sociales: Las plataformas de redes sociales se han convertido en fuentes ricas de datos. Los tweets, en particular, pueden ser una mina de oro para el análisis del sentimiento del mercado. El análisis de sentimientos de los tweets relacionados con el mercado de valores puede proporcionar información en tiempo real sobre las percepciones y reacciones de los inversores.
- Análisis de Sentimiento del Mercado: Esta es una área creciente en el análisis financiero, donde se utilizan algoritmos avanzados para analizar textos y extraer

el sentimiento general del mercado. Esto puede incluir el análisis de noticias, publicaciones en redes sociales y foros financieros.

- Reacciones a Eventos Específicos: Los datos relacionados con eventos específicos, como desastres naturales, cambios políticos o crisis económicas, pueden ser críticos para prever cómo estos eventos afectarán a los mercados financieros.

2.7. Variable Objetivo

En el contexto de la elaboración de predicciones para el mercado financiero, es importante identificar y definir la variable de interés, ya que esta representa un elemento clave en la toma de decisiones y en el proceso de aprendizaje de los modelos de machine learning. Entre las opciones más comunes y directas para predicciones se encuentra la estimación del precio de un activo en el siguiente periodo temporal. Esta ofrece una comprensión clara y específica del movimiento esperado del precio y los resultados son muy sencillos de interpretar. Sin embargo los precios en los mercados suelen ser volátiles debido a diferentes eventos como las noticias, anuncios políticos, por lo que predecir el precio exacto puede ser una tarea difícil de concretar.

Además, una posible alternativa es utilizar los indicadores técnicos como variables objetivo para predecir su valor futuro. Estos indicadores son valiosos porque proporcionan información sobre tendencias de precios y comportamiento del activo. Por lo que predecir su valor futuro puede ser de utilidad para la toma de decisiones. Por ejemplo, podríamos predecir el valor futuro de una media móvil y utilizar esta predicción para tomar decisiones. Sin embargo, una desventaja es que los indicadores técnicos pueden ser más difíciles de interpretar y requieren un conocimiento avanzado de análisis técnico, lo que podría complicar la toma de decisiones para algunos usuarios.

Por otro lado, predecir algún tipo de retorno también es una consideración importante, dado que este aspecto constituye el objetivo final de la participación en el mercado financiero. El retorno sobre la inversión, ya sea en términos de ganancias de

capital o dividendos, es un indicador del éxito de las estrategias de inversión. Por lo tanto, la capacidad de pronosticar los retornos puede influir significativamente en las decisiones de inversión y en la gestión del riesgo. No obstante, este tipo de predicciones requieren modelos más complejos debido a la necesidad de considerar una variedad de factores de riesgo y macroeconómicos.

Un último caso que reviste especial interés para este proyecto, se centra en estimar la variación relativa, en términos porcentuales, de un activo o indicador financiero dentro de un periodo específico. La predicción de cambios porcentuales proporciona una medida intuitiva de la variación del precio del activo. También facilita la comparación entre activos de diferentes escalas o precios. Esto permite una evaluación más uniforme y justa de los movimientos de precios, independientemente del tamaño absoluto del activo. Se profundizará más sobre este tema en las siguientes secciones, destacando su importancia y cómo se aplica en este proyecto.

2.8. Herramientas Tecnológicas

En el presente trabajo se utilizó un gran número de herramientas relacionadas con la programación, análisis de datos y machine learning. A continuación se desarrollan las más importantes.

Python

Python[5] es un lenguaje de programación de alto nivel, interpretado y de propósito general, diseñado para ser legible y sencillo, lo que facilita su uso tanto para desarrolladores principiantes como experimentados. Es el lenguaje preferido para el aprendizaje automático por varias razones clave, que en conjunto contribuyen a su popularidad y adopción generalizada en el campo:

- **Simplicidad y legibilidad**: Python tiene una sintaxis clara e intuitiva, lo que facilita el desarrollo rápido y el mantenimiento del código, especialmente en proyectos complejos.

[5] “Python Software Foundation.” https://www.python.org/

- **Amplio ecosistema de bibliotecas**: Python ofrece un rico ecosistema de bibliotecas y marcos diseñados para el aprendizaje automático y el análisis de datos, como Scikit-learn, TensorFlow, PyTorch, Keras y Pandas. Estas bibliotecas proporcionan funciones y utilidades prediseñadas para operaciones matemáticas, manipulación de datos y tareas de aprendizaje automático, lo que reduce la necesidad de escribir código desde cero.
- **Comunidad activa**: La gran comunidad de Python proporciona soporte, actualizaciones regulares y abundante documentación, asegurando que las bibliotecas estén siempre optimizadas y actualizadas.
- **Productividad y eficiencia**: Con herramientas como *Jupyter Notebooks*, Python facilita el prototipado rápido y la iteración continua de modelos, acelerando el ciclo de desarrollo.

Además de las ventajas técnicas mencionadas [27] el uso del lenguaje se ve reforzada por la experiencia de uso adquirida durante el cursado de la carrera.

Pytorch

Pytorch[6] Es un framework de código abierto basado en el lenguaje de programación Python y la biblioteca Torch. Torch es una biblioteca de ML de código abierto que se utiliza para crear redes neuronales profundas. Es una de las plataformas preferidas para la investigación de aprendizaje profundo debido a que admite más de 200 operaciones matemáticas diferentes. La popularidad de PyTorch sigue aumentando, ya que simplifica la creación de modelos de redes neuronales artificiales. Es utilizado principalmente por científicos de datos para investigación y aplicaciones de inteligencia artificial [28].

[6] “Pytorch.” https://pytorch.org/

Finagg

Para la recolección de datos en este proyecto, se empleó una librería de Python conocida como finagg[7], que se especializa en facilitar la conexión con diversas APIs financieras de acceso gratuito. Finagg no solo simplifica la conexión con estas fuentes de datos, sino que también asiste en el almacenamiento y preprocesamiento de los mismos. Esta herramienta se destaca por su capacidad para integrar datos históricos provenientes de distintas APIs en bases de datos SQL, lo cual es fundamental para el análisis y el procesamiento eficiente de la información. Además, finagg ofrece funcionalidades adicionales para la transformación de estos datos agregados en formatos más adecuados para su uso en análisis avanzados e iniciativas de Inteligencia Artificial y Aprendizaje Automático.

SQLite

Para el almacenamiento de los datos obtenidos de las diferentes APIs, se utiliza el motor de base de datos SQLite[8]. Esta elección se debe a que el paquete finagg ofrece métodos que permiten agregar datos brutos o transformados en una base de datos local de SQLite, lo que resulta ideal para gestionar series de datos personalizados, como tickers o series económicas. También proporciona los metadatos necesarios para la creación automática de tablas para los distintos conjuntos de datos.

Es importante destacar que es una base de datos que no requiere la configuración de un servidor dedicado. Esto facilita su instalación haciéndola ideal para proyectos de pequeña a mediana escala y para entrenar modelos localmente.

Pandas

La biblioteca de software de código abierto Pandas[9] está diseñada específicamente para la manipulación y el análisis de datos en Python. Se puede utilizar para cargar,

[7] “finagg: Financial Aggregation for Python.” https://theogognf.github.io/finagg/build/html/index.html
[8] “SQLite.” https://www.sqlite.org/
[9] “Pandas.” https://pandas.pydata.org/

alinear, manipular o incluso fusionar datos. Es muy útil para trabajar con datos estadísticos, datos tabulares como tablas SQL o Excel, con datos de series temporales y con datos matriciales arbitrarios con etiquetas de filas y columnas [29].

Pandas se utilizó fundamentalmente en este trabajo para manipular el gran volumen de datos extraído de la base de datos, aplicando transformaciones, construyendo el conjunto de datos y realizando operaciones de limpieza y preparación.

3. Selección de Datos

Como se mencionó previamente, los datos representan un aspecto crucial, siendo probablemente el más crítico, dado que constituyen el fundamento sobre el cual los modelos desarrollan su aprendizaje. Por esta razón, se llevó a cabo una búsqueda exhaustiva de datos disponibles gratuitamente que cubran todas las categorías más relevantes. La recolección de datos se enfocó en empresas que cotizan en la bolsa de valores de Estados Unidos, obteniendo esta información a través de APIs proporcionadas por diversos organismos que suministran datos financieros de manera gratuita.

La idea inicial del proyecto incluía el uso de datos de la plataforma X[10] (antiguamente Twitter), reconociendo su potencial para reflejar eventos relevantes y el sentimiento público. Redes sociales como X son una fuente dinámica de reacciones y opiniones que podría haber aportado información valiosa sobre tendencias del mercado y percepciones de las inversiones.

Sin embargo, esta opción se descartó debido a restricciones en el acceso a la API de X, que no es de libre acceso y requiere suscripción. Esto implica limitaciones tanto en términos de costos como en la capacidad de recolección de datos.

Otras alternativas consideradas para obtener información sobre el sentimiento del mercado o noticias fueron plataformas como Bloomberg[11], que ofrecen acceso a noticias y análisis financiero. Sin embargo, estas opciones también requieren suscripción, lo cual representaba una limitación similar a la de X, y por ello no se pudieron utilizar.

Por lo tanto, el proyecto se enfocó en fuentes de datos más accesibles y directamente relevantes para el análisis financiero, como datos de precios, volúmenes de transacciones, variables económicas y reportes financieros trimestrales. Estos datos fueron obtenidos utilizando el paquete finagg. Es importante señalar que las tablas de datos utilizadas son autónomas y no tienen relaciones entre sí, ya que están

[10] "X." https://twitter.com/
[11] "Bloomberg." https://www.bloomberg.com/

diseñadas para almacenar series temporales y datos económicos que se analizan de forma individual.

A continuación se detalla en profundidad los distintos datos recopilados y la estructura de los datos.

3.1. Yahoo Finance

Es una plataforma digital que ofrece noticias financieras, datos, cotizaciones en tiempo real y herramientas para la gestión de carteras de inversión. Yahoo Finance[12] permite acceder a información sobre acciones, índices, materias primas, divisas, criptomonedas y otros activos financieros.

Esta plataforma destaca como una de las fuentes más reconocidas en la recopilación de información sobre precios y volúmenes de operaciones de diversos activos financieros. Los datos obtenidos son particularmente relevantes en el análisis técnico y se pueden ver gráficamente en las denominadas velas japonesas.

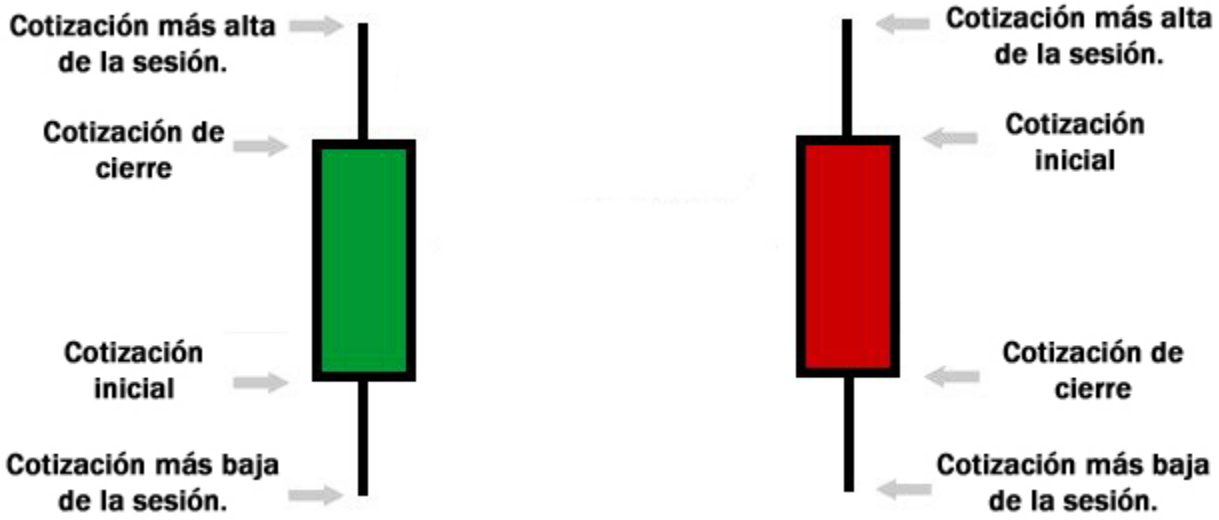

Fig. 2 Representación precios en velas japonesas. De "Patrones de Velas Japonesas en bolsa: Tipos, gráficos e interpretación" por Jose V. Gascó, 2023. https://www.rankia.com/blog/bolsa-desde-cero/3648825-patrones-velas-japonesas-bolsa-tipos-graficos-interpretacion

[12] "Yahoo Finance." https://finance.yahoo.com/

Los datos utilizados son de temporalidad diaria y fueron almacenados en la siguiente tabla:

Table: yfinance.raw.prices

Column: ticker VARCHAR

Column: date VARCHAR

Column: open FLOAT

Column: high FLOAT

Column: low FLOAT

Column: close FLOAT

Column: volume FLOAT

Detalle: Esta tabla contiene información sobre los precios históricos de las acciones en el mercado financiero para diferentes empresas.

- **ticker**: Nombre simbólico o identificador de una empresa en la bolsa de valores.
- **date**: Fecha en la que se registraron los precios de la acción.
- **open**: Precio de la acción al inicio del día de operaciones en el mercado.
- **high**: Precio más alto al que se negoció la acción durante el día.
- **low**: Precio más bajo al que se negoció la acción durante el día.
- **close**: Precio de la acción al cierre del día de operaciones en el mercado.
- **volume**: Número de acciones que se negociaron durante el día.

A partir de la tabla *yfinance.raw.prices* se generó una nueva tabla denominada *yfinance.refined.daily*. Para ello se implementaron varios pasos de transformación y refinamiento de los datos con el objetivo de facilitar el análisis y mejorar la capacidad predictiva de los modelos de Machine Learning. Estos cambios se centraron en la aplicación del cambio logarítmico a los valores de precio y en el cálculo de nuevas variables derivadas.

Los detalles de estas transformaciones son los siguientes:

- Cambio Logarítmico de Valores de Precio: Se aplicó el cambio logarítmico a las columnas open, high, low, close, y volume de la tabla original. Este cambio logarítmico transforma los valores de precio en cambios porcentuales, lo cual es más adecuado para el análisis financiero ya que normaliza las diferencias de precios y permite una mejor comparación entre diferentes acciones y períodos de tiempo.
- Cálculo de Nuevas Variables Derivadas:
    - **LOG_CHANGE(high, open):** Se calculó el cambio logarítmico entre el precio más alto del día (high) y el precio de apertura (open) de la acción. Este valor proporciona una medida porcentual de cuánto se elevó el precio de una acción desde su apertura hasta su punto más alto durante el día de operaciones.
    - **LOG_CHANGE(low, open):** De manera similar, se calculó el cambio logarítmico entre el precio más bajo del día (low) y el precio de apertura (open). Este cálculo ofrece una perspectiva de la disminución máxima del precio de la acción desde su apertura durante el día.

De esta manera la tabla resultante que refleja el comportamiento de los precios queda estructurada de la siguiente manera:

Table: yfinance.refined.daily

Column: ticker VARCHAR

Column: date VARCHAR

Column: LOG_CHANGE(open) FLOAT

Column: LOG_CHANGE(high) FLOAT

Column: LOG_CHANGE(low) FLOAT

Column: LOG_CHANGE(close) FLOAT

Column: LOG_CHANGE(volume) FLOAT

Column: LOG_CHANGE(high, open) FLOAT

Column: LOG_CHANGE(low, open) FLOAT

Detalle: Contiene datos derivados de la tabla yfinance.raw.prices

- **ticker**: El símbolo bursátil de la empresa.
- **date**: Fecha en la que se registraron los precios.
- **LOG_CHANGE(open)**: Cambio logarítmico del precio de apertura con respecto al día anterior.
- **LOG_CHANGE(high**): Cambio logarítmico del precio más alto con respecto al día anterior.
- **LOG_CHANGE(low)**: Cambio logarítmico del precio más bajo con respecto al día anterior.
- **LOG_CHANGE(close)**: Cambio logarítmico del precio de cierre con respecto al día anterior.
- **LOG_CHANGE(volume)**: Cambio logarítmico del volumen con respecto al día anterior.
- **LOG_CHANGE(high, open)**: Cambio logarítmico entre el precio más alto del día y el precio de apertura del mismo día. Es una medida de cuánto aumentó el precio durante el día desde su apertura.
- **LOG_CHANGE(low, open)**: Cambio logarítmico entre el precio más bajo del día y el precio de apertura del mismo día. Es una medida de cuánto disminuyó el precio durante el día desde su apertura.

3.2. The Federal Reserve Economic Data (FRED)

La Reserva Federal de Datos Económicos (FRED)[13] es una base de datos administrada por la Reserva Federal de St. Louis. Se caracteriza por ofrecer una extensa variedad de datos económicos y financieros tanto de Estados Unidos como de otras partes del mundo. Su propósito principal es proporcionar acceso a datos estadísticos para facilitar el análisis económico y la toma de decisiones.

[13] “Federal Reserve Economic Data.” https://fred.stlouisfed.org/

La base de datos incluye variables macroeconómicas como el Producto Interno Bruto (PIB), la inflación, las tasas de interés, el empleo, entre otros. FRED se actualiza regularmente, asegurando que los usuarios tengan acceso a la información más reciente y relevante.

Los siguiente datos fueron obtenidos de la API de la FRED utilizando el paquete finagg:

Table: fred.raw.series:

Column: series_id VARCHAR

Column: realtime_start VARCHAR

Column: realtime_end VARCHAR

Column: date VARCHAR

Column: value FLOAT

Detalle: Esta tabla cuenta con información de diferentes series de datos económicos y financieros, donde la columna **series_id** puede ser:

- **CIVPART**: Tasa de participación civil en la fuerza laboral. Representa la proporción de la población civil en edad laboral que está trabajando o buscando trabajo.
- **CPIAUCNS**: Índice de Precios al Consumidor (CPI) para todos los bienes urbanos. Es una medida de la inflación y muestra cómo cambian los precios al consumidor con el tiempo.
- **CSUSHPINSA**: Índice de precios de viviendas en Estados Unidos. Mide los cambios en el valor de mercado de las viviendas residenciales en Estados Unidos
- **DJIA**: Promedio Industrial Dow Jones. Es uno de los índices bursátiles más conocidos y representa el rendimiento de 30 grandes empresas públicas en EE.UU.
- **FEDFUNDS**: Tasa de fondos federales. Es la tasa de interés a la que las instituciones depositarias prestan dinero a otras en el mercado de fondos federales a corto plazo.

- **GDP**: Producto Interno Bruto. Representa el valor total de bienes y servicios producidos en un país.
- **GDPC1**: Producto Interno Bruto real, ajustado por inflación.
- **GS10**: Tasa de interés de los bonos del Tesoro de EE.UU. a 10 años.
- **M2**: Suministro de dinero M2. Incluye efectivo, cuentas de ahorro, cuentas corrientes, depósitos a plazo, etc.
- **MICH**: Expectativas de los consumidores sobre la inflación en el próximo año. Es una encuesta mensual realizada por la Universidad de Michigan.
- **NASDAQ100**: Índice bursátil NASDAQ 100. Representa el rendimiento de las 100 mayores empresas no financieras en el NASDAQ.
- **NASDAQCOM**: Índice compuesto NASDAQ. Incluye todas las acciones que se negocian en el NASDAQ.
- **PSAVERT**: Tasa de ahorro personal. Representa la proporción del ingreso personal que se ahorra.
- **SP500**: Índice Standard & Poor's 500. Es un índice bursátil que representa el rendimiento de 500 grandes empresas cotizadas en las bolsas de EE.UU.
- **UMCSENT**: Índice de sentimiento del consumidor de la Universidad de Michigan. Es una medida de la confianza de los consumidores en la economía.
- **UNRATE**: Tasa de desempleo. Representa la proporción de la fuerza laboral que está desempleada y buscando activamente empleo.
- **WALCL**: Tamaño del balance de la Reserva Federal. Muestra el total de activos que tiene la Reserva Federal.

A partir de los datos brutos provistos por la API, se llevaron a cabo transformaciones para refinar y adaptar los datos para un análisis económico más efectivo los cuales fueron estructurados en la tabla *fred.refined.economic*:

- Transformación de Columnas en Variables Independientes: Cada serie económica en *fred.raw.series* se representaba en filas separadas bajo la columna series_id. En *fred.refined.economic*, estas series se han transformado en columnas individuales, cada una representando una serie económica diferente. Esto facilita el análisis cruzado y la correlación entre diferentes datos financieros.

- Aplicación de Cambio Logarítmico: Varias columnas en la tabla *fred.refined.economic* han sido preprocesadas utilizando el cambio logarítmico facilitando la comparación y el análisis temporal de los datos. Se exceptuaron aquellas variables que ya estaban expresadas en términos porcentuales.
- Simplificación y Enfoque en Datos Clave: Mientras que la *tabla fred.raw.series* incluye columnas como series_id, realtime_start, realtime_end, la tabla *fred.refined.economic* se centra exclusivamente en las fechas y los valores de las series económicas, eliminando columnas adicionales para simplificar y centrarse en los datos más relevantes.

3.3. The Securities and Exchange Commission's (SEC)

La Comisión de Bolsa y Valores (SEC) es el organismo regulador del mercado de valores de Estados Unidos. Su creación respondió a la necesidad de restaurar la confianza del público en los mercados financieros después de los colapsos y el caos financiero de 1929. Fue creada en 1934 y como entidad gubernamental, la SEC se encarga de garantizar que los mercados operen de manera justa y transparente, lo cual es esencial para atraer y mantener la inversión nacional e internacional.

En su rol de organismo regulador del mercado de valores, genera y hace públicos una amplia variedad de datos esenciales para la transparencia y el buen funcionamiento de los mercados financieros. Estos datos incluyen información detallada sobre las empresas cotizadas, como informes financieros, declaraciones de registro, y divulgaciones de eventos significativos. A través de su sistema EDGAR (Electronic Data Gathering, Analysis, and Retrieval), la SEC pone a disposición del público los informes anuales (10-K), informes trimestrales (10-Q), y otros documentos financieros presentados por las compañías. Esta información es crucial para que los inversores, analistas y otros participantes del mercado puedan tomar decisiones informadas.

Los datos obtenidos a partir de este organismo se organizan en las tablas *sec.raw.submissions* y *sec.raw.tags* que presentan un resumen de los reportes trimestrales y anuales de las empresas.

Table: sec.raw.submissions

Column: cik VARCHAR

Column: ticker VARCHAR

Column: entity_type VARCHAR

Column: sic VARCHAR

Column: sic_description VARCHAR

Column: name VARCHAR

Column: exchanges VARCHAR

Column: ein VARCHAR

Column: description VARCHAR

Column: category VARCHAR

Column: fiscal_year_end VARCHAR

Detalle: Esta tabla contiene para cada empresa una descripción de la misma.

- **cik (Central Index Key)**: Identificador único asignado por la SEC a cada entidad que presenta documentos.
- **ticker**: Es el símbolo bajo el cual una acción de una empresa se negocia en una bolsa de valores. Es una forma corta y única de identificar a una empresa en una bolsa. Por ejemplo, "AAPL" es el ticker para Apple Inc.
- **entity_type**: Identifica el tipo de entidad. Por ejemplo, si es una corporación, una sociedad anónima, una entidad privada, entre otros.
- **sic (Standard Industrial Classification)**: Un sistema de clasificación utilizado en los EE.UU. para categorizar las empresas por el tipo de actividad económica en la que están involucradas.
- **sic_description**: Descripción detallada de lo que significa el código SIC específico.
- **name**: Nombre completo de la empresa o entidad.

- **exchanges**: Bolsa de valores en la que cotiza la empresa. Por ejemplo: NYSE, Nasdaq.
- **ein (Employer Identification Number)**: Número de identificación fiscal de una empresa en los EE.UU.
- **description**: Breve descripción o resumen sobre la empresa.
- **category**: Clasifica a las empresas según su capitalización de mercado y velocidad con la que presentan sus informes a la SEC.
- **fiscal_year_end**: Fecha de cierre del año fiscal de la empresa.

Table: sec.raw.tags

Column: cik VARCHAR

Column: accn VARCHAR

Column: taxonomy VARCHAR

Column: tag VARCHAR

Column: form VARCHAR

Column: units VARCHAR

Column: fy INTEGER

Column: fp VARCHAR

Column: start VARCHAR

Column: end VARCHAR

Column: filed VARCHAR

Column: frame VARCHAR

Column: label VARCHAR

Column: description VARCHAR

Column: entity VARCHAR

Column: value FLOAT

Detalle: Contiene datos relacionados con la presentación de informes financieros para cada empresa en función de los informes presentados ante la SEC.

- **cik** (Central Index Key): Identificador único asignado por la SEC a cada entidad que presenta documentos.
- **accn**: Número de acceso. Es un número único para una presentación específica dentro de una entidad.
- **taxonomy**: Clasificación o esquema que define la estructura y contenido de las etiquetas de un informe.
- **tag**: Nombre técnico o identificador de un dato en particular dentro del informe. Especifica el tipo de dato financiero y puede ser:
    - Assets: Activos totales de la empresa
    - AssetsCurrent: Activos circulantes
    - EarningsPerShareBasic: Ganancias por acción.
    - InventoryNet: Valor del inventario
    - LiabilitiesCurrent: Pasivos circulantes. Deudas y obligaciones a pagar dentro de un año.
    - NetIncomeLoss: Beneficio neto o pérdida neta después de restar todos sus gastos de sus ingresos.
    - StockholdersEquity: Valor contable de la empresa. Diferencia entre activos y pasivos.
    - CommonStockSharesOutstanding: Número de acciones en manos de inversores.
    - Liabilities: Total de deudas a corto y largo plazo.
- **form**: Indica el tipo de formulario que se presenta ante la SEC, como "10-Q" (un informe trimestral) o "10-K" (un informe anual).
- **units**: Unidades de medida en las que se presenta el valor, por ejemplo USD.
- **fy** (Fiscal Year): Año fiscal al que corresponde la información.
- **fp**: Período fiscal. Puede ser informe trimestral ("Q1", "Q2", “Q3”) o informe a fin de año ("FY").
- **start**: Fecha de inicio del período de informe.

- **end**: Fecha de finalización del período de informe.
- **filed**: Fecha en que se presentó el informe a la SEC.
- **frame**: Codificación del período fiscal y el intervalo del informe
- **label**: Es una descripción legible del tag.
- **description**: Descripción detallada o explicación del "tag".
- **entity**: Nombre de la empresa.
- **value**: Valor numérico o cuantitativo asociado con el "tag" para el período específico.

A partir de los datos anteriores se creó la tabla *sec.refined.quarterly*. Este proceso implicó la generación de nuevas variables, almacenadas en esta tabla. Entre las transformaciones clave se incluyó la modificación de la temporalidad de ciertas variables, pasando de anual a trimestral. Esto se logró mediante el método de rellenado forward fill, que consiste en utilizar el último valor disponible y por lo tanto conocido por el mercado para rellenar los datos faltantes. Adicionalmente, se aplicó un cambio logarítmico a las siguiente variables para que todos los valores estén en términos porcentuales:

- AssetsCurrent
- CommonStockSharesOutstanding
- InventoryNet
- Liabilities
- LiabilitiesCurrent
- StockholdersEquity

También se calcularon ratios y métricas financieras adicionales, enriqueciendo así el conjunto de datos con información más detallada y relevante para análisis de periodos trimestrales:

- **AssetCoverageRatio**:
    - **Definición**: Esta relación mide la capacidad de una empresa para cubrir sus deudas utilizando solo sus activos. Es una indicación de la solvencia de la empresa.
    - **Fórmula**: Assets / Liabilities

    - **Interpretación**: Un valor superior a 1 indica que la empresa tiene más activos que deudas, lo que podría considerarse una posición financiera más fuerte. Por otro lado, un valor inferior a 1 sugiere que la empresa tiene más deudas que activos.
- **BookRatio (Price-to-Book Ratio)**:
    - **Definición**: Es una medida que compara el valor de mercado de la empresa con su valor contable.
    - **Fórmula**: StockholdersEquity / CommonStockSharesOutstanding
    - **Interpretación**: Un ratio inferior a 1 podría indicar que la acción está infravalorada, mientras que un valor superior a 1 podría indicar que está sobrevalorada. Sin embargo, esta interpretación puede variar dependiendo del sector y otros factores.
- **DebtEquityRatio**:
    - **Definición**: Esta relación mide la proporción de financiamiento por deuda en comparación con el capital propio o equity.
    - **Fórmula**: Liabilities / StockholdersEquity
    - **Interpretación**: Un ratio alto indica que la empresa está financiada en gran medida por deuda. Aunque la deuda puede impulsar el rendimiento, también puede aumentar el riesgo.
- **EarningsPerShareBasic**:
    - **Definición**: Representa la porción de las ganancias de una empresa asignada a cada acción en circulación.
    - **Fórmula**: NetIncomeLoss / CommonStockSharesOutstanding
    - **Interpretación**: Indica la rentabilidad de la empresa en función del número de acciones. Cuanto más alto sea, más rentable es la empresa para sus accionistas.
- **QuickRatio**:
    - **Definición**: Es una medida de la capacidad de una empresa para cubrir sus pasivos circulantes sin depender de sus inventarios.
    - **Fórmula**: (AssetsCurrent - InventoryNet) / LiabilitiesCurrent
    - **Interpretación**: Una relación superior a 1 indica que la empresa puede

cubrir sus obligaciones a corto plazo sin vender su inventario. Es una medida de liquidez más estricta que el WorkingCapitalRatio.

- **ReturnOnAssets (ROA):**
  - **Definición**: Indica cuán eficientemente una empresa utiliza sus activos para generar beneficios.
  - **Formula**: NetIncomeLoss / Assets
  - **Interpretación**: Un ROA más alto indica una mayor eficiencia en la utilización de los activos.
- **ReturnOnEquity (ROE):**
  - **Definición**: Muestra cuánto beneficio genera una empresa con el dinero invertido por los accionistas.
  - **Fórmula**: NetIncomeLoss / StockholdersEquity
  - **Interpretación**: Un ROE más alto indica una mayor rentabilidad para los accionistas.
- **WorkingCapitalRatio (Ratio Corriente):**
  - **Definición**: Mide la capacidad de una empresa para cubrir sus pasivos circulantes con sus activos circulantes.
  - **Fórmula**: AssetsCurrent / LiabilitiesCurrent
  - **Interpretación**: Una relación superior a 1 sugiere que la empresa tiene suficientes activos circulantes para cubrir sus deudas a corto plazo. Un valor inferior a 1 puede indicar problemas de liquidez.

De manera que la tabla resultante tiene la siguiente estructura:

Table: sec.refined.quarterly

Column: cik VARCHAR

Column: filed VARCHAR

Column: fy INTEGER

Column: fp VARCHAR

Column: LOG_CHANGE(Assets) FLOAT

Column: LOG_CHANGE(AssetsCurrent) FLOAT

Column: LOG_CHANGE(CommonStockSharesOutstanding) FLOAT

Column: LOG_CHANGE(InventoryNet) FLOAT

Column: LOG_CHANGE(Liabilities) FLOAT

Column: LOG_CHANGE(LiabilitiesCurrent) FLOAT

Column: LOG_CHANGE(StockholdersEquity) FLOAT

Column: AssetCoverageRatio FLOAT

Column: BookRatio FLOAT

Column: DebtEquityRatio FLOAT

Column: EarningsPerShareBasic FLOAT

Column: QuickRatio FLOAT

Column: ReturnOnAssets FLOAT

Column: ReturnOnEquity FLOAT

Column: WorkingCapitalRatio FLOAT

3.4. The Bureau of Economic Analysis (BEA)

La Oficina de Análisis Económico (BEA) es una agencia del Departamento de Comercio de los Estados Unidos encargada de producir importantes estadísticas económicas que influyen en las decisiones de política monetaria y de inversión tanto de empresas privadas como del gobierno. La información proporcionada por la BEA abarca diversos aspectos de la economía, como el Producto Interno Bruto (PIB), el ingreso y gasto personal, el balance de pagos y otros indicadores que reflejan la salud y dirección de la economía estadounidense.

A partir de la API de la BEA se obtuvieron y estudiaron las siguientes variables macroeconómicas:

Table: input_output

Column: table_id INTEGER

Column: year INTEGER

Column: row_code VARCHAR

Column: row_description VARCHAR

Column: row_type VARCHAR

Column: col_code VARCHAR

Column: col_description VARCHAR

Column: col_type VARCHAR

Column: value FLOAT

Detalle: Los datos de esta tabla son parte de un modelo Entrada-Salida que describe las interacciones económicas entre diferentes sectores o industrias. Estos modelos son comúnmente usados en economía para analizar cómo las salidas de un sector se convierten en entradas para otro sector.

- **table_id**: Identificador de la tabla o variable económica
- **year**: Año de los datos.
- **row_code**: Código que identifica a la industria.
- **row_description**: Nombre textual de row_code. Ejemplo: farms, retail trade.
- **row_type**: Especifica si pertenece a Industry o Commodity.
- **col_code**: Sector o industria que interactúa con el sector o industria representado en row_code.
- **col_description**: descripción textual de col_code
- **col_type**: Especifica si pertenece a Industry o Commodity.
- **value**: Magnitud de la transacción expresada en millones de dólares.

Los datos presentados poseen un nivel de agregación extremadamente bajo, lo que conlleva a un volumen considerable y a una complejidad significativa tanto en su interpretación como en su procesamiento. Debido a estas características, no se procesaron y, por ende, no se utilizaron.

Table: fixed_assets

Column: table_id VARCHAR

Column: series_code VARCHAR

Column: line INTEGER

Column: line_description VARCHAR

Column: year INTEGER

Column: metric VARCHAR

Column: units VARCHAR

Column: e INTEGER

Column: value FLOAT

Detalle: Contiene información de 108 tablas determinadas por **table_id**. Estas tablas ofrecen un análisis exhaustivo de activos fijos y bienes duraderos en distintas esferas: privada, gubernamental y del consumidor. Cubren aspectos como el valor neto, la depreciación, la inversión y la edad promedio de estos activos. Las métricas se desglosan según el costo actualizado, el costo histórico y los índices de cantidad ajustados por inflación. Además, las tablas se organizan por tipo de activo, industria y forma legal de la organización. Se encuentran en temporalidad anual y contienen datos desde 1925 a 2022.

Las tablas se pueden enmarcar en 9 secciones

- Section 1 - fixed assets and consumer durable goods
- Section 2 - private fixed assets by type
- Section 3 - private fixed assets by industry
- Section 4 - nonresidential fixed assets
- Section 5 - residential fixed assets
- Section 6 - private fixed assets
- Section 7 - government fixed assets
- Section 8 - consumer durable goods
- Section 9 - chained dollar tables

Cada sección cuenta con muchas tablas (*table_id*) y dentro de estas tablas se encuentran distintas columnas (*series_code*) con una serie de valores en un rango de

tiempo. En algunos casos para un mismo *table_id* y *series_code*, se tienen 2 rangos de valores que se identifican por *line.*

Es por ello que se crearon tablas separadas para cada combinación única de *table_id*, *series_code*, y *line.* Cada tabla se nombró de acuerdo a la convención fix_as_[table_id]_[series_code]_[line]. La tabla resultante:

Table: fix_as_[table_id][series_code][line]

Column: year INTEGER

Column: line_description VARCHAR

Column: metric VARCHAR

Column: e FLOAT

Column: value FLOAT

Column: log_change FLOAT

Column: norm FLOAT

Column: norm_log_change FLOAT

Donde metric es la unidad de medida y e es el exponente de value.

Mediante esta desagregación de la tabla **fixed_assets,** se obtuvieron un total de 6261 tablas específicas y procesadas.

Table: gdp_by_industry

Column: table_id INTEGER

Column: freq VARCHAR

Column: year INTEGER

Column: quarter INTEGER

Column: industry VARCHAR

Column: industry_des

Detalle: Estos datos representan un conjunto detallado de estadísticas económicas relacionadas con la producción, el valor agregado y los insumos por industria. Ofrece una visión agregada y de alto nivel de la actividad económica por industria, a diferencia de la tabla *input_output* que proporciona una visión más granular y específica.

Algunas de las tablas contienen valores anuales y otras trimestrales. En este caso solo se conservan las trimestrales debido a que la cantidad de datos en temporalidad anual se reduce demasiado como para entrenar un modelo. El valor agregado de las distintas tablas está expresado en miles de millones de dólares.

Las tablas (*table_id*) provistas por la BEA son las siguientes:

- Value Added by Industry (1)
- Value added by Industry as a Percentage of Gross Domestic Product (5)
- Chain-Type Quantity Indexes for Value Added by Industry (8)
- Percent Changes in Chain-Type Quantity Indexes for Value Added by Industry (9)
- Real Value Added by Industry (10)
- Chain-Type Price Indexes for Value Added by Industry (11)
- Percent Changes in Chain-Type Price Indexes for Value Added by Industry (12)
- Contributions to Percent Change in Real Gross Domestic Product by Industry (13)
- Contributions to Percent Change in the Chain-Type Price Index for Gross Domestic Product by Industry (14)
- Gross Output by Industry (15)
- Chain-Type Quantity Indexes for Gross Output by Industry (16)
- Percent Changes in Chain-Type Quantity Indexes for Gross Output by Industry (17)
- Chain-Type Price Indexes for Gross Output by Industry (18)
- Percent Changes in Chain-Type Price Indexes for Gross Output by Industry (19)

- Intermediate Inputs by Industry (20)
- Chain-Type Quantity Indexes for Intermediate Inputs by Industry (21)
- Percent Changes in Chain-Type Quantity Indexes for Intermediate Inputs by Industry (22)
- Chain-Type Price Indexes for Intermediate Inputs by Industry (23)
- Percent Changes in Chain-Type Price Indexes for Intermediate Inputs by Industry (24)
- Real Gross Output by Industry (208)
- Real Intermediate Inputs by Industry (209)

Cada una de estas tablas cuenta con valores trimestrales desde 2005 a 2022 para más de 100 industrias (*industry*). Las distintas categorías de industrias son las siguientes:

- Agriculture, forestry, fishing, and hunting (11)
- Farms (111CA)
- Forestry, fishing, and related activities (113FF)
- Mining (21)
- Oil and gas extraction (211)
- Mining, except oil and gas (212)
- Support activities for mining (213)
- Utilities (22)
- Construction (23)
- Food and beverage and tobacco products (311FT)
- Textile mills and textile product mills (313TT)
- Apparel and leather and allied products (315AL)
- Manufacturing (31G)
- Nondurable goods (31ND)
- Wood products (321)
- Paper products (322)
- Printing and related support activities (323)
- Petroleum and coal products (324)
- Chemical products (325)

- Plastics and rubber products (326)
- Nonmetallic mineral products (327)
- Primary metals (331)
- Fabricated metal products (332)
- Machinery (333)
- Computer and electronic products (334)
- Electrical equipment, appliances, and components (335)
- Motor vehicles, bodies and trailers, and parts (3361MV)
- Other transportation equipment (3364OT)
- Furniture and related products (337)
- Miscellaneous manufacturing (339)
- Durable goods (33DG)
- Wholesale trade (42)
- Motor vehicle and parts dealers (441)
- Food and beverage stores (445)
- Retail trade (44RT)
- General merchandise stores (452)
- Air transportation (481)
- Rail transportation (482)
- Water transportation (483)
- Truck transportation (484)
- Transit and ground passenger transportation (485)
- Pipeline transportation (486)
- Other transportation and support activities (487OS)
- Transportation and warehousing (48TW)
- Warehousing and storage (493)
- Other retail (4A0)
- Information (51)
- Publishing industries, except internet (includes software) (511)
- Motion picture and sound recording industries (512)
- Broadcasting and telecommunications (513)

- Data processing, internet publishing, and other information services (514)
- Finance and insurance (52)
- Federal Reserve banks, credit intermediation, and related activities (521CI)
- Securities, commodity contracts, and investments (523)
- Insurance carriers and related activities (524)
- Funds, trusts, and other financial vehicles (525)
- Real estate and rental and leasing (53)
- Real estate (531)
- Rental and leasing services and lessors of intangible assets (532RL)
- Professional, scientific, and technical services (54)
- Legal services (5411)
- Miscellaneous professional, scientific, and technical services (5412OP)
- Computer systems design and related services (5415)
- Management of companies and enterprises (55)
- Administrative and waste management services (56)
- Administrative and support services (561)
- Waste management and remediation services (562)
- Educational services, health care, and social assistance (6)
- Educational services (61)
- Health care and social assistance (62)
- Ambulatory health care services (621)
- Hospitals (622)
- Hospitals and nursing and residential care facilities (622HO)
- Nursing and residential care facilities (623)
- Social assistance (624)
- Arts, entertainment, recreation, accommodation, and food services (7)
- Arts, entertainment, and recreation (71)
- Performing arts, spectator sports, museums, and related activities (711AS)
- Amusements, gambling, and recreation industries (713)
- Accommodation and food services (72)
- Accommodation (721)

- Food services and drinking places (722)
- Other services, except government (81)
- Finance, insurance, real estate, rental, and leasing (FIRE)
- Government (G)
- Gross domestic product (GDP)
- Federal (GF)
- Government enterprises (GFE)
- General government (GFG)
- National defense (GFGD)
- Nondefense (GFGN)
- State and local (GSL)
- Government enterprises (GSLE)
- General government (GSLG)
- Housing (HS)
- Information-communications-technology-producing industries (ICT)
- Information-communications-technology-producing industries (ICT)
- All industries (II)
- Not allocated by industry (NABI)
- Other real estate (ORE)
- Private goods-producing industries (PGOOD)
- Professional and business services (PROF)
- Private services-producing industries (PSERV)

Los datos asociados a las industrias GDP, II y NABI fueron descartadas debido a que por inspección se pudo observar que hay datos faltantes.

Esta estructura un poco compleja fue desagregada en un total de 2058 tablas. De manera que se tiene una tabla por cada dato económico e industria específica. Por ejemplo, la tabla: *gdp_15_21* contiene la *producción bruta (tabla 15)* para la industria de *minería (key 21).* Como resultados se obtienen tablas de este tipo:

Table: gdp_[table_id]_[industry]

Column: year INTEGER

Column: quarter INTEGER

Column: value FLOAT

Column: norm FLOAT

4. Dataset

En este estudio, se adoptaron dos metodologías distintas para la estructuración y entrenamiento de la red neuronal. La primera metodología implicó la conversión de todos los datos a una escala de tiempo diaria, seguida de su integración simultánea en la red, sin realizar distinciones basadas en el tipo de datos. La segunda metodología se enfocó en procesar distintamente los datos según su categoría, utilizando subredes para introducirlos de manera independiente al modelo.

Este capítulo explica el procesamiento de datos aplicado en la primera metodología, mientras que en el capítulo Aprendizaje se describen las adaptaciones que se realizaron a los datos para la segunda metodología.

4.1. Temporalidad de los datos

La temporalidad de los datos influye significativamente en la selección del conjunto de datos para el entrenamiento de modelos analíticos. Si se dispone de datos con distintas temporalidades, por ejemplo diarios y anuales, la utilización conjunta presenta desafíos, dado que la granularidad de los datos no es compatible directamente.

En lo que respecta al rango de fechas, este está determinado por el conjunto de datos con el rango más restrictivo. Es decir, si un conjunto de datos abarca desde 2014 hasta 2023 y otro de 2016 hasta 2023, el rango efectivo para el análisis combinado sería de 2016 a 2023, dado que solo en este período se disponen de datos de ambas series. Esta limitación en el rango de fechas implica que cualquier análisis o modelo de entrenamiento estará confinado a la ventana temporal donde se intersectan todos los conjuntos de datos disponibles.

En la siguiente tabla se puede visualizar una breve descripción de los datos seleccionados y procesados.

Nombre	Temporalidad	Tipo Dato	Fecha Inicio	Fecha Fin	Detalle
fred.refined.economic	Diaria	Económico	2014-10-02	2023-08-28	
sec.refined.quarterly	Trimestral	Empresa	2009	2023	Rango datos depende de la empresa
yfinance.refined.daily	Diaria	Empresa	-	2023-08-29	fecha inicio depende empresa
fix_as_[table_id]_[series_code]_[line]	Anual	Económico	1901	2021	rango depende del dato
gdp_[table_id]_[industry]	Trimestral	Económico	2005-Q1	2023-Q1	

Tabla I: Descripción del dataset

El conjunto de datos *fred.refined.economic* ofrece una visión diaria del comportamiento económico desde octubre de 2014 hasta agosto de 2023, pero se debe mencionar que algunos datos de la tabla fueron desagregados a partir de datos trimestrales.

En cuanto a la serie *sec.refined.quarterly*, presenta datos trimestrales desde 2009 hasta 2023, cuya variabilidad en el rango de datos depende de cada empresa específica, lo que puede ser un limitante debido a la cantidad de datos disponibles.

La tabla *yfinance.refined.daily* contiene datos diarios hasta agosto de 2023. El

rango de datos es amplio y su fecha de inicio varía según la empresa.

Las tablas *fix_as_[table_id][series_code][line]* cuenta con datos dentro de un amplio rango, pero en temporalidad anual. Esto hace que sea difícil utilizarlos en conjunto con los demás datos trimestrales y diarios por lo que fueron descartados.

Las tablas *gdp_[table_id]_[industry]* que se derivan a partir de la tabla *gdp_by_industry* de la BEA, son de temporalidad trimestral y dentro de un rango similar a los datos de la sec y de fred.

La alineación de estas series temporales, junto con la interpolación necesaria para datos faltantes y la normalización para permitir comparaciones homogéneas, son pasos fundamentales en la preparación del conjunto de datos.

4.2. Datos de entrada

Para optimizar el entrenamiento de los modelos, se emplean conjuntos de datos con frecuencias diarias y trimestrales, es por ello que se descarta la tabla fix_as_[table_id]_[series_code]_[line] del conjunto de datos.

De esta manera que las tablas utilizadas son las siguientes:

- fred.refined.economic
- yfinance.refined.daily
- sec.refined.quarterly
- gdp_[table_id]_[industry]

Las tablas con datos trimestrales son convertidas a una frecuencia diaria mediante una técnica de interpolación conocida como "forward fill", que consiste en llenar los intervalos diarios con el último valor trimestral registrado. Esta metodología asegura la continuidad de los datos y permite una mayor coherencia temporal para el análisis diario, facilitando así el proceso de aprendizaje automático al proporcionar una secuencia completa y detallada de información a lo largo del tiempo. También, todas las variables económicas que no están representadas de forma porcentual, fueron expresadas mediante el cambio logarítmico.

En el caso de las distintas tablas del tipo gdp_[table_id]_[industry], los datos son trimestrales, pero no se dispone de la fecha exacta de publicación, es por ello que fueron desplazados un trimestre al momento de ingresar a la red, para evitar utilizar información futura o desconocida. Este desplazamiento se estimó a partir de las fechas de las últimas publicaciones.

Además fue necesario reducir el número de tablas gdp_[table_id]_[industry] debido a las limitaciones en capacidad de cómputo y para evitar una sobrerrepresentación de características de una naturaleza específica, por lo que se seleccionaron aquellas que se consideran más relevantes:

- Real Value Added by Industry (10): Presenta el valor agregado por industria ajustado por inflación. Entender el valor agregado real de una industria puede ofrecer pistas sobre la salud económica y la productividad de esa industria. Si una industria está agregando valor significativo al PIB, esto podría señalar un rendimiento financiero sólido que potencialmente podría traducirse en precios de acciones más altos para las empresas dentro de esa industria. La red neuronal podría identificar patrones históricos donde el aumento del valor agregado en una industria específica se correlaciona con movimientos en los precios de las acciones.
- Real Gross Output by Industry (208): El producto bruto real de una industria refleja el volumen de producción y ventas, proporcionando una medida de su rendimiento económico. Un aumento en la producción bruta puede reflejar una demanda más fuerte y eficiencias operativas, lo cual puede ser un indicador positivo para los inversores, potencialmente elevando los precios de las acciones.
- Chain-Type Price Indexes for Value Added by Industry (11): Esta tabla puede proporcionar información adicional sobre las tendencias de precios dentro de la industria específica. Entender cómo los precios dentro de la industria están cambiando puede ser útil para predecir los márgenes de beneficio y, por ende, el rendimiento de las acciones.
- Real Intermediate Inputs by Industry (209): Esta tabla ofrece información sobre los insumos intermedios ajustados por inflación, que son los bienes y servicios

utilizados en la producción. Los cambios en los costos de estos insumos pueden afectar directamente los márgenes de beneficio de las empresas, y por lo tanto, sus precios de acciones.

4.3. Variable Objetivo

En el diseño de modelos predictivos financieros, la elección de una variable de respuesta adecuada es crucial. Los datos de entrada se deben complementar con una variable de respuesta que se alinee con los objetivos analíticos, en este caso, la variación futura de precio en activos financieros. Si bien el precio en términos nominales puede parecer la elección más directa, enfrenta limitaciones significativas en análisis estadísticos debido a su potencial falta de estacionariedad. Es decir, la media y la varianza no son constantes, algo que muchos modelos presuponen.

Además, los precios nominales no reflejan adecuadamente la dinámica de los retornos porcentuales, que son de mayor relevancia para los inversores y analistas, pues estos últimos buscan comprender y capitalizar las variaciones relativas de los precios, no los cambios absolutos.

En este contexto, el cambio logarítmico del precio emerge como una alternativa para el target de modelos predictivos. Este enfoque se centra en la tasa de cambio relativa de los precios, y es una práctica estándar para analizar y modelar la rentabilidad de los activos.

Además, el uso del cambio logarítmico en el análisis de series temporales financieras ofrece la ventaja de producir datos que se aproximan a una distribución normal [30]. Esta característica es particularmente ventajosa ya que numerosos modelos estadísticos y algoritmos de aprendizaje automático parten del supuesto de normalidad en la distribución de los errores. Cumplir con este supuesto potencia la efectividad y la convergencia de los modelos en la predicción de datos futuros. En la siguiente gráfica se puede observar que la distribución de probabilidad para el cambio

logarítmico de precio de la empresa Apple[14] sigue una distribución normal, evidenciando la validez de este enfoque.

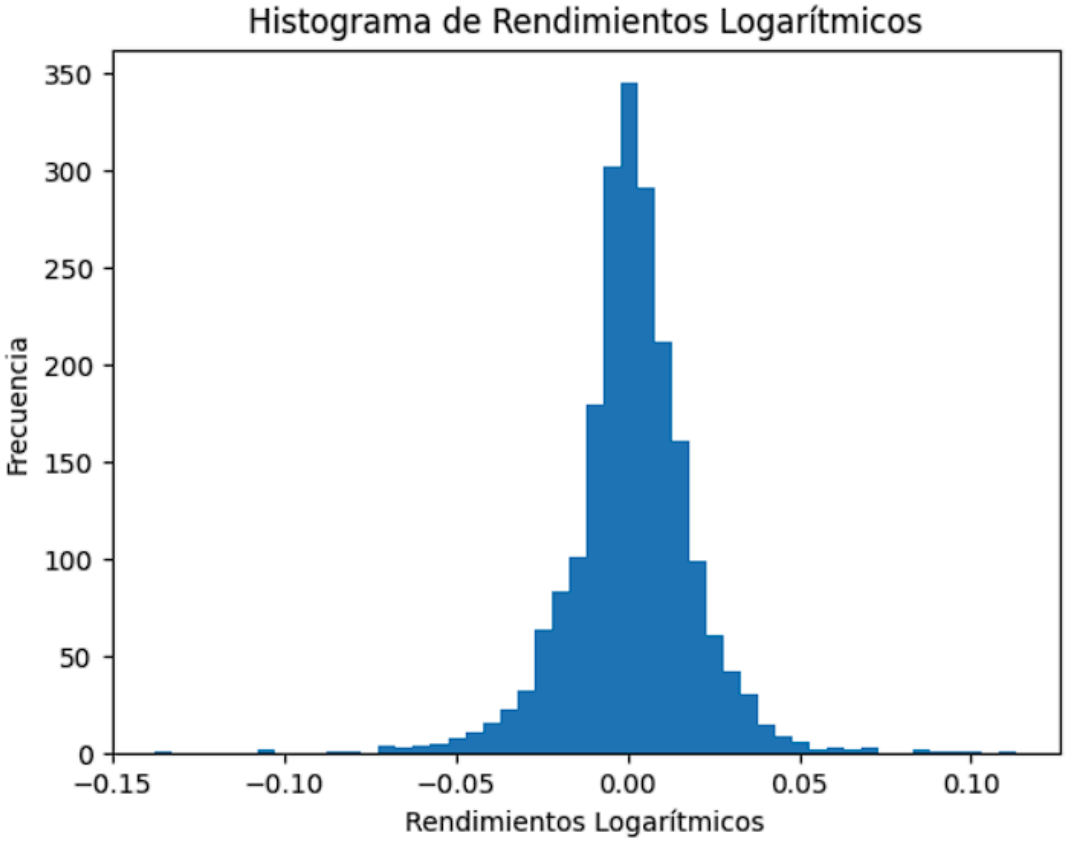

Fig. 3 Histograma de cambio logarítmico del precio (Apple)

Adicionalmente, se llevó a cabo una simulación de operaciones de trading utilizando los cambios logarítmicos en el precio del día siguiente para la toma de decisiones de compra o venta. En este caso, si el cambio logarítmico es mayor a 0.5 se realiza una compra y si es menor a -0.5 una venta.

Esta simulación permite visualizar el rendimiento que se obtendría en un escenario hipotético donde las predicciones sean completamente acertadas. Se observa cómo el modelo ejecutaría las compras y ventas en los momentos óptimos de incremento y disminución de precios. En el caso de aplicar esta metodología con las acciones de APPLE, la simulación indicó un retorno potencial del 584.57% en un periodo de 299

[14] "Apple." https://www.apple.com/

días. Por supuesto este nivel de acierto es una utopía, pero nos permite ver que entre distintos enfoques a la hora de definir la variable objetivo, este probablemente es el que maximiza las ganancias.

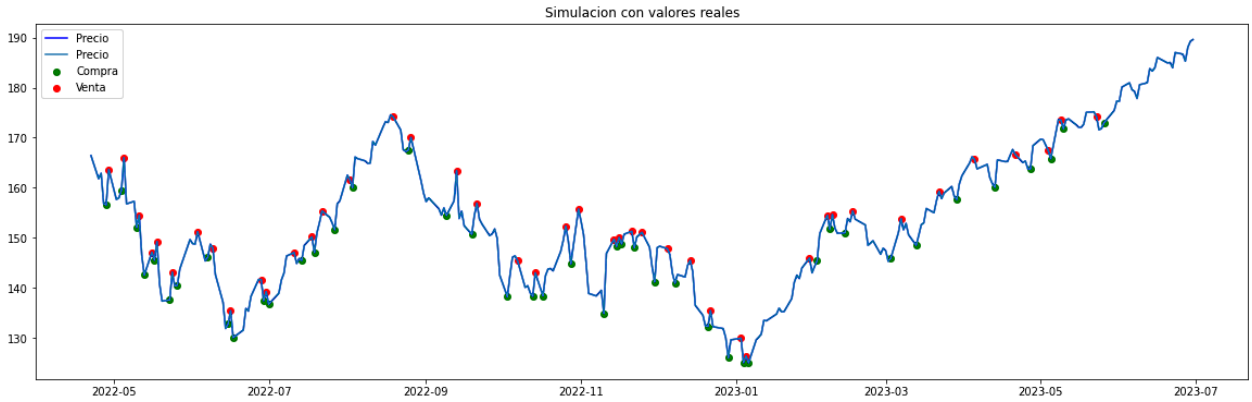

Fig. 4 Simulacion de trading

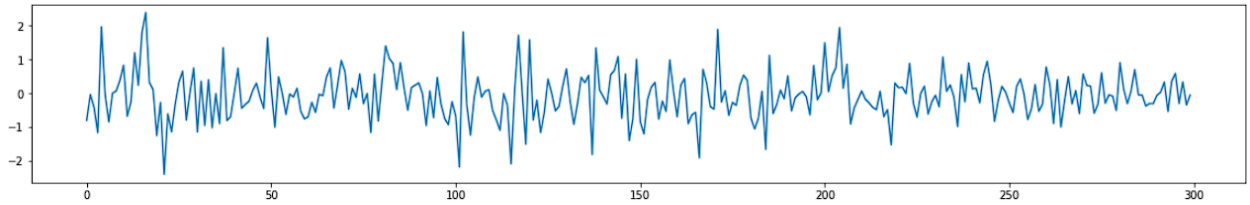

Fig. 5 Cambio logarítmico de precio

4.4. Armado de Dataset

Tras la determinación de los patrones de entrada y salida para la red neuronal, el siguiente paso es establecer la forma en que estos datos alimentarán al modelo en el proceso de entrenamiento. En este contexto, se diseñó y codificó una función que opera como una ventana deslizante que determina el rango de datos que contiene cada patrón de entrada, el corrimiento de la ventana y el rango de valores a predecir.

Esta función recibe las variables de entrada y los objetivos a predecir y se pueden especificar los siguientes parámetros:

- Definición del Rango Temporal de la Entrada: Esta función permite establecer el período temporal que abarca cada patrón de entrada. Por ejemplo, al trabajar

con 20 variables económicas, se puede fijar una ventana de 50 días. Esto significa que cada vector de entrada en la red neuronal tendrá una dimensión de 20 variables por 50 días, resultando en un conjunto de datos multidimensional que refleja las tendencias y patrones durante ese período específico.

- Horizonte de predicción: La función también posibilita la definición de la dimensión de la variable objetivo, esencial para determinar el horizonte temporal de las predicciones. Por ejemplo, se puede establecer que cada predicción contenga datos de los siguientes n periodos de tiempo.
- Desplazamiento de la Ventana Temporal: Otro aspecto clave es la capacidad de definir el desplazamiento de la ventana temporal entre cada muestra de datos. Este desplazamiento, medido en unidades de tiempo, determina cómo se actualiza la ventana para incluir datos nuevos y descartar los más antiguos.

De esta manera, al utilizar datos de varios días para realizar las predicciones, se alinea el dataset con la capacidad de las redes LSTM para capturar y aprender de las dependencias temporales y las secuencias de datos. Esto es especialmente valioso en el análisis financiero, donde las tendencias y patrones a lo largo del tiempo son fundamentales.

4.5. Escalado de Datos

Antes de comenzar con el entrenamiento es importante escalar los datos para que todas las características tengan la misma importancia. Esto es necesario ya que las redes neuronales son algoritmos que pueden ser muy sensibles a la escala de los datos. Existen múltiples opciones de escalado [31]. En este caso, debido a que muchos datos tienden a tener una distribución normal, una buena opción es utilizar un escalado estándar o también llamado z-score [32]. Con esta técnica, cada valor en un conjunto de datos se normaliza de tal manera que la media de todos los valores es 0 y la desviación estándar es 1.

La fórmula es la siguiente:

$$V(n) = \frac{x(n)-\mu}{\sigma}$$

Donde:

- V(n) es el valor normalizado en el tiempo n.
- x(n) es el valor original en el tiempo.
- μ es el promedio.
- σ es la desviación estándar.

4.6. Validacion

Para el entrenamiento de la red se utiliza una metodología denominada Walk Forward Validation [33]. Es una herramienta de validación cruzada diseñada para evaluar modelos predictivos, especialmente útil en el contexto de series temporales. Esta técnica divide un conjunto de datos en múltiples muestras de entrenamiento y prueba a lo largo del tiempo de una manera que simula cómo se utilizaría el modelo en la vida real para hacer predicciones hacia el futuro.

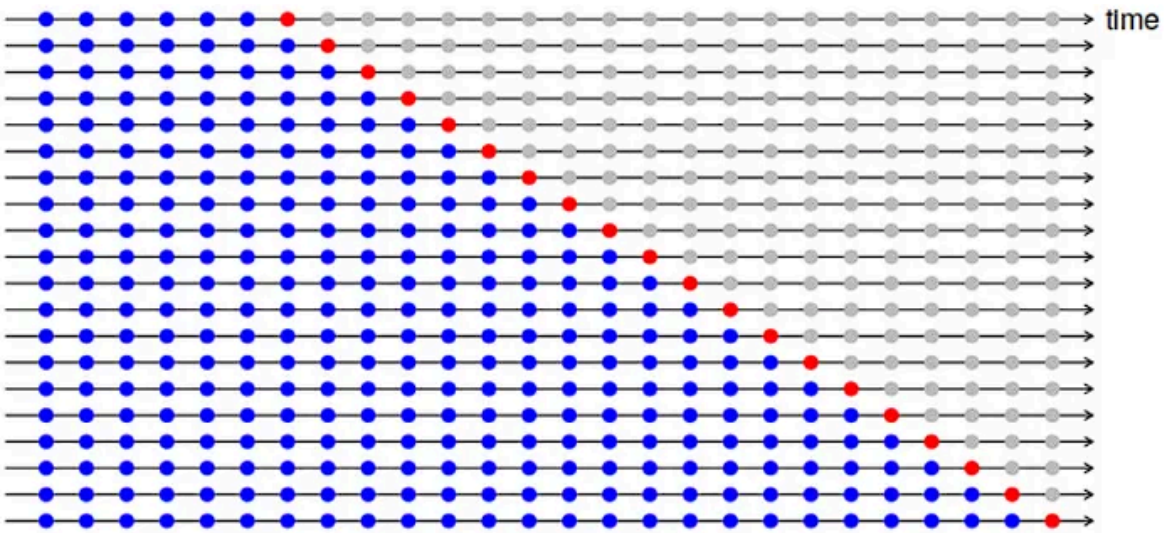

Fig. 6 Representación Walk Forward Validation. De "Time Series Cross-Validation: A Walk Forward Approach in Python" por germayne, 2019.
https://medium.com/eatpredlove/time-series-cross-validation-a-walk-forward-approach-in-python-8534dd1db51a

En Walk Forward Validation, el conjunto de datos se divide en dos partes: una ventana de entrenamiento que se expande con cada iteración y una ventana de prueba que sigue a la ventana de entrenamiento. Al principio, se entrena un modelo en un segmento inicial de datos y luego se prueba en el siguiente segmento, llamado horizonte de predicción. Después de esta primera evaluación, la ventana de entrenamiento se expande para incluir los datos hasta el horizonte y el modelo se reentrena antes de hacer la siguiente prueba en el nuevo horizonte. Este proceso se repite, avanzando a través del conjunto de datos y permitiendo que el modelo se adapte a medida que se dispone de más información.

El propósito de este enfoque es asegurarse de que el modelo pueda predecir con precisión los datos futuros basándose en la información aprendida de los datos pasados.

Debido a que algunos datos originalmente son de naturaleza trimestral, se eligió un tamaño de ventana mayor a 90, de manera que pueda capturar los cambios trimestrales para cada predicción. Además, en cada iteración del entrenamiento se utilizó un horizonte de predicción de un día, y un corrimiento de ventana de un lugar. Esto permite utilizar el máximo de información posible para cada predicción y además facilita la comparación de las predicciones contras los valores objetivo o conocidos.

Un aspecto importante a destacar sobre el método de validación Walk Forward es que, si bien sobresale por simular de manera realista el uso del modelo y por maximizar el aprovechamiento de los datos, también presenta ciertas desventajas. Entre estas, se encuentra la dificultad en implementar técnicas de validación convencionales para evaluar problemas como el sobreajuste (overfitting) y en ajustar parámetros de forma óptima, como la modificación de la taza de aprendizaje. Estas limitaciones se deben principalmente a la necesidad de respetar el orden cronológico de los datos, inherente a las series temporales. Además, reservar una cantidad significativa de datos para conjuntos de prueba y validación puede implicar realizar predicciones para puntos demasiado distantes en el tiempo, lo cual muy posiblemente conlleve una disminución en la precisión de dichas predicciones.

Otra desventaja es que esta metodología implica reentrenar el modelo de manera continua cada vez que se incorpora un nuevo dato al conjunto de entrenamiento. A diferencia de otros métodos de validación que requieren un único entrenamiento, la naturaleza iterativa y progresiva de Walk Forward significa que con cada nuevo dato añadido, el modelo debe ser ajustado y entrenado de nuevo. Esto no solo incrementa el tiempo de procesamiento, sino que también eleva el consumo de recursos computacionales.

4.7. Métricas

Error Cuadrático Medio

Como primera métrica para evaluar el desempeño del modelo se utilizó el Error Cuadrático Medio (MSE). Esta es una métrica tradicional y ampliamente utilizada para evaluar modelos en tareas de predicción. Su popularidad se debe a su capacidad para cuantificar de manera efectiva el rendimiento de un modelo al medir el promedio de los cuadrados de los errores, es decir, la diferencia cuadrática entre los valores predichos por el modelo y los valores reales observados. El MSE es particularmente valioso porque penaliza errores más grandes de manera más severa que los pequeños, lo que resulta en una métrica sensible a variaciones significativas en el rendimiento del modelo.

En este estudio, dada la metodología de validación Walk Forward empleada, se obtuvo el error promedio de entrenamiento y test para cada iteración. Es decir, cada vez que se ampliaba el dataset de entrenamiento con un nuevo valor.

Simulacion de Trading

El segundo método implementado fue una simulación de estrategia de trading, utilizando las predicciones generadas por el modelo para realizar operaciones de compra y venta. Se definieron umbrales de acción basados en la desviación estándar de las predicciones, ajustada por un factor específico. De manera que las predicciones que superaban el umbral establecido indicaban un momento para

comprar, mientras que aquellas por debajo del umbral mínimo sugerían una venta. La eficacia de esta estrategia se evaluó en términos del porcentaje de retorno de la inversión, proporcionando una medida tangible del rendimiento del modelo en un entorno de mercado real y dinámico.

Este enfoque práctico es crucial, ya que en el mundo real del trading, la rentabilidad final es el indicador más relevante del éxito. Aunque un bajo MSE indica una buena correspondencia entre las predicciones y los valores reales, no necesariamente garantiza resultados financieros positivos. La simulación de trading ayuda a cerrar esta brecha, proporcionando una validación adicional y específica del modelo en términos de su capacidad para generar decisiones de trading rentables.

5. Entorno de trabajo

Una de las complejidades más notables en el campo del entrenamiento de modelos de Machine Learning es la necesidad de contar con una capacidad de cómputo significativa. Este desafío fue particularmente relevante en el desarrollo de este proyecto. La naturaleza intensiva en recursos de los modelos avanzados, especialmente aquellos involucrados en el análisis de series temporales y la implementación de redes neuronales como LSTM, requiere un entorno de trabajo que cuente con hardware con alta capacidad de cómputo para este tipo de tareas.

Las GPU (Unidades de Procesamiento Gráfico) han adquirido una importancia crucial en el campo del entrenamiento de modelos de Machine Learning, particularmente para tareas como el entrenamiento y la implementación de redes neuronales profundas. Su arquitectura especializada permite manejar eficientemente grandes cantidades de operaciones paralelas, una característica esencial para el procesamiento de algoritmos de aprendizaje profundo. Las GPU son capaces de ejecutar miles de hilos de cálculo simultáneamente, lo que las hace significativamente más rápidas que las CPU tradicionales para tareas específicas de Machine Learning. Esta capacidad para realizar operaciones matemáticas complejas y manejar grandes volúmenes de datos de manera eficiente no sólo acelera el proceso de entrenamiento, sino que también posibilita la experimentación con modelos más grandes y complejos, facilitando avances significativos en la precisión y capacidad predictiva de los modelos de Machine Learning.

5.1. Colab

Durante gran parte del desarrollo de este proyecto el entorno utilizado fue el de Google Colab[15]. Esta es una plataforma de notebooks interactivos que ofrece un entorno de codificación en la nube. Aunque Google Colab ofrece acceso a recursos como GPUs y TPUs que pueden acelerar significativamente el entrenamiento de

[15] “Google Colab.” https://colab.research.google.com/

modelos complejos, en este caso, no se utilizaron dichos recursos debido a las restricciones de la versión gratuita.

La ejecución inicial de los modelos se realizó utilizando únicamente la CPU disponible en Colab. Esta limitación obligó a reducir la escala del dataset y a simplificar la complejidad del modelo para adecuarse a la capacidad de cómputo limitada. Esta adaptación tuvo un impacto notable en la cantidad de datos que se podían procesar simultáneamente, lo que a su vez repercutió en la precisión del modelo y en su habilidad para generalizar basándose en los datos disponibles.

Inicialmente, se lograron resultados prometedores utilizando solamente los valores de precio como entrada y manteniendo un tiempo de entrenamiento razonable. Sin embargo, al intentar enriquecer el modelo con más datos y aumentar su complejidad, nos encontramos con que los tiempos de entrenamiento se extendían considerablemente. Esto hizo impracticable el uso de un modelo más avanzado y ajustado a la mayor dimensión de los datos de entrada, resaltando las limitaciones inherentes al uso exclusivo de la CPU en contextos de entrenamiento de modelos más sofisticados y datos extensos.

Esta limitación llevó a considerar alternativas, incluyendo la posibilidad de acceder a recursos computacionales más potentes, como GPUs, mediante la suscripción de pago de Colab, para manejar la carga de datos de manera más efectiva y mejorar el rendimiento del modelo.

5.2. Entorno Personal

Enfrentando las limitaciones de capacidad de cómputo al utilizar Google Colab, se llevó a cabo una investigación para encontrar alternativas viables. De esta búsqueda emergieron dos opciones principales.

La primera opción consistía en optar por el plan premium de Google Colab, que proporciona acceso a una GPU específicamente diseñada para el entrenamiento de modelos de redes neuronales. Aunque este plan representa una mejora significativa al ofrecer recursos especializados, sigue presentando limitaciones en términos de

unidades de cómputo disponibles. Estas restricciones pueden resultar insuficientes para llevar a cabo entrenamientos extensivos y pruebas exhaustivas de modelos.

La segunda opción es aprovechar los recursos de hardware personales. En este caso, se cuenta con un tarjeta gráfica AMD Radeon RX 580 de 8GB. Aunque esta GPU no se equipara en rendimiento a las ofrecidas por Colab, es significativamente más potente que un procesador CPU convencional. La configuración completa del ordenador es la siguiente:

- Procesador: AMD Ryzen 1500X
- GPU: AMD SAPPHIRE NITRO+ Radeon RX 580 8GB GDDR5
- Ram: 16 GB DDR4
- HDD: 320 GB
- SO: Ubuntu 20.04.6 LTS

Un aspecto importante a considerar es que la GPU Radeon RX 580 dejó de recibir soporte oficial desde el año 2021 en su software para aprendizaje automático. Esta situación conlleva ciertas complicaciones, como la falta de actualizaciones de drivers y potenciales problemas de compatibilidad con software nuevo o actualizado. Esto llevó a investigar cómo configurar y optimizar un entorno de trabajo que permita el uso efectivo de esta tarjeta gráfica en proyectos de cómputo intensivo.

En este proceso, se deben considerar factores como la compatibilidad del sistema operativo, la instalación de drivers adecuados, y la configuración de entornos de desarrollo que soporten la GPU. Además, es esencial explorar la compatibilidad de esta tarjeta con las bibliotecas y herramientas específicas que se planea utilizar, especialmente aquellas relacionadas con el aprendizaje automático y el procesamiento de datos a gran escala.

Finalmente, se decidió que, como primer paso, se intentarían realizar las configuraciones utilizando la GPU disponible, y si esto no fuera viable, se optaría por el plan premium de Google Colab.

Configuración del entorno

Dadas las limitaciones y requerimientos descritos anteriormente, la configuración del entorno de trabajo involucra consideraciones meticulosas en cuanto a drivers, software, etc. Una solución eficaz para manejar estas complejidades es el uso de Docker[16]. Esta herramienta ofrece un entorno controlado y consistente, facilitando la replicabilidad y la gestión de dependencias, lo que resulta crucial en proyectos de cómputo intensivo como el aprendizaje automático. Además, Docker permite la separación entre la configuración del hardware y el software, lo que significa que los ajustes en el sistema operativo o los drivers no afectan directamente el entorno de desarrollo.

Para utilizar Docker fue necesario instalar Ubuntu como sistema operativo, eligiendo la versión 20.04.6 LTS. Esta versión fue seleccionada por su estabilidad, soporte a largo plazo (LTS), y su buena compatibilidad con herramientas de aprendizaje automático y entornos de desarrollo como Docker.

En cuanto a Docker, se utilizó una imagen[17] de entorno ya configurada y optimizada para trabajar con la GPU AMD Radeon RX 580 en tareas de aprendizaje automático, frente a la falta de soporte oficial reciente.

La configuración del entorno utilizado a partir de la imagen de docker es la siguiente:

- **ROCm version 3.7**: ROCm (Radeon Open Compute)[18] es una plataforma de cómputo abierto que proporciona soporte para una gama de GPUs AMD. Actualmente se encuentran en la versión 5.6, pero debido a que a partir de la versión 4.0 AMD eliminó el soporte para la RX 580 fue necesario utilizar una versión anterior a esta.
- **Ubuntu 18.04 LTS**: La imagen docker cuenta con esta versión antigua de Ubuntu la cual es compatible con los drivers y softwares utilizados en el entorno.

[16] “Docker.” https://www.docker.com/
[17] “rocm/pytorch:rocm3.7_ubuntu18.04_py3.6_pytorch_gcc.” https://hub.docker.com/layers/rocm/pytorch/rocm3.7_ubuntu18.04_py3.6_pytorch_gcc/images/sha256-8fc2b7ccb8f3560d63716af5dcd95171c2a4b198e78dd319fa49751d65e28ccf
[18] “Radeon Open Compute.” https://www.amd.com/en/products/software/rocm.html

- **Python 3.6.9**: Versión más antigua de python compatible con los sistemas y bibliotecas específicos utilizados en este entorno.
- **PyTorch 1.7.0a0+19fb0b8**: Una versión de desarrollo de PyTorch a partir de la versión 1.7.0. PyTorch es una biblioteca de aprendizaje automático de código abierto. El identificador 19fb0b8 sugiere una versión específica dentro del proceso de desarrollo, asociada a una serie de cambios o adiciones con la finalidad de adaptarla a este entorno.

6. Aprendizaje

En la fase de entrenamiento del modelo, se adoptaron dos metodologías distintas. La primera metodología no establece distinciones entre las diversas categorías de datos, incorporándolos en una sola agrupación para ser procesados por una red LSTM. Este enfoque integral procesa todos los datos de manera uniforme, sin segmentación específica.

Posteriormente, se desarrolló una segunda metodología como una evolución del enfoque inicial, basada en las observaciones y resultados obtenidos. Esta metodología avanzada implica la utilización de un modelo ensamblado. Esta estrategia permite diferenciar y manejar de manera individualizada cada conjunto de datos. Tras esta segmentación y tratamiento específico, los datos se reintegran para dar una única salida. Además, se efectuaron ajustes y mejoras en el procesamiento de los datos para optimizar los resultados. Este enfoque diferenciado busca mejorar la precisión y eficacia del modelo, aprovechando las características de cada tipo de dato.

6.1. Enfoque de Datos Integrados

Esta primera aproximación se llevó a cabo en el entorno de Colab. Dadas las restricciones inherentes a la capacidad de cómputo de este entorno, se optó por modelos de menor complejidad dimensional y por el uso de un conjunto limitado de datos, específicamente centrado en un rango temporal acotado. A pesar de estas limitaciones, este enfoque preliminar resultó eficaz para evaluar el comportamiento del modelo frente a los datos.

Con el objetivo de analizar el efecto que cada nuevo conjunto de datos tenía sobre el desempeño del modelo, se procedió a la integración progresiva de los distintos conjuntos de datos en cada entrenamiento. Los datos utilizados para estos análisis proceden de la empresa Apple Inc., la cual es la empresa con mayor capitalización de mercado actualmente, lo que proporcionó una valiosa oportunidad para estudiar el

comportamiento del modelo en el contexto de datos financieros de una empresa tecnológica líder en el mercado.

Para el proceso de testeo, se optó por un periodo específico caracterizado por la ausencia de tendencias pronunciadas en el precio de las acciones. Esta selección fue estratégica, ya que en escenarios de tendencia alcista, por ejemplo, resulta más sencillo generar ganancias en comparación con un mercado bajista. El rango de fechas de testeo es desde el 22 de abril de 2022 hasta 30 de junio de 2023. Es importante mencionar que los datos expresados a continuación son en días hábiles o que haya operado el mercado.

La configuración utilizada en todos los casos fue la siguiente:

- Dataset:
    - Cantidad total de datos: 700 días.
    - Dimensión de cada patrón de entrada: 100 días
    - Dimensión de cada variable objetivo: 1 día
- Modelo:
    - Arquitectura: LSTM con una capa oculta.
    - Tamaño de la capa oculta: 50 unidades.
    - Número de capas: 1.
    - Función de pérdida: Error Cuadrático Medio (MSE).
    - Optimizador: Adam con tasa de aprendizaje inicial de 0.001.
- Entrenamiento
    - Número de épocas: 70.
    - Tamaño del lote: 64.
    - Validación: Ventana expansiva con 450 datos iniciales de entrenamiento, un horizonte de predicción de 1 y un período de expansión de 1.

Con esta configuración se realizaron distintos entrenamientos, en los cuales se fue ampliando el conjunto de datos.

Entrenamiento 1: Se empleó como variable de entrada la variación logarítmica (log_change) del precio, que coincide con el objetivo de predicción (target), para establecer una línea base de desempeño del modelo. Las siguientes gráficas

muestran la evolución del error cuadrático medio durante el entrenamiento y el resultados de las predicciones.

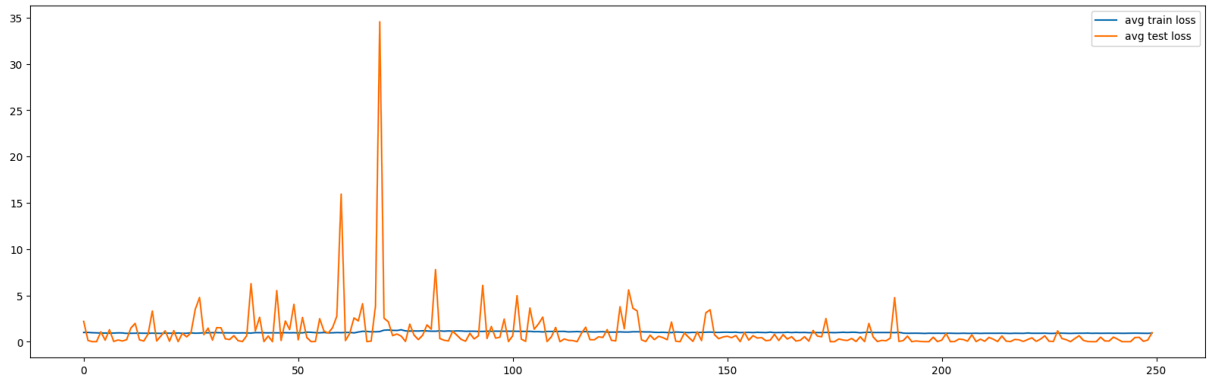

Fig. 7 Comparación de la Pérdida Promedio en Entrenamiento y Prueba

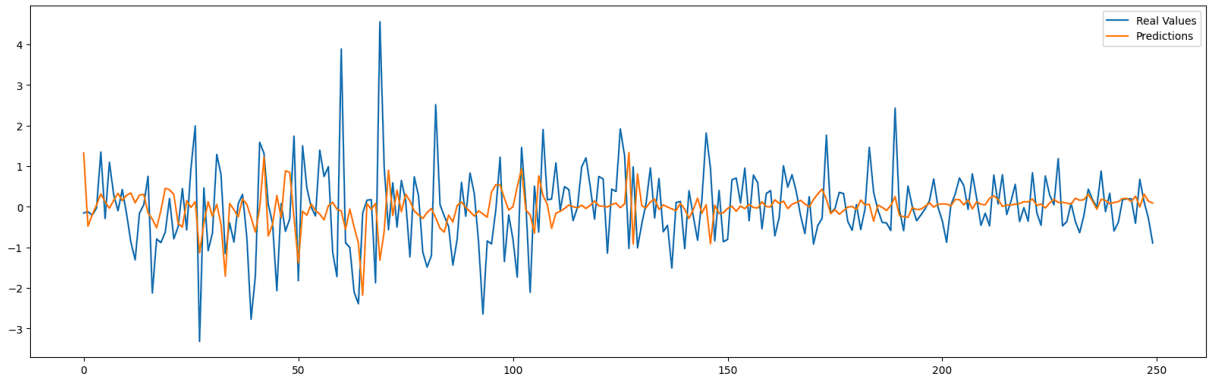

Fig. 8 Predicciones y valores reales del cambio logarítmico del precio.

Luego se realizó una simulación de trading automático, ejecutando compras y ventas de acciones de acuerdo a los valores predichos. En caso de que el cambio logarítmico supere un umbral definido se realiza una compra. De manera similar, cuando el cambio logarítmico decae cierto umbral se realiza una venta. A partir de estas operaciones se obtuvo un porcentaje de retorno.

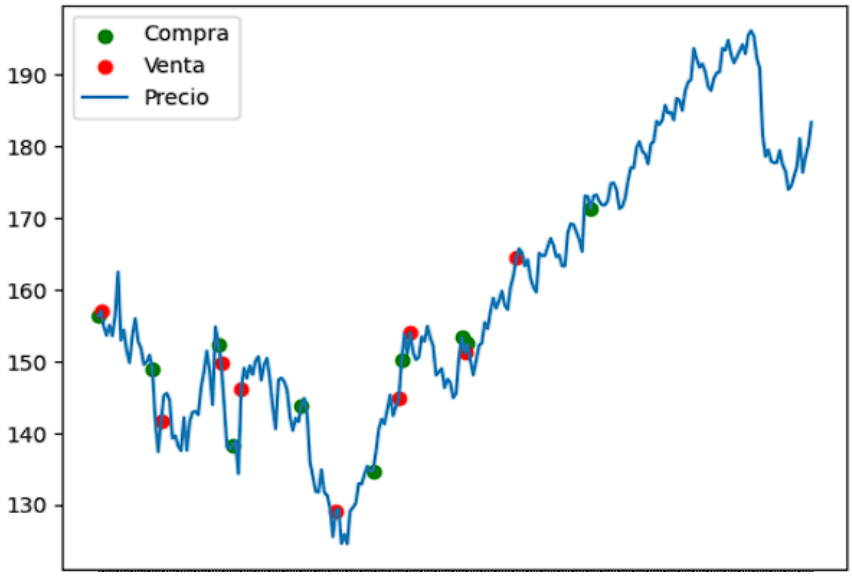

Fig. 9 Compras y ventas de la simulación

Porcentaje de retorno: 17.15% en 250 días

Entrenamiento 2: A la variación logarítmica del precio, se añadió el conjunto de datos obtenido de Yahoo Finance, variables macroeconómicas obtenidas de la Reserva Federal (FRED).

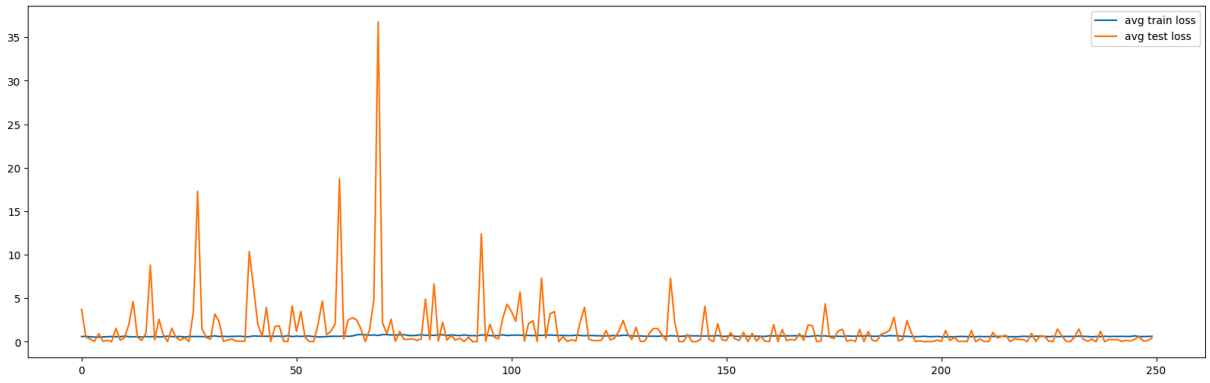

Fig. 10 Comparación de la Pérdida Promedio en Entrenamiento y Prueba

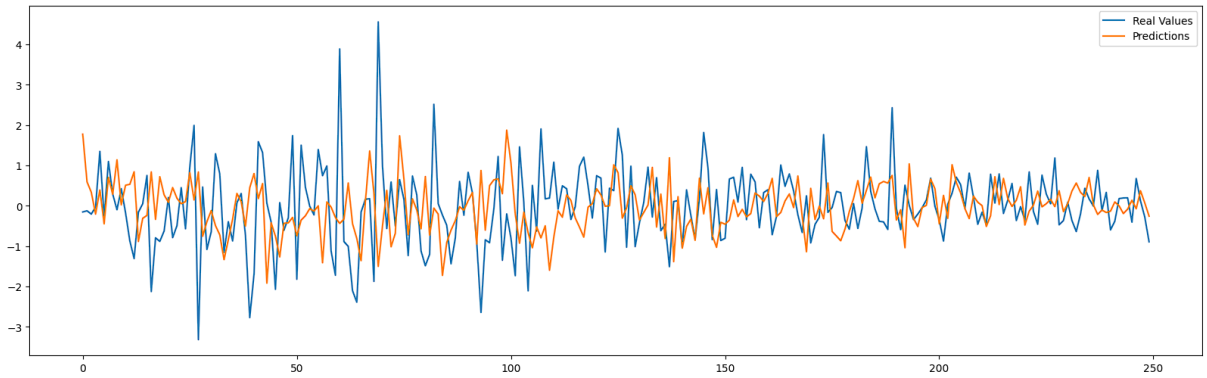

Fig. 11 Predicciones y valores reales del cambio logarítmico del precio.

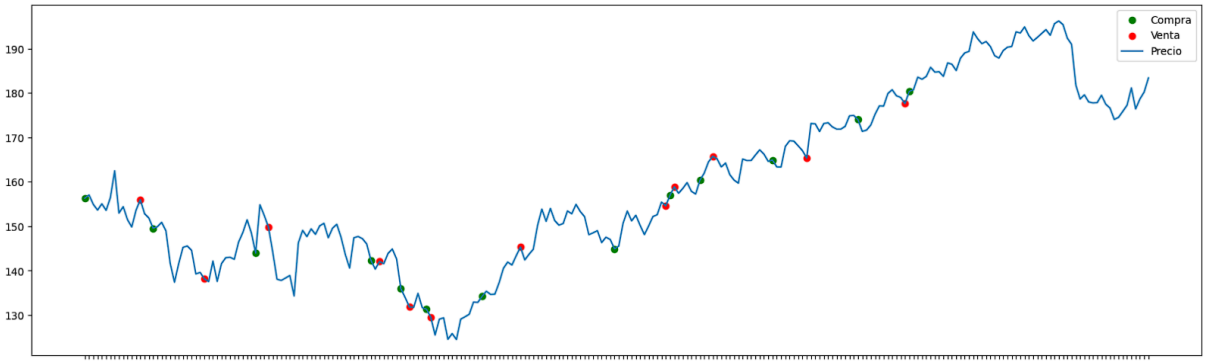

Fig. 12 Compras y ventas de la simulación

Porcentaje de retorno: -2.03% en 250 días

Entrenamiento 3: Se extendió el conjunto de datos anterior con información proporcionada por la Comisión de Bolsa y Valores (SEC), integrando así indicadores fundamentales de la empresa.

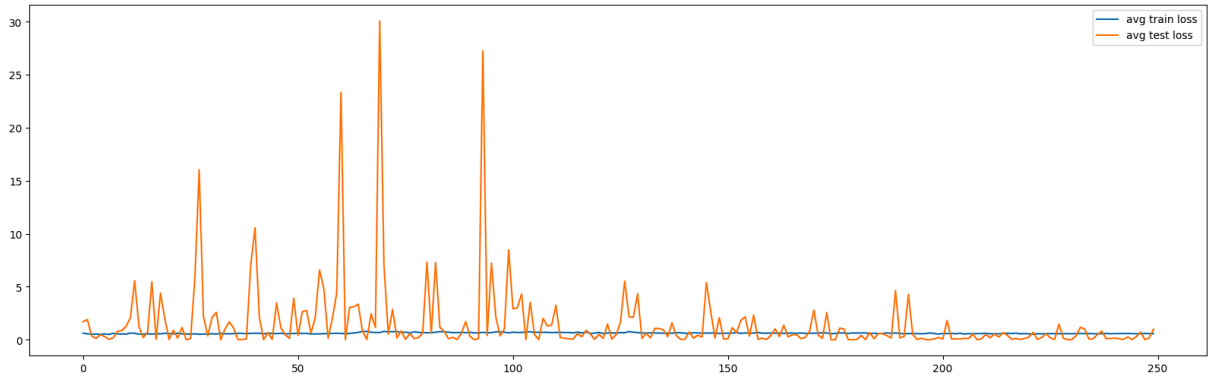

Fig. 13 Comparación de la Pérdida Promedio en Entrenamiento y Prueba

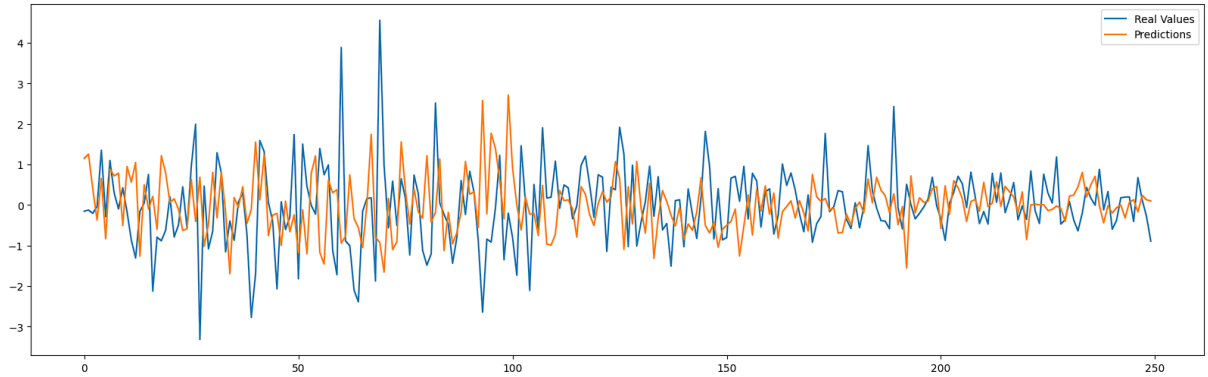

Fig. 14 Predicciones y valores reales del cambio logarítmico del precio.

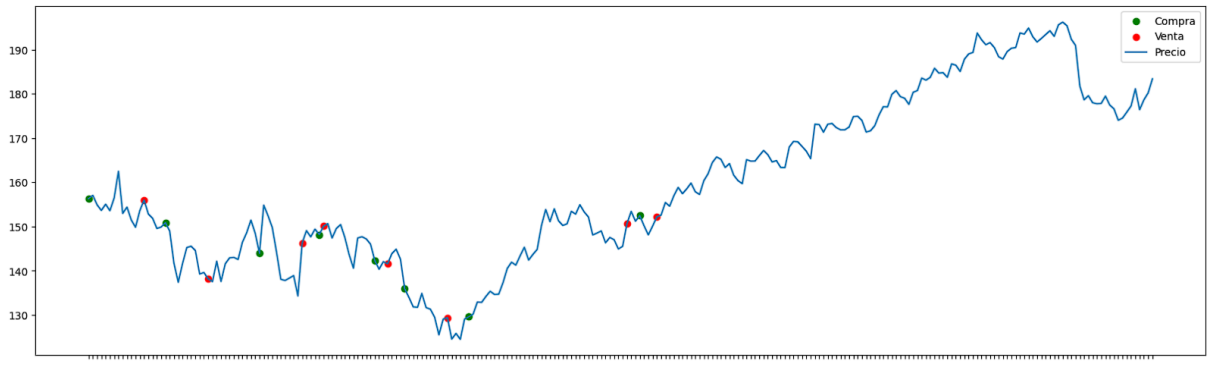

Fig. 15 Compras y ventas de la simulación

Porcentaje de retorno: 16.28% en 250 días

Entrenamiento 4: Finalmente, se añadieron datos sobre el producto bruto interno de las distintas industrias, buscando capturar aún más aspectos macroeconómicos que pudieran influir en la dinámica del mercado.

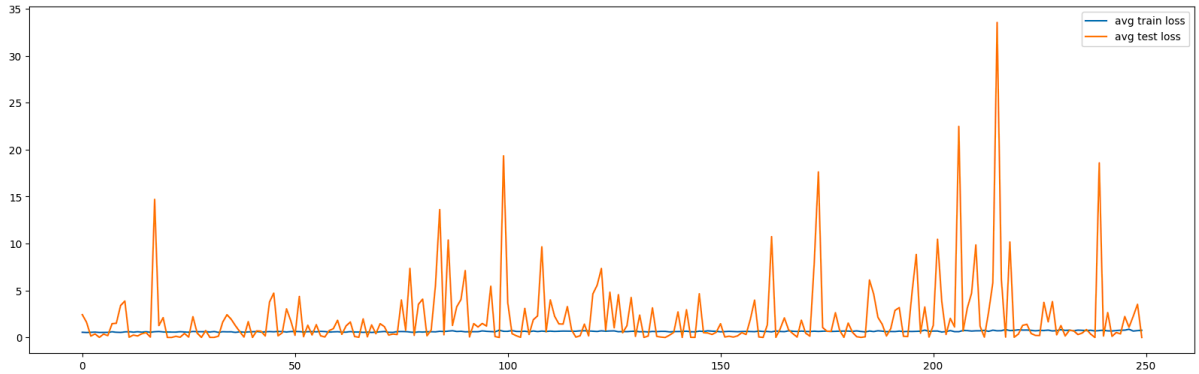

Fig. 16 Comparación de la Pérdida Promedio en Entrenamiento y Prueba

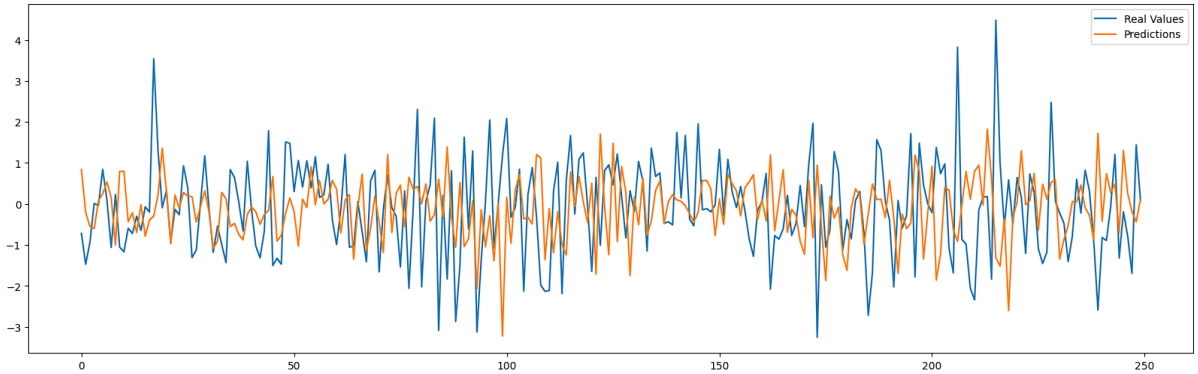

Fig. 17 Predicciones y valores reales del cambio logarítmico del precio.

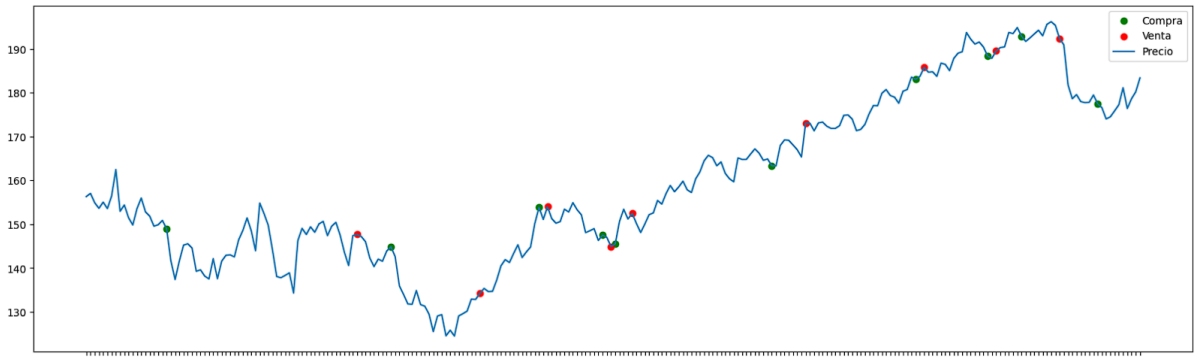

Fig. 18 Compras y ventas de la simulación

Porcentaje de retorno: 7.83% en 250 días

Cada etapa de entrenamiento tenía como objetivo evaluar cómo la inclusión de diferentes tipos y fuentes de datos afecta la precisión de las predicciones del modelo. Es importante señalar que los porcentajes de retorno obtenidos en las simulaciones de trading son sensibles a los umbrales definidos. Si bien deben considerarse con prudencia al evaluar el aprendizaje del modelo, estos umbrales son útiles para desarrollar y ajustar estrategias de trading.

En esta fase inicial, se detectó que los resultados del modelo presentan una variabilidad considerable. Generalmente, se anticiparía que al añadir nuevos datos al modelo, su capacidad de predicción se mantendría o incluso mejoraría, dado que dispone de más información para el aprendizaje. Sin embargo, esta expectativa no se cumplió en la práctica. Por lo tanto, la sección siguiente se enfoca en un análisis más detallado tanto de los datos como de la arquitectura del modelo y se utiliza hardware con mayor capacidad de cómputo lo que permite implementar modelos más complejos.

6.2. Enfoque con Modelo Ensamblado

Procesamiento de datos trimestrales

A partir de los resultados obtenidos, se llevaron a cabo modificaciones significativas en el tratamiento de los datos. Una de las etapas cruciales revisadas fue el análisis de procesamiento de datos. Anteriormente, se había destacado la utilidad de la transformación logarítmica en los datos, especialmente por sus ventajas inherentes. No obstante, se identificó un desafío particular en esta metodología: dado que el conjunto de datos incluye numerosas entradas trimestrales convertidas a un formato diario mediante el método de relleno con el último valor conocido, la aplicación de la transformación logarítmica resulta en una predominancia de ceros. Esto se debe a que los valores repetidos, tras la transformación, se convierten en ceros, excepto en los momentos en que se actualiza con nuevos datos trimestrales.

Esta situación conduce a un patrón de entrada donde predominan los ceros, provocando un sobreajuste del modelo a este patrón específico.

Para abordar este problema, se decidió omitir la transformación logarítmica en los datos que originalmente no tienen temporalidad diaria. Esto permitió mantener los valores en su escala original y evitar la generación de ceros consecutivos, que era una consecuencia directa de aplicar la transformación logarítmica a datos constantes. En el enfoque anterior, debido al relleno con el último valor conocido, todos los valores dentro de un trimestre se convertían en ceros tras la transformación, excepto en los días en que se actualizaba con un nuevo dato trimestral. Esto resultaba en un patrón donde únicamente el valor correspondiente a la nueva observación trimestral era distinto de cero, mientras que el resto del trimestre seguía mostrando ceros. Con el nuevo enfoque, aunque todavía persisten valores repetidos debido al rellenado, se elimina este comportamiento de ceros predominantes, lo que contribuye a reducir el riesgo de sobreajuste del modelo.

Reduccion datos GDP por industria

En una sección previa del informe, se destacó la importancia de ciertos conjuntos de datos, especialmente aquellos que incluyen información sobre la actividad económica por industria, medida a través del Producto Bruto Interno (PBI) o, en su terminología inglesa, Gross Domestic Product (GDP). Estos datos se organizan en tablas nombradas bajo la nomenclatura *gdp_[table_id]_[industry]*.

El número total de estas tablas supera las 2000 unidades. En fases anteriores del proyecto, la selección de tablas se realizaba mediante la identificación de un *table_id* específico. Sin embargo, para esta etapa, se adoptó un enfoque diferente centrado en la relevancia de las industrias en relación con la predicción del activo de interés. El método empleado para esta selección fue el uso del algoritmo Random Forest Regressor[19], provisto por la biblioteca sklearn en Python, con los datos de precios trimestrales como variable objetivo.

[19] "Random Forest Regressor."
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor

La eficacia de este método reside en su habilidad para identificar, de entre un conjunto amplio de variables potencialmente influyentes, aquellas que tienen un impacto significativo en el modelo predictivo. Como resultado de este proceso, se logró una selección más específica y enfocada, limitando las tablas a las 20 industrias más relevantes. Esta estrategia es crucial para prevenir la sobrecarga del modelo con características relacionadas exclusivamente al PBI, lo cual podría minimizar la importancia de otras variables clave.

Arquitectura ensamblada

La Arquitectura de Ensamblado representa un enfoque avanzado en el procesamiento de datos para modelos de aprendizaje automático, particularmente útil cuando se manejan conjuntos de datos de naturaleza y temporalidad diversas. Esta arquitectura implica la integración de múltiples subredes, cada una especializada en el tratamiento de tipos específicos de datos, lo cual es esencial en el contexto del análisis financiero, donde la variedad y la complejidad de los datos son significativas.

Como se pudo observar los datos utilizados varían en su naturaleza y frecuencia. Por ejemplo, se manejan datos diarios como el precio y el volumen de las acciones, que ofrecen una perspectiva a corto plazo de la actividad del mercado. Estos datos son dinámicos, reflejando las fluctuaciones del mercado en tiempo real o casi real. Por otro lado, también se utilizan datos trimestrales asociados a los reportes financieros de las empresas, que proporcionan una visión más profunda y a largo plazo del desempeño y la estabilidad financiera de las corporaciones. Además, se incorporan variables macroeconómicas, las cuales ofrecen una perspectiva más amplia del entorno económico en el que operan las empresas y los mercados financieros.

Un Arquitectura Ensamblada es particularmente eficaz en este entorno debido a su capacidad para tratar cada tipo de dato de manera diferencial. En lugar de procesar toda la información de forma homogénea, este tipo de arquitectura permite que cada subred se especialice en un tipo específico de datos, optimizando el procesamiento y

la interpretación. En este caso, se definieron 4 subredes, una para cada conjunto de datos.

La primera subred, implementada con una arquitectura de red neuronal recurrente LSTM, se dedica al análisis de datos de precios y volúmenes de activos financieros. Esta subred está especialmente diseñada para capturar tendencias y patrones a corto plazo, aprovechando la capacidad de las LSTM para discernir relaciones temporales complejas en series de tiempo.

Por otro lado, contamos con 3 subredes construidas con capas densamente conectadas. La primera se encarga del análisis de los datos de informes financieros trimestrales (indicadores fundamentales). Esto permite extraer insights valiosos sobre la salud financiera y las proyecciones a largo plazo de las empresas. Al procesar estos datos trimestrales, la subred puede identificar indicadores de rendimiento y tendencias estratégicas.

La segunda, está específicamente orientada hacia el análisis de variables macroeconómicas. Esta subred evalúa cómo los cambios en el entorno económico global y local pueden influir en los mercados financieros, ayudando a entender mejor el comportamiento de los activos financieros bajo diferentes condiciones económicas.

Mientras que la última subred está tambien destinada a procesar y analizar variables macroeconomicas, pero específicamente del Producto Bruto Interno (PBI) por industria. Esta subred aporta una comprensión profunda de las tendencias económicas a nivel sectorial, lo cual es fundamental para comprender las dinámicas del mercado y la interacción entre distintos sectores económicos.

Finalmente, una vez que cada subred ha procesado sus datos y generado su salida, estas salidas se combinan en una única representación. Esto se logra concatenando las salidas de las subredes, formando así un tensor unificado que contiene la información procesada por todas las subredes. Este tensor combinado se introduce entonces en una capa densa adicional que realiza una transformación para

luego producir la salida final del modelo. En la Fig. 19 se puede ver un diagrama del modelo ensamblado.

Este enfoque de integración garantiza que el modelo no solo considere las características únicas de cada conjunto de datos, sino que también aprenda cómo estas características interactúan, lo que resulta en predicciones más precisas.

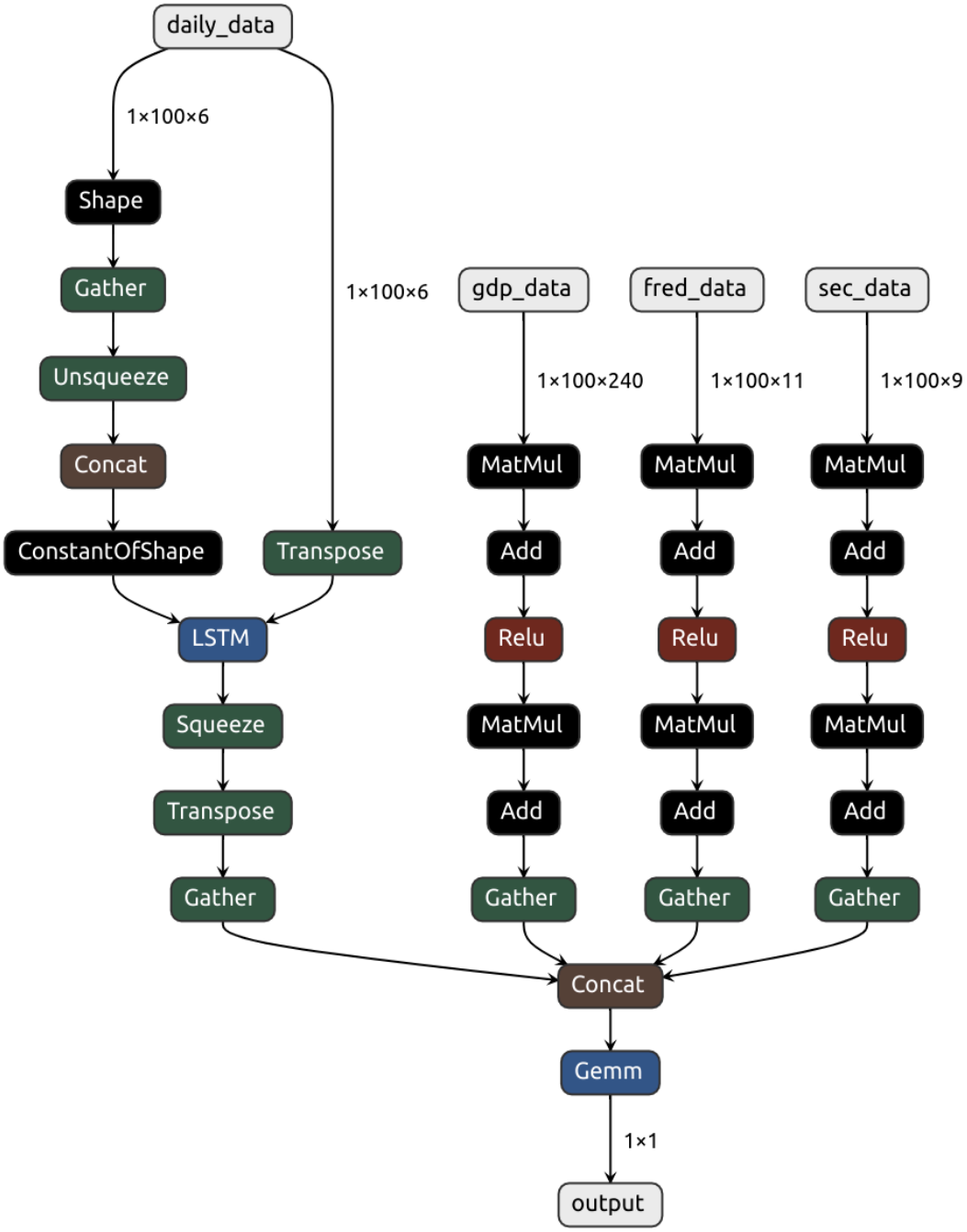

Fig. 19: Diagrama modelo ensamblado utilizado.

7. Descripción y Funcionamiento del Sistema

En este capítulo se presenta el diseño y descripción detallada del sistema desarrollado para la predicción de la evolución de activos financieros mediante el enfoque de modelo ensamblado. El sistema está conformado por múltiples módulos que interactúan entre sí para recolectar, procesar y analizar datos financieros, con el objetivo final de entrenar un modelo de red neuronal que permita realizar predicciones sobre el comportamiento futuro de precios de activos. A continuación, se describen los componentes del sistema, sus especificaciones y el modo de operación.

7.1. Módulo de Configuración

Este módulo se encarga de configurar el entorno de ejecución y las dependencias necesarias.

Tareas Principales

- Verificación de disponibilidad de CUDA para acelerar el procesamiento con GPU.
- Instalación de librerías y paquetes necesarios, como sqlalchemy, pandas_ta, yahoo_fin, torch entre otros.
- Configuración de las variables de entorno con las claves de acceso a las APIs utilizadas, leyendo un archivo config.json.

Especificaciones

- **Lenguaje:** Python.
- **Herramientas Utilizadas:** sqlalchemy, pandas, yahoo_fin, json, os.
- **Entrada:** Archivo config.json con las credenciales de las APIs.
- **Salida:** Entorno configurado y listo para la ejecución de los módulos posteriores.

Gestión de Credenciales

- Las claves de acceso a las APIs se manejan de manera segura mediante un archivo config.json, evitando exponer información sensible en el código.

7.2. Módulo de Adquisición de Datos

Este módulo se encarga de recolectar datos financieros de diferentes fuentes externas.

Fuentes de datos

- **Yahoo Finance**: Obtiene datos históricos de precios de acciones.
- **FRED (Federal Reserve Economic Data)**: Recolecta variables macroeconómicas.
- **BEA (Bureau of Economic Analysis)**: Obtiene datos del Producto Interno Bruto (PIB) por industria (variables macroeconómicas).
- **SEC (Securities and Exchange Commission)**: Recupera información financiera reportada por las empresas (Indicadores fundamentales).

Especificaciones

- **Herramientas Utilizadas**: Librerías específicas como yahoo_finance y conexiones a APIs.
- **Datos Obtenidos**:
    - Precios históricos de acciones para el ticker especificado.
    - Indicadores macroeconómicos (tasas de interés, índices de producción, etc.).
    - Datos del PIB por industria.
    - Estados financieros trimestrales de empresas.
- **Formato de Datos**: Datos en formato de DataFrames de Pandas.
- **Frecuencia de Datos**: Datos diarios y trimestrales.
- **Entrada**: Ticker del activo financiero y rango de fechas.
- **Salida:** DataFrames con los datos recolectados.

Modo de Operación

- Se realiza una conexión a la API de yahoo finance utilizando las credenciales y parámetros necesarios.
- Se envían solicitudes a la base de datos local para obtener los datos de BEA, FRED y SEC mediante conexiones establecidas con SQLAlchemy.

- Los datos obtenidos se convierten en DataFrames para facilitar su manipulación.

7.3. Módulo de Procesamiento de Datos

Este módulo se encarga de preparar los datos para su uso en el modelo predictivo, aplicando una serie de transformaciones y técnicas de integración a los datos.

Especificaciones

- **Herramientas Utilizadas:** Pandas, NumPy, Scikit-learn.
- **Entrada:** DataFrames con datos brutos de las diferentes fuentes.
- **Salida:** DataFrames procesados y listos para ser utilizados en el entrenamiento del modelo.

Modo de Operación

- **Transformaciones de Datos y Preparación**
    - **Resampling datos trimestrales**: Conversión de datos trimestrales a temporales diarios para alinear con los datos de precios.
    - **Resampling datos de precio**: Conversión de datos diarios a trimestrales para alinear con la frecuencia de los datos trimestrales de la BEA en el análisis con Random Forest..
    - **Cálculo de Cambios Logarítmicos**: Normaliza las series temporales y facilita el modelado de tasas de cambio.
    - **Relleno y Manejo de Datos Faltantes**: Uso de métodos como forward-fill para completar datos faltantes, eliminación de columnas innecesarias y manejo de valores nulos.
    - **Estandarización y Normalización**: Aplicación de escaladores (StandardScaler) para ajustar las variables a una escala común.
- **Integración de datos**
    - **Combinar Datos de Diferentes Fuentes:** Integración de precios históricos, variables macroeconómicas, PIB e información financiera en conjuntos de datos unificados.

    - **Alineación de Índices Temporales:** Sincronización de fechas y alineación de los índices temporales de los diferentes conjuntos de datos para asegurar la coherencia temporal entre las series.
- **Selección de Características**
    - **Análisis de Correlación y Random Forest**: Se utilizan modelos de Random Forest para identificar las variables más relevantes.
    - **Filtrado de Características Relevantes**: Se filtran los conjuntos de datos para incluir solo las columnas correspondientes a las industrias más relevantes.
- **Armado de dataset**
    - **Creación de Ventanas Deslizantes:** Se crea el dataset de entrenamiento y validación para utilizar el método de ventanas deslizantes.
    - **Parámetros**:
        - Número total de días utilizados para entrenamiento y validación.
        - Cantidad de datos usados para cada predicción.
        - Cantidad de días a predecir en cada predicción.
        - Salto de ventana.

7.4. Módulo de Red Neuronal

Este módulo implementa y entrena un modelo de red neuronal para la predicción del cambio de precios de activos.

Especificaciones

- **Herramientas Utilizadas:** Torch, Pandas, Numpy**.**
- **Entradas:** Conjuntos de datos procesados y divididos en secuencias temporales para entrenamiento.
- **Salidas:** Modelo entrenado y predicciones generadas.
- **Arquitectura del modelo:** Modelo jerárquico que combina datos diarios y trimestrales.
    - **Subredes**:

        - **LSTM**: Para capturar dependencias temporales en los datos diarios (precios de acciones).
        - **Capas Densas**: Para procesar los datos trimestrales (indicadores fundamentales y variables macroeconómicas).
    - **Entradas del Modelo:**
        - Datos de precios diarios (cambios logarítmicos).
        - Datos macroeconómicos trimestrales (PIB, indicadores de FRED).
        - Datos financieros de empresas (información de la SEC).
    - **Salidas del modelo**: Predicciones de cambios en el precio del activo.
    - **Parámetros del modelo:**
        - **Función de Pérdida:** Error cuadrático medio (MSE).
        - **Optimizador:** Adam.
        - **Número de Épocas:** Configurable
        - **Tamaño del Lote:** 24.

Modo Operación

- **Entrenamiento del Modelo:**
    - El modelo se entrena iterativamente, ajustando sus pesos para minimizar la función de pérdida.
- **Validación Cruzada:**
    - Uso de ventanas expandibles (expanding window) para evaluar el modelo de manera más realista en series temporales.
- **Evaluación y Almacenamiento:**
    - Se evalúa el desempeño del modelo en conjuntos de prueba.
    - Las predicciones generadas se almacenan para análisis posterior.

7.5. Módulo de Simulación y Resultados

Este módulo utiliza las predicciones generadas por el modelo para simular estrategias de trading y evaluar el desempeño del sistema.

Especificaciones

- **Función de simulación:** Simula una estrategia de compra/venta basada en umbrales definidos aplicados a las predicciones del modelo.
- **Herramientas Utilizadas**: Python, Pandas, Matplotlib, NumPy.
- **Entradas**: Predicciones generadas por el modelo, precios reales del activo financiero, y parámetros de la estrategia (umbrales de compra y venta).
- **Salidas**: Métricas de desempeño de la estrategia (porcentaje de retorno) y visualizaciones gráficas que ilustran las señales de compra y venta sobre la serie de precios

Modo Operación

- **Preparación de los Datos:**
    - Se alinean las predicciones y los precios reales del activo financiero en el mismo marco temporal.
    - Se ajustan los índices de tiempo para asegurar la correspondencia entre las predicciones y los precios.
- **Definición de la Estrategia de Trading:**
    - Se establecen los umbrales de compra y venta que determinarán las señales de entrada y salida del mercado a partir de la desviación estándar de los datos.
- **Simulación de la Estrategia:**
    - Se recorre la serie de predicciones y, en cada punto temporal, se evalúa si se cumplen las condiciones para generar una señal de compra o venta.
    - Las reglas básicas de la estrategia son:
        - Compra: Si el estado del algoritmo se encuentra en "compra" y la predicción supera el umbral de compra.
        - Venta: Si el estado es "venta" y la predicción cae por debajo del umbral de venta.
        - En cada operación de compra o venta se cambia el estado.
    - Se actualiza el estado de la estrategia y se calcula el retorno en caso de realizar una venta.
- **Cálculo del Retorno**:

    - Se calcula el retorno acumulado de la estrategia multiplicando los rendimientos obtenidos.
    - El porcentaje de retorno se expresa como el incremento o decremento porcentual del capital inicial.
- **Generación de Visualizaciones:**
    - Se genera una gráfica que muestra la serie de precios del activo financiero, indicando con marcadores los puntos de compra y venta según las señales generadas.
    - Se grafican las predicciones del modelo para visualizar su comportamiento a lo largo del tiempo.
- **Análisis e Interpretación de Resultados:**
    - Se evalúa la efectividad de los umbrales utilizados y se pueden ajustar para optimizar la estrategia.

En la siguiente imagen se muestra un diagrama que ilustra el funcionamiento del sistema según la descripción previa, destacando el flujo de datos y la interacción entre los distintos módulos

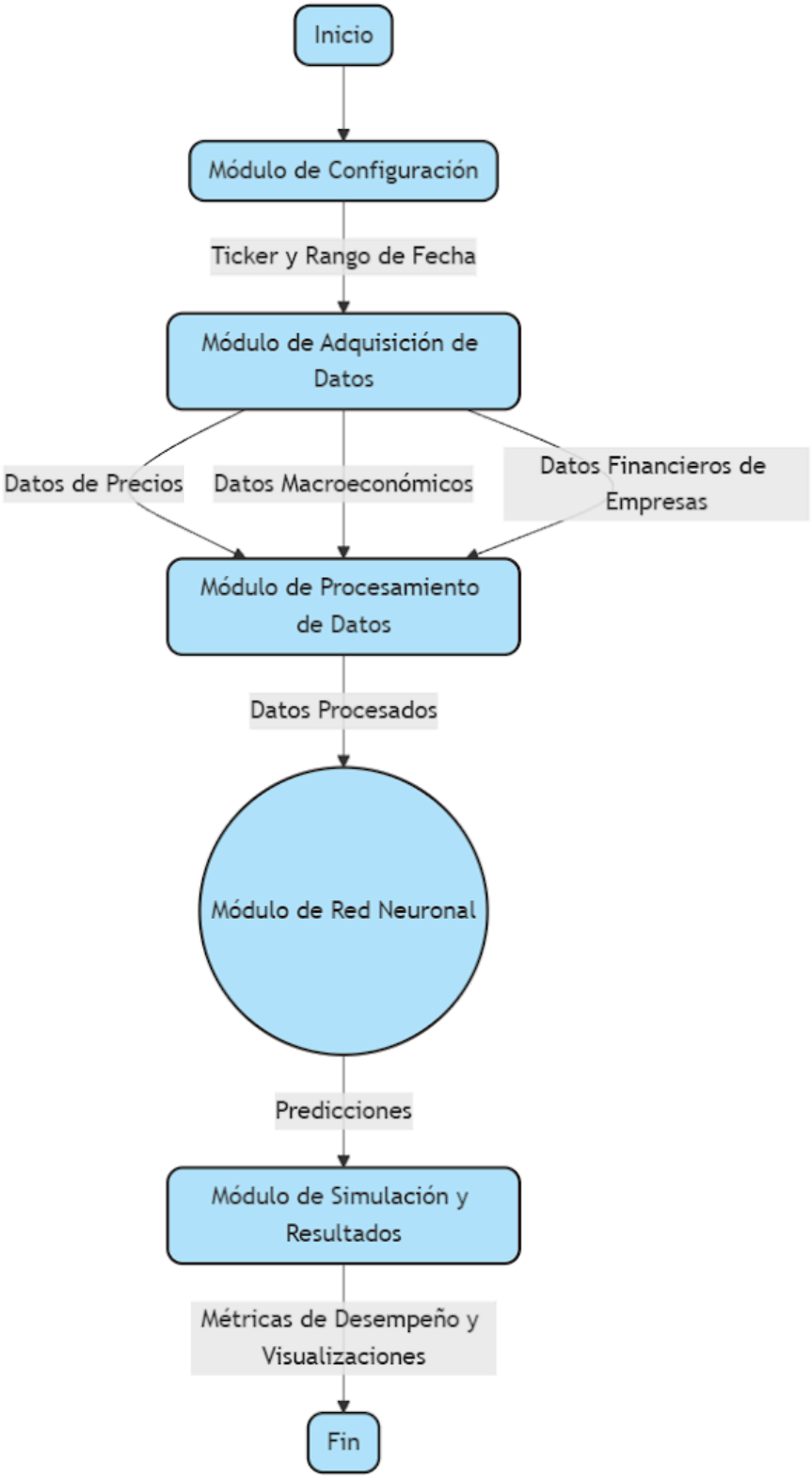

Fig. 20 Diagrama de flujo del sistema

8. Resultados

La presente sección expone los resultados derivados de la implementación de las modificaciones en los conjuntos de datos, tal como se detalló en el capítulo precedente. Además, se aborda la eficacia del modelo ensamblado tras su entrenamiento en un entorno personalizado. Estos resultados son cruciales para evaluar el impacto de las estrategias de preprocesamiento de datos y la metodología de entrenamiento adoptada.

El rango de fechas de testeo es desde el 22 de abril de 2022 hasta 30 de junio de 2023. Es importante mencionar que los datos expresados a continuación son en días hábiles o que haya operado el mercado.

8.1. Yahoo Finance - GDP

En el primer ensayo, se emplearon datos relacionados con precios y volúmenes extraídos del conjunto de datos de Yahoo Finance. Además, se incorporaron tablas de Producto Interno Bruto (GDP) suministradas por la Oficina de Análisis Económico (BEA), correspondientes a las 20 industrias más significativas en la predicción del cambio logarítmico del precio de la empresa en estudio.

Configuración

La configuración utilizada en este caso fue la siguiente:

- Dataset:
    - Cantidad total de datos: 1000 días
    - Dimensión de cada patrón de entrada: 100 días
    - Dimensión de cada variable objetivo: 1 día
    - Paso de ventana: 1 día
- Modelo:
    - Arquitectura: Modelo Ensamblado (LSTM + 1 Capa densa)
    - Tamaño capa oculta LSTM: 200

- Tamaño capa densa: 200
- Función de pérdida: Error Cuadrático Medio (MSE).
- Optimizador: Adam con tasa de aprendizaje de 0.001.

- Entrenamiento
  - Número de épocas: 80.
  - Tamaño del lote: 32.
  - Validación: Ventana expansiva con 700 datos iniciales de entrenamiento, 300 días de test, predicción de 1 día y un período de expansión de 1.

Resultados

En este escenario, el tiempo total empleado en el entrenamiento y validación alcanzó las 3 horas, 38 minutos y 15 segundos. A lo largo de la fase de entrenamiento, el error cuadrático medio promedio registrado fue de 0.571, en comparación con un promedio de 1.380 en la fase de test.

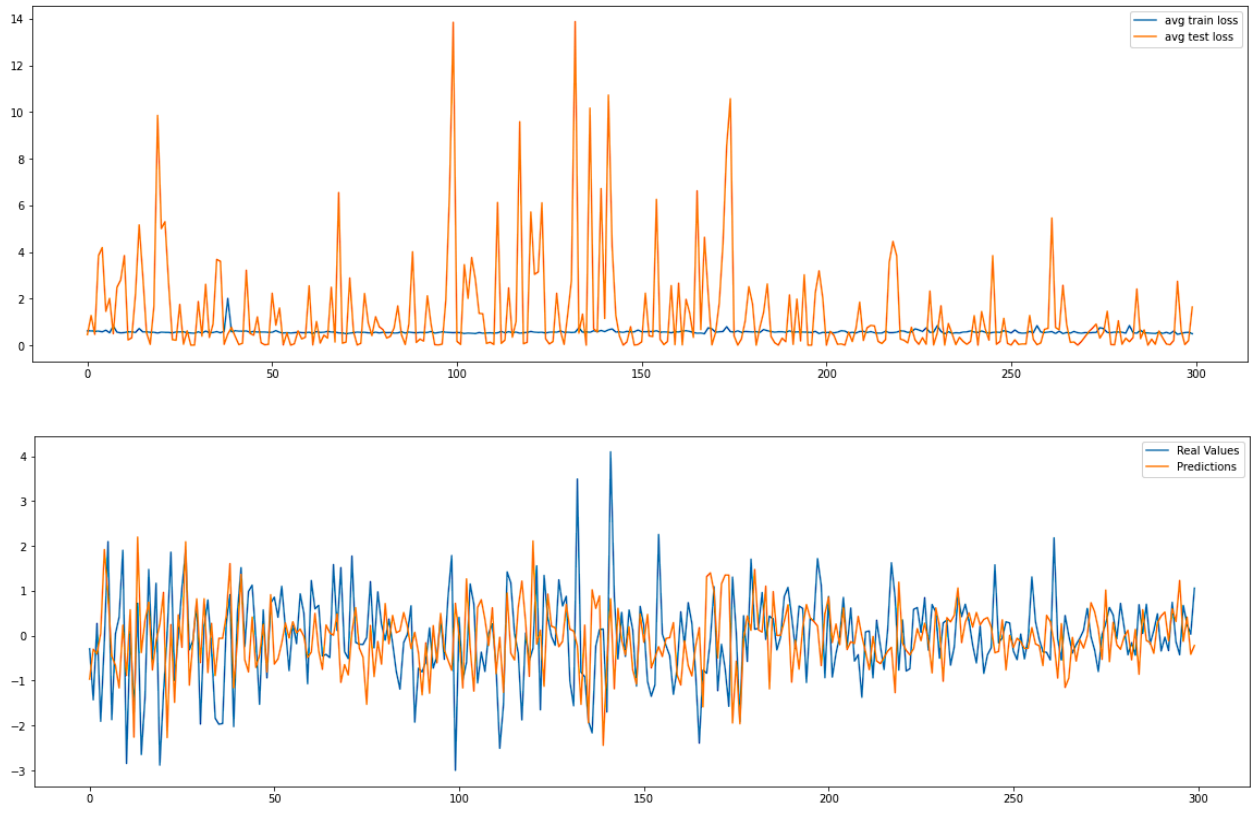

Fig. 21 MSE train y test - Cambio logarítmico del precio (YF - GDP)

Al examinar la primera gráfica, se puede notar que hay picos significativos durante la fase de testeo. Si observamos la gráfica que muestra el cambio logarítmico del precio, se puede ver que los picos en la primera gráfica están asociados con fluctuaciones significativas en el precio, sugiriendo que el modelo lucha para predecir correctamente estos eventos. Esta dificultad puede ser indicativa de una variedad de factores, como la ausencia de características relevantes que podrían proporcionar contexto adicional al modelo, la necesidad de una arquitectura de red más compleja capaz de capturar la dinámica de los cambios de precios, o la posibilidad de eventos atípicos o ruido en los datos que son intrínsecamente difíciles de predecir.

Simulación

Para evaluar el rendimiento del modelo predictivo, se codificó una función para realizar una simulación de trading utilizando las salidas del modelo para realizar operaciones de compra y venta. El objetivo es estimar el retorno porcentual de una estrategia de trading basada en las predicciones del modelo.

La simulación sigue un enfoque sistemático donde las decisiones de trading se basan en señales generadas a partir de las predicciones del modelo. Se simulan operaciones de compra cuando las predicciones superan un umbral de compra predefinido y operaciones de venta cuando caen por debajo de un umbral de venta correspondiente

Los umbrales para las señales de compra y venta en este caso se determinan multiplicando la desviación estándar de las predicciones por un coeficiente fijo. Para las simulaciones el coeficiente se establece en 0.5 para ambas operaciones. De esta manera el umbral de compra se establece en medio desvío estándar por encima de la media, y el de venta a medio desvío estándar por debajo.

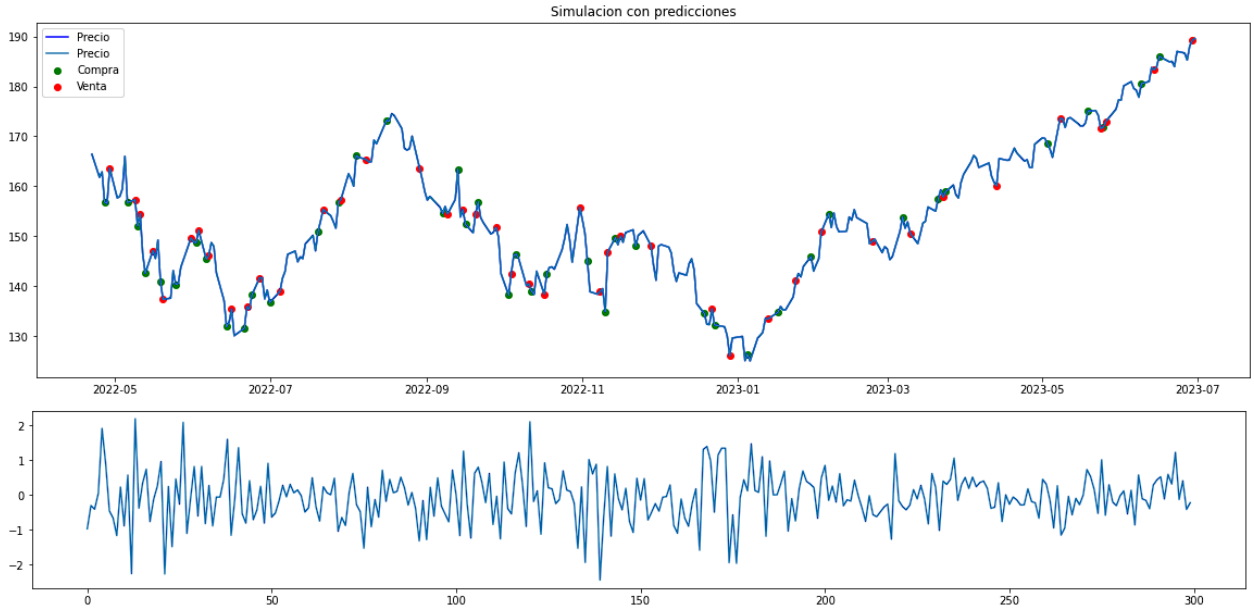

Fig. 22 Simulación Trading - Cambio logarítmico del precio (YF - GDP)

En este caso con las configuraciones especificadas se obtuvo un porcentaje de retorno de 43.93% en 299 días. Este es un excelente número, pero es importante mencionar que los resultados varían de acuerdo a como se definen los umbrales. De todas maneras, al aplicar distintas configuraciones se pudo ver que el porcentaje de retorno siempre fue positivo.

8.2. Yahoo Finance - GDP - FRED

En esta segunda instancia, a la información concerniente a precios, volúmenes y producto interno bruto, se le sumaron los datos macroeconómicos suministrados por la Reserva Federal (FRED). Estos contienen datos financieros como Índice de precio del consumidor, tasa de interés de fondos federales, tasa de participación civil en la fuerza laboral, oferta monetaria, entre otros.

Para el tratamiento de los datos mencionados, se incorporó una capa densa en la estructura del modelo. Se llevaron a cabo dos evaluaciones experimentales distintas: en la primera, se estableció una dimensión de 200 en cada subred del modelo; en la segunda evaluación, se incrementó esta dimensión a 350. Esta variación en las

dimensiones permitió observar los efectos de diferentes configuraciones en el rendimiento del modelo.

Configuración

La configuración utilizada en este caso fue la siguiente:

- Dataset:
    - Cantidad total de datos: 1000 días
    - Dimensión de cada patrón de entrada: 100 días
    - Dimensión de cada variable objetivo: 1 día
    - Paso de ventana: 1 día
- Modelo:
    - Arquitectura: Modelo Ensamblado (LSTM + 2 Capas densas)
    - Tamaño capa oculta LSTM:
        - Configuración 1: 200
        - Configuración 2: 350
    - Tamaño capas densas:
        - Configuración 1: 200
        - Configuración 2: 350
    - Función de pérdida: Error Cuadrático Medio (MSE).
    - Optimizador: Adam con tasa de aprendizaje de 0.001.
- Entrenamiento
    - Número de épocas: 80.
    - Tamaño del lote: 32.
    - Validación: Ventana expansiva con 700 datos iniciales de entrenamiento, 300 días de test, horizonte de predicción de 1 día y período de expansión de 1 día.

Resultados

En la primera evaluación el tiempo total de entrenamiento y validación fue de 3 horas 54 minutos, con un error cuadrático medio promedio en el entrenamiento de 0.592, en comparación con un promedio de 1.443 en la fase de test.

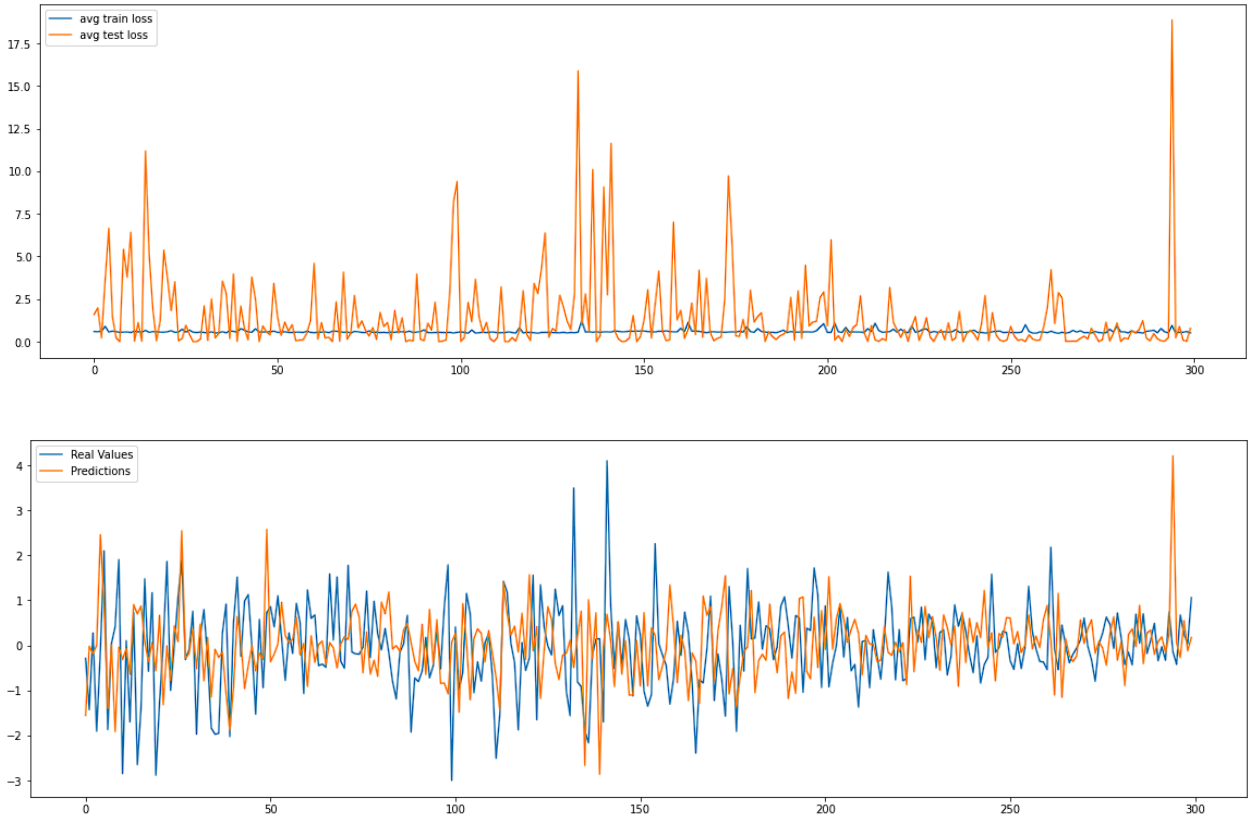

Fig. 23 MSE train y test - Cambio logarítmico del precio (YF - GDP - FRED)

El error cuadrático en las fases de predicción y test ha aumentado comparado con el caso anterior, lo que indica un rendimiento inferior. En la segunda gráfica, se observa visualmente hacia el final un error notable en la predicción del cambio logarítmico. Este fenómeno podría sugerir un problema con los datos utilizados, como la influencia de una variable que está desorientando al modelo, o la necesidad de un modelo más complejo.

En la segunda evaluación, se efectuó un incremento en las dimensiones de cada subred del modelo, pasando de 200 a 350. Esta modificación resultó en un tiempo

total de ejecución de 5 horas, 45 minutos y 52 segundos. Durante la fase de entrenamiento, se registró un error cuadrático medio promedio de 0.568. En contraste, la fase de prueba presentó un promedio de 1.245 en esta métrica. Estos resultados sugieren una mejora significativa en el rendimiento del modelo.

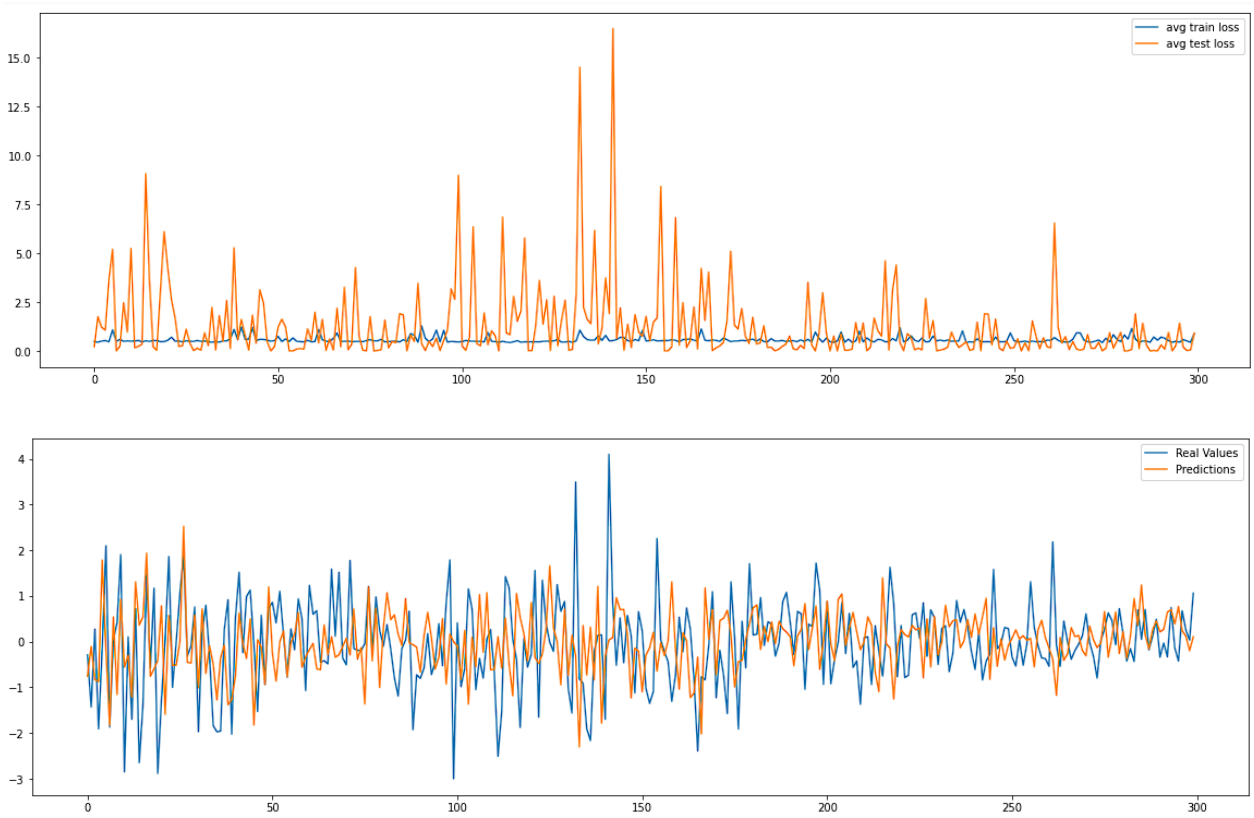

Fig. 24 MSE train y test - Cambio logarítmico del precio (YF - GDP - FRED)

La observación detallada de la gráfica revela la ausencia del pico identificado en el caso anterior, lo cual sugiere que un incremento en la profundidad del aprendizaje ha contribuido a una mejora sustancial. Este fenómeno subraya la importancia de implementar modelos de mayor complejidad, con el fin de capturar de manera más efectiva las características intrínsecas de los datos.

Simulación

Para realizar la simulación de trading se utilizó la misma configuración que en el caso anterior, es decir, el umbral de compra en medio desvío estándar por encima de la media, y el de venta a medio desvío estándar por debajo.

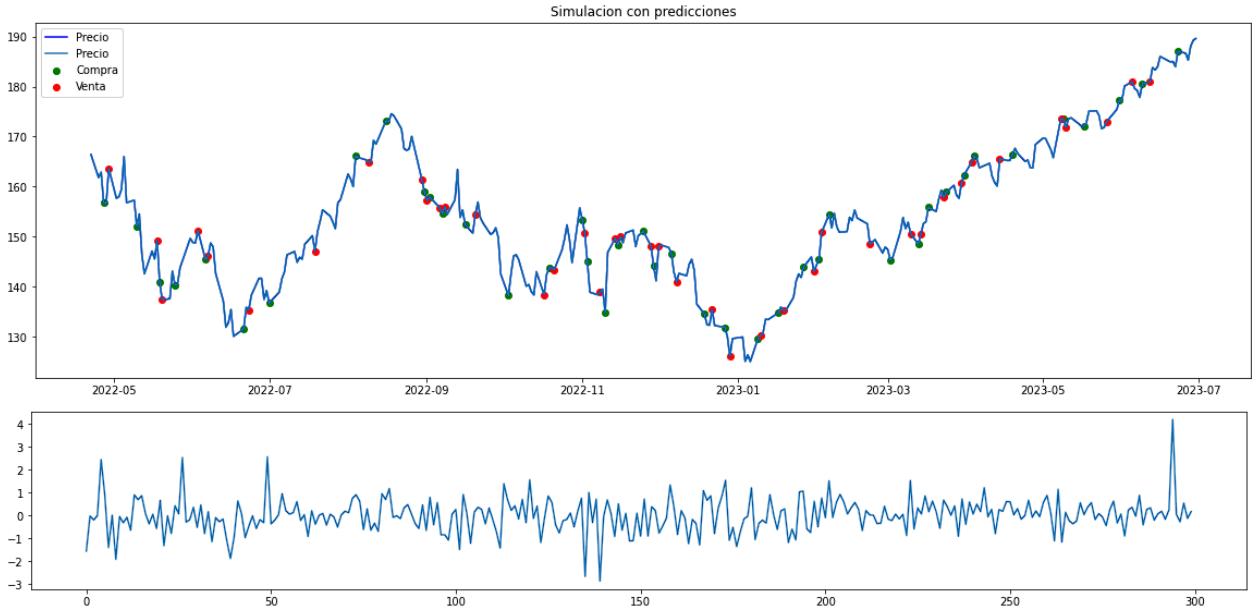

Fig. 25 Simulación Trading - Cambio logarítmico del precio (YF - GDP - FRED)

En el contexto del primer ensayo con estos datos, se alcanzó un retorno del 25.23% durante un periodo de 299 días. Este resultado, aunque inferior al rendimiento previamente obtenido, aún constituye un retorno considerablemente positivo. La posibilidad de optimizar el desempeño mediante la modificación de los umbrales se presenta como una estrategia viable, resaltando la necesidad de calibrar estos umbrales de acuerdo con las condiciones específicas de cada escenario. Sin embargo, es importante considerar que estos resultados podrían estar influenciados por una capacidad de aprendizaje limitada del modelo, como se refleja en el aumento del error cuadrático medio. Este indicador sugiere que el modelo podría estar enfrentando desafíos en la captura y procesamiento eficiente de los patrones subyacentes en los datos.

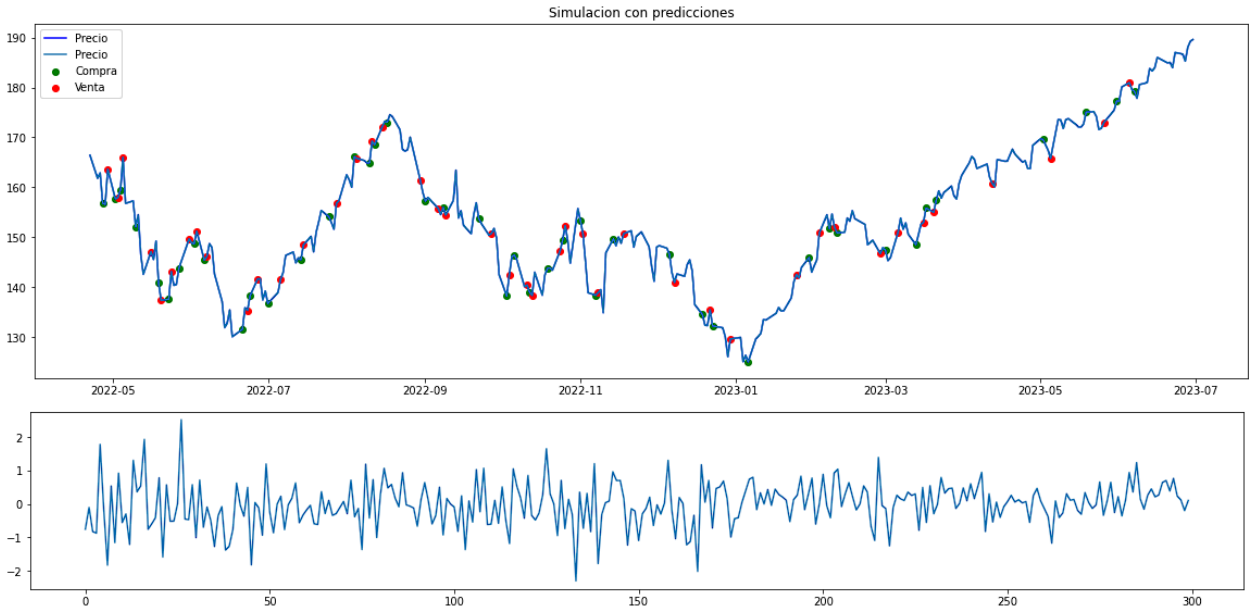

Fig. 26 Simulación Trading - Cambio logarítmico del precio (YF - GDP - FRED)

En el segundo ensayo, utilizando los mismos umbrales que en el primero, se logró un retorno significativamente mayor, específicamente un 42.08% en un periodo de 299 días. Este resultado notoriamente superior subraya la eficacia de incrementar la dimensión del modelo. La mejora en el rendimiento, evidenciada por este aumento en el retorno, sugiere que una mayor complejidad o capacidad del modelo puede ser clave para captar de manera más efectiva las sutilezas y variaciones en los datos

8.3. Yahoo Finance - GDP - FRED - SEC

En la última instancia, se realizó una integración completa de datos, incluyendo indicadores fundamentales proporcionados por la Comisión de Bolsa y Valores (SEC), que aportan información valiosa obtenida de los reportes empresariales. Además, para abordar adecuadamente la complejidad y el volumen aumentado de estos datos, se optó por expandir las dimensiones del modelo. Asimismo, fue imprescindible disminuir el tamaño del lote de procesamiento de 32 a 24, debido a limitaciones del hardware que provocaba interrupciones en el funcionamiento.

Configuración

La configuración utilizada en este caso fue la siguiente:

- Dataset:
    - Cantidad total de datos: 1000 días
    - Dimensión de cada patrón de entrada: 100 días
    - Dimensión de cada variable objetivo: 1 día
    - Paso de ventana: 1 día
- Modelo:
    - Arquitectura: Modelo Ensamblado (LSTM + 3 Capas densas)
    - Tamaño capa oculta LSTM: 500
    - Tamaño capas densas: 500
    - Función de pérdida: Error Cuadrático Medio (MSE).
    - Optimizador: Adam con tasa de aprendizaje de 0.001.
- Entrenamiento
    - Número de épocas: 80.
    - Tamaño del lote: 24.
    - Validación: Ventana expansiva con 700 datos iniciales de entrenamiento, 300 días de test, horizonte de predicción de 1 día y período de expansión de 1 día.

Resultados

El tiempo total empleado para completar el proceso de entrenamiento y validación fue de 12 horas 5 minutos y 48 segundos. Durante la fase de entrenamiento, el modelo alcanzó un error cuadrático medio promedio de 0.793, mientras que en el test se incrementó a 1.715.

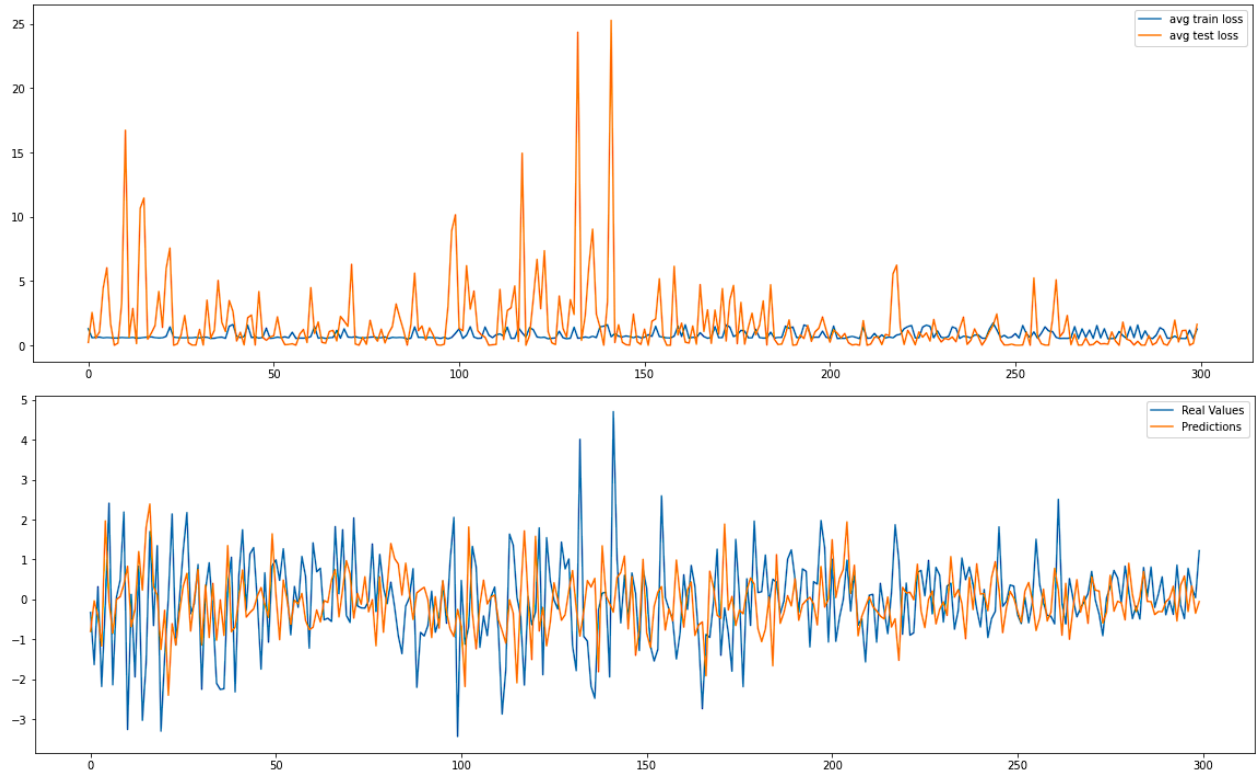

Fig. 27 MSE train y test - Cambio logarítmico del precio (YF - GDP - FRED - SEC)

Simulación

Para realizar la simulación de trading se utilizó la misma configuración que en el caso anterior, es decir, el umbral de compra en medio desvío estándar por encima de la media, y el de venta a medio desvío estándar por debajo.

Bajo estas configuraciones, se alcanzó un retorno del 28.54% en un periodo de 299 días. Este resultado, aunque no alcanza el rendimiento superior observado en la segunda configuración del caso previo, sí supera al de la primera configuración del caso 1. Una estrategia potencial para mejorar estos resultados podría ser la experimentación con modelos de mayor dimensión, tal como se observó en el caso anterior, donde el incremento en la dimensión condujo a un mejor rendimiento. No obstante, se debe considerar que los tiempos de entrenamiento, que en este caso ascienden a 12 horas, comienzan a representar una restricción significativa dada la capacidad del hardware utilizado. Este factor limitante plantea un desafío en el

equilibrio entre la búsqueda de un rendimiento óptimo y la viabilidad práctica en términos de recursos y tiempo de procesamiento.

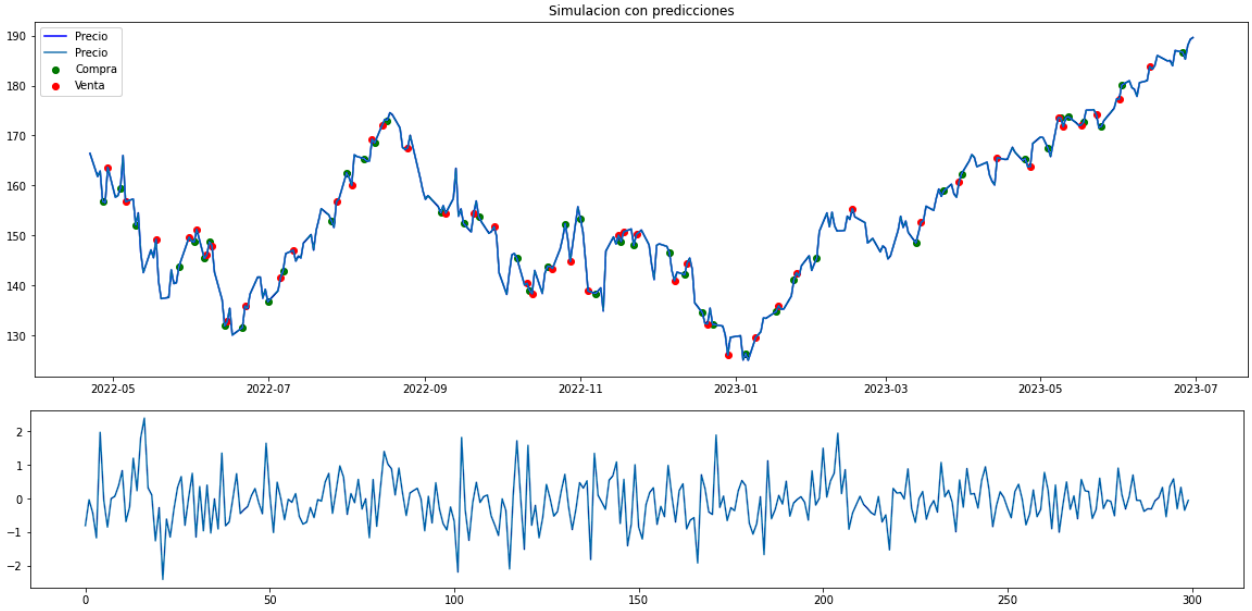

Fig. 28 Simulación Trading - Cambio logarítmico del precio (YF - GDP - FRED - SEC)

9. Conclusiones

El estudio realizado ha demostrado que el uso del Aprendizaje Automático en el mercado financiero es una herramienta de gran valor, ofreciendo oportunidades significativas para mejorar la toma de decisiones y la anticipación de tendencias del mercado. El modelo superó con creces el objetivo establecido, al ofrecer tasas de retorno superiores a la inflación de EEUU, que en el último año (2022) fue del 8% [34]. Sin embargo, es crucial reconocer y abordar las limitaciones inherentes a estos modelos para asegurar su efectividad y confiabilidad.

9.1. Evaluación de Modelos y Tecnologías Utilizadas

Durante el desarrollo del proyecto, se optó por un entorno local con GPU para ejecutar los modelos. A pesar de enfrentar problemas iniciales de compatibilidad, este enfoque resultó beneficioso, ya que brindó la capacidad computacional necesaria para entrenar modelos de complejidad moderada sin incurrir en los costos de servicios en la nube. De otro modo, habría sido necesario reducir tanto la complejidad de los modelos como la cantidad de datos procesados. Esto permitió realizar pruebas que demostraron el potencial del modelo para su futura aplicación en un producto.

No obstante, es importante destacar que para llevar en un futuro este sistema a un producto orientado al consumidor final, debe poder ejecutarse en cualquier PC, por lo que sería necesario migrar la ejecución de los modelos a servicios en la nube. Esto permitiría acceder a una base centralizada con las redes ya entrenadas, optimizando tiempos y recursos de procesamiento. Cabe destacar que los servicios en la nube son costosos, lo cual representa una limitación importante en términos de viabilidad económica.

En cuanto a los modelos utilizados, la arquitectura elegida demostró ser acertada, ya que los resultados obtenidos fueron satisfactorios en términos de precisión de las predicciones. Sin embargo, como se mencionó anteriormente, las exigencias de hardware para ejecutar estos modelos son factores a considerar como también la disponibilidad de datos recientes.

Una alternativa viable sería entrenar una red LSTM exclusivamente con datos de precios y variables diarias para generar señales de compra o venta, ya que estos datos suelen estar disponibles de forma instantánea. Los demás datos podrían aprovecharse para obtener insights valiosos mediante análisis de datos. Esta propuesta podría materializarse en un panel visual interactivo que facilite la toma de decisiones del usuario, aunque para implementarla sería necesaria la colaboración de profesionales con experiencia en finanzas.

9.2. Limitaciones

Una de las principales limitaciones identificadas es la no inclusión de comisiones en las simulaciones. Esto sugiere que los resultados obtenidos no reflejan completamente la rentabilidad real, ya que las comisiones pueden tener un impacto significativo en los rendimientos. Además, la demora en la ejecución de las operaciones, un factor no considerado en el modelo, puede afectar la efectividad de las estrategias propuestas, dado que el mercado financiero es altamente sensible al tiempo.

Otra consideración importante es la dificultad de los modelos para predecir cambios bruscos en el mercado. Estos eventos, a menudo impulsados por factores externos impredecibles o por la naturaleza volátil del mercado, pueden provocar desviaciones significativas entre las predicciones del modelo y los resultados reales. Esta limitación subraya la importancia de no depender exclusivamente de las salidas del modelo para tomar decisiones financieras.

9.3. Logro de objetivos

En cuanto a las metas planteadas en el anteproyecto se cumple el objetivo principal de desarrollar un sistema que predice las tendencias futuras de precios en el mercado financiero. Pero se debe destacar algunas limitaciones o ausencias que se pueden considerar como trabajos futuros.

Una de las principales ausencias es la incorporación de información proveniente de noticias y redes sociales. Estos datos son fundamentales porque pueden reflejar el sentimiento del mercado y el impacto de eventos externos en tiempo real, lo que enriquecería las predicciones del modelo. Inicialmente, se planeó utilizar la plataforma X para obtener dicha información, pero fue descartada debido a las restricciones de su API, que requiere un pago por acceso. A futuro, sería recomendable investigar fuentes confiables y gratuitas que permitan integrar este tipo de datos en el modelo, con el fin de mejorar la precisión de las predicciones.

En cuanto al objetivo de implementar un módulo para ejecutar un análisis de los fundamentos de un activo o empresa utilizando técnicas de análisis de datos, se puede considerar que fue parcialmente alcanzado. El modelo recibe datos fundamentales como parte de su entrenamiento, y varios de estos datos fueron preprocesados para generar variables adicionales que se utilizaron en el modelo predictivo. Sin embargo, este módulo no proporciona una puntuación o evaluación específica de los fundamentos de la empresa, ya que forma parte de una subarquitectura dentro de la red neuronal que contribuye al resultado final de las predicciones.

Respecto al objetivo de evaluar el sistema en un entorno de simulación de trading, fue cumplido de manera parcial. Se ejecutaron simulaciones de compras y ventas basadas en las predicciones del sistema, y se calculó la rentabilidad obtenida al final del proceso. Esta estrategia permitió estimar una rentabilidad porcentual pero utilizando datos históricos. Por lo que el modelo no fue evaluado en un entorno de simulación en tiempo real, lo que limita la capacidad de replicar condiciones del mercado en vivo.

9.4. Trabajos futuros

El presente proyecto ha sentado una base para el desarrollo y la aplicación del Machine Learning en el análisis del mercado financiero. Sin embargo, dado el carácter iterativo y evolutivo de los proyectos de Machine Learning, hay varias áreas clave para la expansión y mejora futuras.

- Operatividad en tiempo real: Un paso crucial será adaptar y probar el modelo en un entorno de mercado real. Esto no solo implica actualizar y expandir constantemente la base de datos para reflejar las condiciones actuales del mercado, sino también reentrenar el modelo regularmente para mantener su precisión y relevancia.
- Optimización de Hiperparámetros: Es esencial implementar mecanismos de optimización de hiperparámetros que se adapten específicamente al activo financiero a predecir, considerando factores como su volatilidad. Este enfoque personalizado mejorará la precisión y la eficacia del modelo para diferentes tipos de activos.
- Balance entre Precisión y Eficiencia Temporal: Un desafío importante consiste en encontrar el equilibrio óptimo entre la precisión de las predicciones del modelo y el tiempo requerido para su entrenamiento. Este balance es clave para maximizar las ganancias debido a las fluctuaciones de precio durante el tiempo que el modelo tarda en realizar sus predicciones.
- Incorporación de Análisis de Sentimiento de Redes Sociales: Dada la influencia de las redes sociales en los mercados financieros, se plantea la incorporación de análisis de sentimiento basado en datos de estas plataformas. Esto podría proporcionar insights valiosos sobre las tendencias del mercado y la percepción pública, lo que podría mejorar significativamente la precisión de las predicciones del modelo.
- Ampliar Conjuntos de Datos: Un paso importante será explorar la posibilidad de realizar un preentrenamiento más exhaustivo del modelo utilizando un conjunto de datos más amplio. En este proyecto, se limitó a los últimos 1000 puntos de datos debido a restricciones de capacidad de cómputo. Sin embargo, utilizar un conjunto de datos más extenso podría mejorar significativamente la capacidad del modelo para capturar y aprender de las tendencias a largo plazo. Esto puede incluir utilizar datos de múltiples empresas.
- Desarrollo de Interfaz Gráfica: Para aumentar la accesibilidad y utilidad del modelo, es necesario diseñar una interfaz gráfica intuitiva. Esta interfaz facilitaría

a los usuarios con menos conocimientos técnicos, el acceso a las predicciones y análisis del modelo, mejorando así su aplicabilidad práctica.

Estas mejoras propuestas representan pasos esenciales hacia la optimización y eficacia del modelo desarrollado. También son parte del proceso necesario para llegar a un producto final que sea más adaptable y útil para una amplia variedad de aplicaciones en el dinámico mundo de las inversiones financieras, permitiendo que el modelo evolucione hacia una solución robusta y accesible para los usuarios.

Referencias

[1] Hayed, A. (2023). Dow Theory Explained: What It Is and How It Works. Investopedia https://www.investopedia.com/terms/d/dowtheory.asp

[2] Graham, B., Dodd, D. L. (1934). Security Analysis: The Classic 1934 Edition. New York: McGraw-Hill.

[3] Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance.

[4] Caballar, R., Stryker C. (2024) What is forecasting?. IBM https://www.ibm.com/think/topics/forecasting

[5] Amazon Web Services. (2024). What is deep learning?. https://aws.amazon.com/what-is/deep-learning/

[6] Logunova, I. (2023). Time Series Analysis in ML. Serokell https://serokell.io/blog/time-series-analysis-in-ml

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Capítulo 6: Deep Feedforward Networks. MIT Press. https://www.deeplearningbook.org/

[8] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Capítulo 9: Convolutional Networks. MIT Press. https://www.deeplearningbook.org/

[9] Cao, J., & Wang, J. (2019). Stock price forecasting model based on modified convolution neural network and financial time series analysis. International Journal of Communication Systems.

[10] Stewart, K. (2024). mean squared error. Britannica https://www.britannica.com/science/mean-squared-error

[11] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Capítulo 10: Sequence Modeling: Recurrent and Recursive Nets. MIT Press.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need.

[13] Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2022). Are Transformers Effective for Time Series Forecasting? The Chinese University of Hong Kong; International Digital Economy Academy (IDEA)

[14] Kamalov, F., Smail, L., & Gurrib, I. (2020). Stock price forecast with deep learning. Canadian University Dubai

[15] Zhu, Y. (October 2020). Stock price prediction using the RNN model. Journal of Physics Conference Series.

[16] De Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Capítulo 10.10: Sequence Modeling: Recurrent and Recursive Nets. MIT Press.
https://www.deeplearningbook.org/

[17] Datosmacro (2024). PIB.
https://datosmacro.expansion.com/diccionario/pib

[18] Maunsell, T. (2024) How Does GDP Affect the Stock Market?. Blueberry
https://blueberrymarkets.com/en/market-analysis/how-does-gdp-affect-the-stock-market/

[19] Zucchi, K. (2023). Inflation's Impact on Stock Returns. Investopedia.
https://www.investopedia.com/articles/investing/052913/inflations-impact-stock-returns.asp

[20] Hall, M. (2024). How Do Interest Rates Affect the Stock Market? Investopedia.
https://www.investopedia.com/investing/how-interest-rates-affect-stock-market/

[21] Plus500. (2024). Qué es el Índice de Confianza del Consumidor (ICC). Plus500.
https://www.plus500.com/es/newsandmarketinsights/consumer-confidence-index-explained

[22] Segal, T (2024). Fundamental Analysis: Principles, Types, and How to Use It. Investopedia
https://www.investopedia.com/terms/f/fundamentalanalysis.asp

[23] Nau, R. (2020). The logarithm transformation. Duke University.
https://people.duke.edu/~rnau/411log.htm#changelog

[24] Wertz, R. (2023). What Is Gaussian Distribution In Machine Learning. Robots.Net.
https://robots.net/fintech/what-is-gaussian-distribution-in-machine-learning/

[25] Statisticseasily. (2024). Qué es la transformación logarítmica.
https://es.statisticseasily.com/glossario/what-is-log-transform/

[26] Mailagaha Kumbure, M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review.

[27] GeeksforGeeks. (2024). Python for Machine Learning.
https://www.geeksforgeeks.org/python-for-machine-learning/

[28] Yazar, K. (2024). What is PyTorch?. Techtarget
https://www.techtarget.com/searchenterpriseai/definition/PyTorch

[29] Datascientest. (2022). Pandas : La biblioteca de Python dedicada a la Data Science.
https://datascientest.com/es/pandas-python

[30] Chen, J. (2024). Normal Distribution: What It Is, Uses, and Formula.
https://www.investopedia.com/terms/n/normaldistribution.asp

[31] Khoong, W. (2023). Why Scaling Your Data Is Important.
https://medium.com/codex/why-scaling-your-data-is-important-1aff95ca97a2

[32] Nevil, S. (2024). Z-Score: Meaning and Formula.
https://www.investopedia.com/terms/z/zscore.asp

[33] germayne. (2019). Time Series Cross-validation — a walk forward approach in python.
https://medium.com/eatpredlove/time-series-cross-validation-a-walk-forward-approach-in-python-8534dd1db51a

[34] CPI Inflation calculator. (2023). 2022 CPI and Inflation Rate for the United States.
https://cpiinflationcalculator.com/2022-cpi-and-inflation-rate-for-the-united-states/

Anexo I: Otras resultados

1. Predicción cambio logarítmico del precio de APPLE

Datos en temporalidad diaria utilizando el cambio logarítmico del precio como variable de entrada.

Configuracion

- Dataset:
    - Cantidad total de datos: 1000 días
    - Dimensión de cada patrón de entrada: 91 días
    - Dimensión de cada variable objetivo: 1 día
    - Paso de ventana: 1 día
- Modelo:
    - Arquitectura: Modelo LSTM con 2 capas
    - Tamaño capa oculta LSTM: 150
    - Función de pérdida: Error Cuadrático Medio (MSE).
    - Optimizador: Adam con tasa de aprendizaje de 0.001.
- Entrenamiento
    - Número de épocas: 120.
    - Tamaño del lote: 64.
    - Validación: Ventana expansiva con 300 datos iniciales de entrenamiento, 700 días de test, horizonte de predicción de 1 día y período de expansión de 1 día.

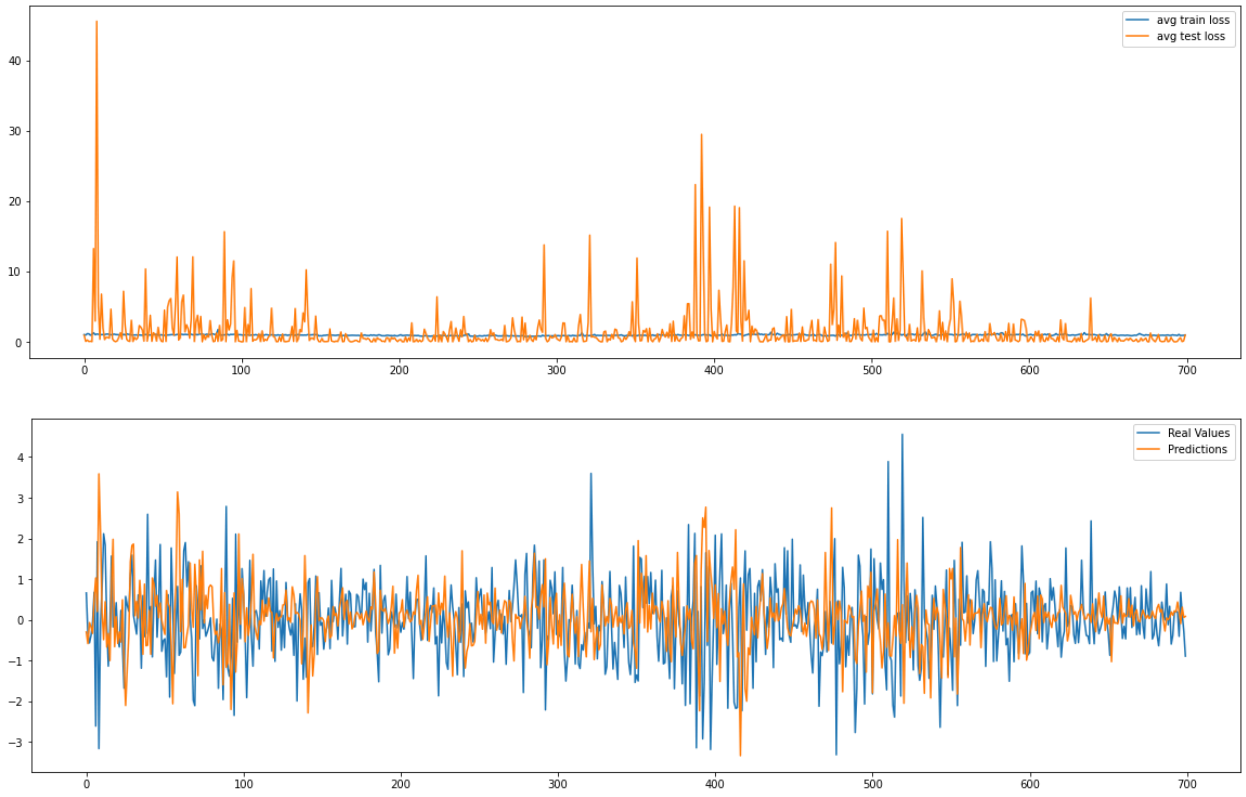

Fig. 29 Resultados AAPL

Simulación:

Umbral de compra en 1 desvío estándar por encima de la media, y el de venta a 1 desvío estándar por debajo de la media.

Retorno: 77.97% en 700 días

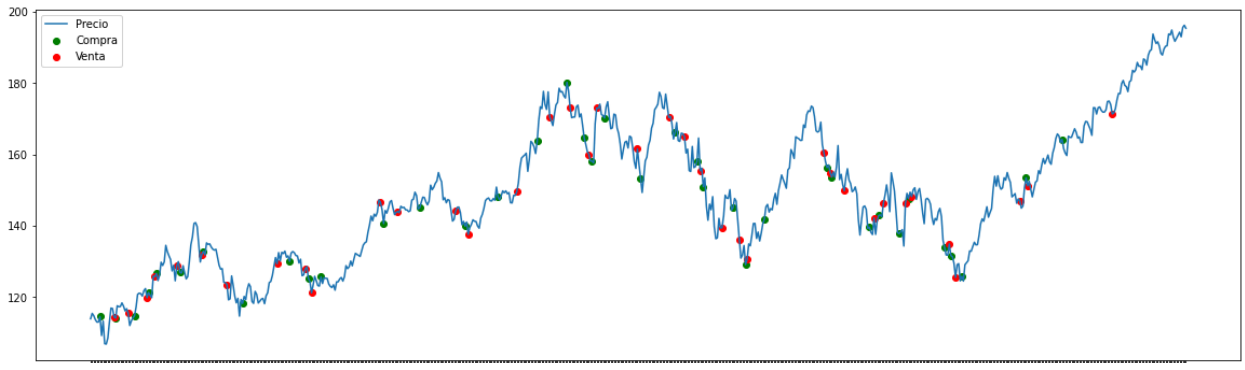

Fig. 30 Simulacion AAPL

2. Predicción de cambio logarítmico de precio de SP&500

Datos en temporalidad trimestral usando como patrón de entrada GDP por industria.

Configuración

- Dataset:
    - Cantidad total de datos: 72 trimestres
    - Dimensión de cada patrón de entrada: 12 trimestres
    - Dimensión de cada variable objetivo: 1 trimestre
    - Paso de ventana: 1 trimestre
- Modelo:
    - Arquitectura: Modelo LSTM con 2 capas
    - Tamaño capa oculta LSTM: 1000
    - Función de pérdida: Error Cuadrático Medio (MSE).
    - Optimizador: Adam con tasa de aprendizaje de 0.001.
- Entrenamiento
    - Número de épocas: 150.
    - Tamaño del lote: 64.
    - Validación: Ventana expansiva con 12 datos iniciales de entrenamiento, 47 trimestres de test, horizonte de predicción de 1 trimestre y período de expansión de 1 trimestre.

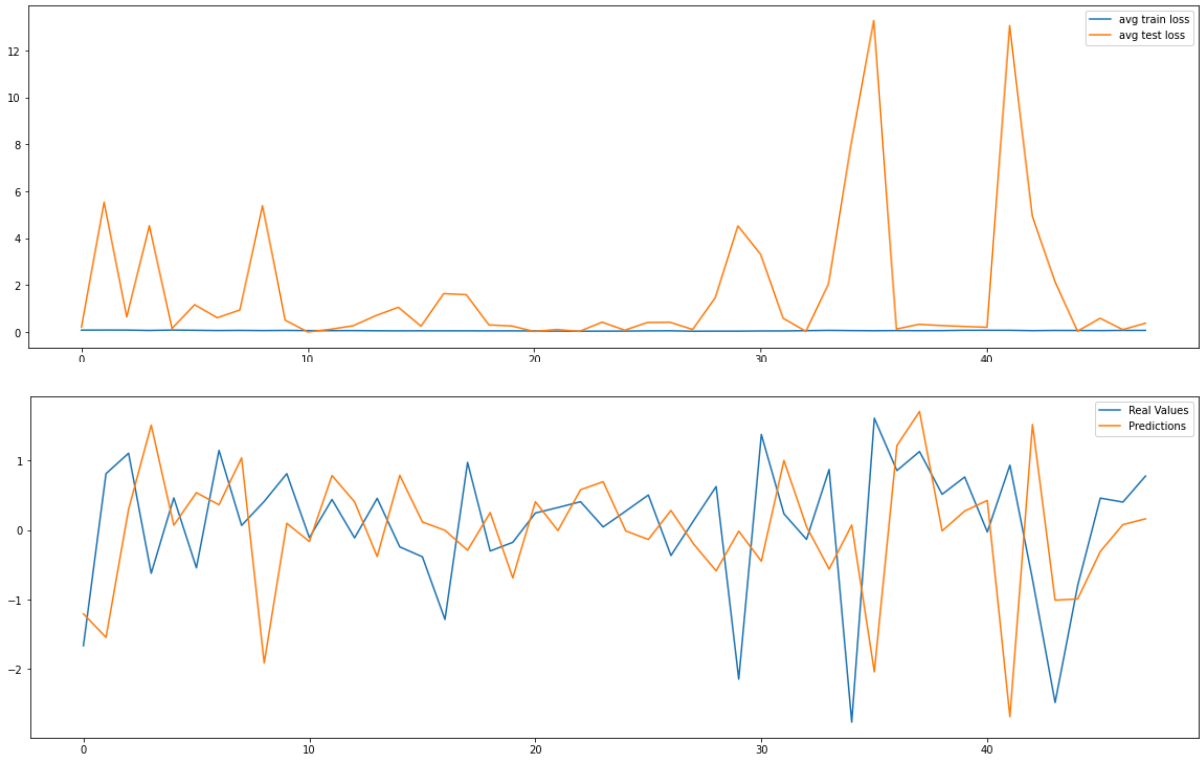

Fig. 31 Resultados S&P 500 Trimestral

Train loss: 0.0604, Test loss: 1.7341

Simulación:

Umbral de compra en medio desvío estándar por encima de la media de las predicciones, y el de venta a medio desvío estándar por debajo de la media.

Retorno: 119.39% en 47 trimestres/4288 dias/11.75 anios

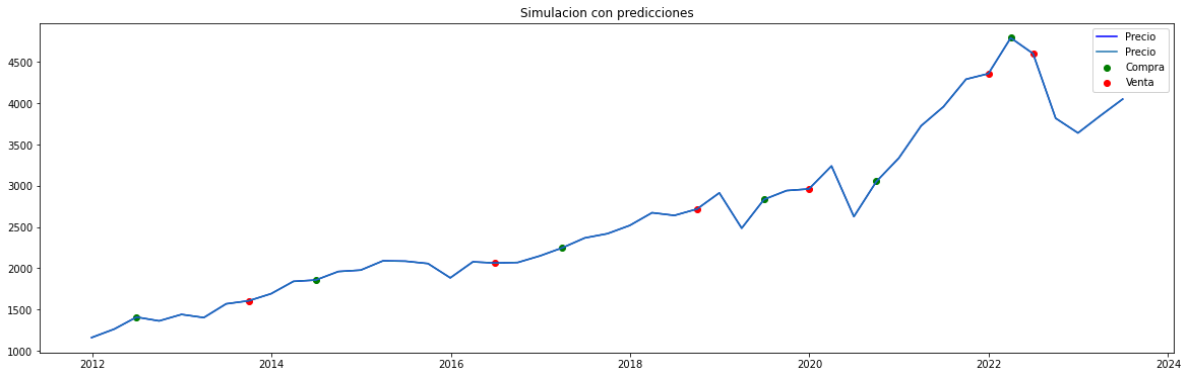

Fig. 32 Simulacion S&P500 Trimestral

Anexo II: Anteproyecto

UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas

PROPUESTA DE PROYECTO FINAL DE CARRERA

INGENIERÍA INFORMÁTICA

DESARROLLO DE UN SISTEMA INTELIGENTE PARA LA TOMA DE DECISIONES DE INVERSIÓN EN EL MERCADO FINANCIERO UTILIZANDO HERRAMIENTAS DE MACHINE LEARNING

Alumno: Bartezaghi, Catriel

Director: Robledo, Miguel Angel

Santa Fe, Abril de 2023

Resumen

El objetivo de este proyecto consiste en el desarrollo de un sistema inteligente que sea capaz de proporcionar recomendaciones personalizadas de inversión en activos financieros, con la finalidad de mejorar la capacidad de los inversores para identificar patrones y tendencias en el mercado financiero.

Para ello se realizará un análisis de datos relevantes que influyen en el precio de los activos mediante el uso de herramientas de machine learning. Además, se realizarán pruebas y validaciones de los modelos en una plataforma de simulación de trading para evaluar su desempeño en condiciones reales del mercado.

Palabras claves

Sistema inteligente, inversión, análisis de datos, machine learning, activos financieros.

Justificación

El mercado financiero es uno de los mercados más grandes y complejos del mundo, y las decisiones de inversión tienen un impacto significativo en la economía global y en las finanzas personales de los individuos. Por lo tanto, la necesidad de contar con información precisa y actualizada para tomar decisiones de inversión es crucial.

Este mercado se encuentra en constante evolución y cada vez es mayor el número de personas interesadas en invertir en activos financieros para obtener rentabilidad de sus ahorros. En el caso de Estados Unidos, el número de ciudadanos que posee acciones es tan alto que llega al 58% de la población [1].

Además, en la mayoría de países de Europa y América, se ha observado un aumento en los índices de inflación en los últimos años [2], lo que ha llevado a la necesidad de buscar alternativas de inversión que permitan preservar el valor de los ahorros de las personas. Pero la relación entre la inflación y las acciones es compleja y depende de las características propias de cada acción [3], por ello realizar un análisis correcto de inversión resulta en una tarea desafiante para personas inexpertas en la materia.

Por otra parte, el componente tecnológico ha revolucionado la forma en que se opera en el mercado financiero, a tal punto que algunas plataformas reportan que entre el 70% y 80% de las acciones negociadas provienen de sistemas automáticos de negociación [4]. Esto permite tomar decisiones de inversión de forma más rápida y precisa, y tiene un impacto significativo en la competitividad del mercado.

En el ámbito del análisis de activos financieros, existen diversas metodologías que buscan evaluar el valor y desempeño de los mismos. Dos de las metodologías más utilizadas históricamente son el análisis fundamental [5] y el análisis técnico [6]. El análisis fundamental pretende determinar el valor auténtico o intrínseco del activo a partir de las distintas variables que influyen en su precio. Esto supone un indicador del rendimiento futuro que se espera del activo, y se basa en el estudio detallado de factores económicos, financieros y de mercado que afectan el valor de los activos, tales como las condiciones macroeconómicas, la posición competitiva de la empresa, su situación financiera y sus perspectivas de crecimiento.

Por otro lado, se encuentra el análisis técnico, el cual se centra en el análisis del precio del activo y el volumen negociado, además de los indicadores que surgen de estos, con el propósito de anticipar con mayor probabilidad cambios en la estructura del mercado. Su objetivo principal es identificar patrones y tendencias en el movimiento de los precios, con el fin de predecir alteraciones en la dinámica del mercado.

Las desventajas de estas dos metodologías radican en que, al realizarlas de forma clásica mediante inspección o análisis de gráficos de precios, los inversores pueden verse limitados por su capacidad para analizar grandes cantidades de datos, lo que dificulta la toma decisiones de inversión y el no aprovechamiento de todos los datos disponibles.
Además, el inversor puede verse influenciado por sus propias emociones y percepciones, esto puede llevar a decisiones de inversión sesgadas basadas en juicios personales y no en datos objetivos.

Por estas razones, existe una creciente demanda por parte de los operadores de mercado de contar con herramientas tecnológicas que les permitan tomar decisiones de inversión más informadas y acertadas en un mercado financiero cada vez más complejo [7]. En este contexto, las herramientas de aprendizaje automático para el análisis bursátil han surgido como una solución prometedora, ya que permiten procesar grandes volúmenes de datos de manera eficiente y proporcionan información valiosa para la toma de decisiones de inversión. De esta manera, el uso de estas herramientas puede mejorar significativamente la capacidad de los inversores para identificar patrones y tendencias en el mercado financiero, lo cual puede traducirse en una mayor rentabilidad y una reducción de los riesgos asociados a la inversión.

También, el empleo de recursos de aprendizaje automático para el análisis bursátil puede democratizar el acceso a la información y la toma de decisiones de inversión, esto los hace de utilidad tanto para inversores de distintos niveles de experiencia, así como para instituciones financieras como los fondos de inversión.

El aporte que presenta el presente proyecto consiste en proveer información al inversor desde tres aspectos claves en el análisis de un activo financiero: técnico, fundamental y social. Esta información se obtendrá a través de la implementación de tres módulos de

análisis, los cuales utilizarán técnicas avanzadas de machine learning. En cuanto al módulo del aspecto social del mercado, en este caso se centra en realizar análisis de los sentimientos [8] de textos u opiniones asociadas a los activos de interés. De esta manera, se podrá lograr una visión integral y precisa que ayude a los inversores a tomar decisiones informadas en sus operaciones.

La realización del presente proyecto también tendrá diversos aportes para el autor en términos de su crecimiento profesional y personal. Por un lado, tendrá la oportunidad de aplicar conocimientos teóricos adquiridos en su formación académica en un proyecto práctico y real, el cual le permitirá afianzar su comprensión y habilidades técnicas, así como también su capacidad para la resolución de problemas y toma de decisiones.

Por otro lado, el proyecto permitirá al autor adquirir y mejorar sus habilidades en el campo del machine learning y análisis de datos, áreas en las que tiene un gran interés y que son muy demandadas en el mercado laboral actual.

Objetivos

Objetivos Generales

- Desarrollar un sistema de aprendizaje automático que, utilizando datos financieros, técnicos y de análisis de sentimiento, sea capaz de analizar y predecir la evolución futura de activos financieros, con el objetivo de ayudar al inversor en la toma de decisiones.

Objetivos Específicos

- Determinar los modelos de machine learning a utilizar para cada módulo.
- Implementar un módulo que realice el análisis del precio histórico de un activo empleando redes neuronales.
- Implementar un módulo que ejecute un análisis de los fundamentos de un activo o empresa utilizando técnicas de análisis de datos.

- Desarrollar un módulo que pueda identificar el tono emocional de un conjunto de textos, como tweets o titulares de noticias.
- Implementar un método de adquisición de los datos necesarios para los 3 módulos mencionados.
- Recopilar y limpiar los datos necesarios para el entrenamiento de los modelos.
- Implementar un método de almacenamiento de los datos históricos utilizados para el análisis y los resultados obtenidos por cada módulo de análisis.
- Evaluar los modelos para determinar su precisión y eficacia.
- Implementar y evaluar los modelos en un entorno de simulación de trading.

Alcances

El producto final del presente proyecto será el desarrollo de tres módulos que utilizarán datos de precios históricos, datos fundamentales de la empresa o activo y datos de análisis de sentimiento relevantes para analizar y predecir el comportamiento del mercado a futuro.
En el caso de los módulos que serán entrenados a partir de precios históricos y fundamentales, se llevará a cabo la selección de los modelos adecuados para cada caso, evaluando su arquitectura e hiperparámetros.

La obtención de los datos se realizará mediante interfaces de programación de aplicaciones (APIs) [9] de sitios web externos. También se realizará una búsqueda exhaustiva para considerar otras fuentes de información relevantes.

En relación con el módulo de análisis de sentimiento, se obtendrá información reciente de redes sociales como Twitter y de titulares de noticias proporcionados por motores de búsqueda. Luego, se utilizarán modelos preentrenados para determinar emociones generales que la sociedad y los inversores tienen hacia los activos financieros de interés.

La salida de todos los modelos será una puntuación en una escala del 0 al 10, que indicará el grado de negatividad o positividad del mercado en relación con el análisis

efectuado, sin descartar la posibilidad de otras salidas como la evolución del precio a futuro.

Para almacenar los datos obtenidos a partir de las APIs y otras fuentes, se empleará una base de datos. La información se almacenará de manera estructurada para facilitar su uso y análisis posterior.

El principal propósito del proyecto es brindar un sistema que sirva de apoyo en la toma de decisiones de inversión de diferentes activos. Es el inversor quien definirá su estrategia de operación a partir de las salidas informativas que proporcionan los módulos de análisis. De todas maneras, los módulos serán desplegados en un entorno de simulación de mercado para evaluar su desempeño al operar de forma automática.
Para las pruebas de simulación de trading, se realizará la conexión con un entorno que replique las condiciones y características del mercado financiero, y se efectuarán operaciones a partir de las predicciones de los modelos entrenados.

Entre las restricciones del proyecto se debe mencionar que la calidad y cantidad de los datos disponibles para el análisis pueden ser limitadas o incompletas, lo que podría afectar la precisión de las recomendaciones generadas. También es importante tener en cuenta que puede haber limitaciones tecnológicas que impidan entrenar una arquitectura de red neuronal compleja que satisfaga las necesidades del problema, debido a la capacidad insuficiente de cómputo disponible.

Por último, para evaluar la eficacia del sistema, se llevarán a cabo simulaciones de compras y ventas durante un período posterior a los datos de entrenamiento, a partir de las salidas de los módulos. Se espera que en la mayoría de los casos el modelo logre un retorno superior a la tasa de inflación anual del último año en EEUU como indicador de éxito.

Metodología

Dado que el objetivo del proyecto implica un análisis utilizando técnicas de aprendizaje automático, la metodología más adecuada es la basada en el ciclo de vida de un proyecto de machine learning. Este enfoque de trabajo consiste en una metodología

que cuenta con etapas iterativas y que incluye pasos específicos para este tipo de proyectos. Algunos de estos pasos incluyen la preparación de datos, selección y evaluación de modelos, entre otros, los cuales son esenciales para garantizar un proceso riguroso y eficiente para la construcción de modelos precisos y útiles.

La principal ventaja de esta metodología es que permite realizar iteraciones de manera gradual para mejorar el modelo progresivamente. Esto es fundamental en un proyecto de machine learning, debido a que la evaluación del modelo es una fase crucial y continua a lo largo del proyecto. Además, en el entorno de análisis financiero donde las condiciones del mercado pueden cambiar drásticamente, es necesario considerar estas variaciones y ajustar los modelos para obtener resultados precisos.

Por último, esta metodología hace hincapié en la interpretación de los resultados de los modelos, esto permite obtener un mayor conocimiento del problema abordado e identificar posibles mejoras en etapas anteriores.

Las etapas en que se desarrollará el proyecto son:

1. Análisis del problema.
2. Recopilación de datos.
3. Procesamiento de datos.
4. Selección de modelos.
5. Entrenamiento y ajuste de modelos.
6. Puesta en producción.

Plan de tareas

Se presenta un plan detallado de las actividades específicas que se realizarán para completar el proyecto. Las actividades serán realizadas por el autor del proyecto. Se estima una carga horaria de 20 horas semanales distribuidas entre los días lunes y viernes. Para las estimaciones de tiempo de cada actividad se utilizará el juicio de expertos.

Etapa 1: Análisis del problema

1. Investigación del problema y contexto del proyecto (12 horas).
2. Identificación de los objetivos y requisitos del proyecto (8 horas).
3. Definición del alcance del proyecto (8 horas).

Hito: Definición del problema y objetivos del proyecto.

Criterio de aceptación: Documento de justificación, objetivos y alcance.

Etapa 2: Recopilación de Datos

1. Búsqueda de datos relevantes para el análisis bursátil. (20 horas)
2. Obtención de los datos necesarios (16 horas).
3. Organización y almacenamiento de los datos (20 horas).

Hito: Datos recolectados y organizados en una base de datos.

Criterio de aceptación: Los datos deben estar almacenados en una base de datos ordenados por categoría. Deben ser legibles y modificables para su posterior procesamiento.

Etapa 3: Procesamiento de datos.

1. Limpieza y preprocesamiento de los datos (24 horas).
2. Análisis exploratorio de datos (12 horas).
3. Selección de características y variables relevantes. (16 horas).
4. Normalización de los datos (12 horas).

Criterio de aceptación: Creación de un conjunto de datos preprocesado y limpio que se haya explorado y analizado a fondo. Los datos deben ser adecuados para ser consumidos por los modelos de machine learning.

Hito: Set de datos limpios con características más relevantes para el problema.

Etapa 4: Selección de modelos.

1. Investigación y evaluación de diferentes modelos (12 horas).
2. Selección de los modelos más adecuados para cada análisis (8 horas).
3. Estudio de los modelos seleccionados (12 horas).
4. Implementación de modelos de prueba (16 horas).

Criterio de aceptación: Identificación de los modelos más adecuados para el proyecto con su correspondiente justificación. Implementación de una arquitectura de prueba para cada modelo y comprobación de su correcto funcionamiento.

Hito: Selección final de los modelos de machine learning que se utilizarán en el proyecto.

Etapa 5: Entrenamiento y ajuste de modelos.

1. Partición de los datos de entrenamiento, test y validación (8 horas).
2. Definición de métricas de evaluación para cada modelo (8 horas).
3. Entrenamiento inicial de los modelos (12 horas).
4. Ajuste de los modelos y optimización de sus hiperparámetros (40 horas).
5. Evaluación del rendimiento de los modelos (12 horas).

Criterio de aceptación: El rendimiento del modelo debe ser evaluado en un conjunto de pruebas independiente para asegurar su capacidad de generalización. Debe cumplir con las métricas de evaluación establecidas en el proyecto, de lo contrario se deben realizar ajustes adicionales.

Hito: Modelos entrenados y ajustados con un rendimiento aceptable.

Etapa 6: Puesta en producción

1. Estudio del entorno de producción (16 horas).

2. Despliegue de los modelos en el entorno de producción (20 horas).
3. Pruebas en un sistema de trading con dinero ficticio (20 horas).
4. Monitoreo y ajuste de los modelos (32 horas).

Criterio de aceptación: Los modelos se han integrado correctamente con la API y se ha desplegado en el entorno de producción. Se han realizado pruebas en el sistema de trading y se han resuelto los problemas identificados. También se ajustaron los modelos de acuerdo al rendimiento obtenido en el entorno de trading.

Hito: Modelo en producción y operativo en el sistema de trading.

Cronograma

El proyecto iniciará el 14 de agosto de 2023 y se espera que finalice el 18 de diciembre de 2023. Se estima un total de 364 horas de trabajo. El cronograma del proyecto se presenta de manera secuencial, esto significa que cada etapa se llevará a cabo en un orden específico. Sin embargo, dentro de cada etapa se pueden realizar iteraciones para mejorar los resultados, siempre y cuando se respete la fecha de finalización establecida.

Es importante destacar que en los proyectos de machine learning el ciclo de vida completo también es iterativo. A medida que se obtienen nuevos resultados, es común que se genere una retroalimentación constante en la que se aprende y se entiende mejor el problema que se está abordando, lo que permite identificar nuevas oportunidades para mejorar el modelo. En este caso, debido a las limitaciones de calendario, se realizará un primer prototipo el cual puede ser mejorado en el futuro con los conceptos adquiridos durante el desarrollo de este proyecto.

En la siguiente página se puede visualizar un diagrama que resume el cronograma del proyecto.

Entregables

- Primer entregable
    - Fecha: 31/08/23
    - Detalle: Se realizará la entrega de un informe que describa los datos a procesar para entrenar los modelos, la fuente y de qué forma serán consumidos.
- Segundo entregable
    - Fecha: 5/10/23
    - Detalle: Se entregará el conjunto de datos limpios y procesados. También un informe de análisis de datos donde se detallen los resultados del análisis exploratorio y la selección de características y variables relevantes.
- Tercer entregable
    - Fecha: 9/11/23
    - Detalle: El entregable para esta fecha consistirá en el código con los modelos implementados. También se proporcionará un informe que describa:
        - Los modelos de machine learning utilizados.
        - Las métricas para evaluar el desempeño de cada modelo.
        - Los resultados obtenidos para cada modelo hasta la fecha.

Gestión de riesgos

A continuación se detallan diferentes riesgos identificados para el proyecto.

ID #	Probabilidad	Impacto
1	Media	Alto

Riesgo: Datos de baja calidad.
Descripción: Los datos utilizados para entrenar el modelo pueden ser incompletos, inexactos o no representativos, lo que puede afectar la precisión de los modelos.
Indicador: Análisis de la consistencia e integridad de los datos.
Respuesta: Mitigar. Realizar una exhaustiva evaluación de la calidad de los datos disponibles y tomar medidas para mejorar o corregir los problemas identificados.
Contingencia: Reducir el número de instrumentos a predecir de acuerdo a la disponibilidad de datos confiables.

ID #	**Probabilidad**	**Impacto**
2	Media	Medio
Riesgo: Sobreajuste del modelo.		
Descripción: El modelo de machine learning puede estar demasiado ajustado a los datos de entrenamiento y no generalizar correctamente en nuevos datos, lo que lleva a predicciones incorrectas con datos de series nunca vistos.		
Indicador: Comparación de rendimiento del modelo con los datos de entrenamiento y con los datos de prueba.		
Respuesta: Mitigar. Utilizar técnicas de regularización, validación cruzada y ajuste de hiperparámetros para evitar el sobreajuste.		

Contingencia: Volver a entrenar los modelos con diferentes configuraciones y/o algoritmos de aprendizaje para mejorar su capacidad de generalización.

ID #	**Probabilidad**	**Impacto**
3	Media	Alto
Riesgo: Limitaciones de capacidad de cómputo.		
Descripción: El proyecto enfrenta restricciones en términos de la capacidad de cómputo disponible para entrenar los modelos de machine learning en tiempo y forma.		
Indicador: Tiempo de entrenamiento del modelo en relación con los plazos y requisitos del proyecto.		
Respuesta: Mitigar. Reducir el volumen de datos utilizados para el entrenamiento y las dimensiones de los modelos.		
Contingencia: Explorar la posibilidad de utilizar modelos pre-entrenados para reducir la carga computacional necesaria.		

ID #	**Probabilidad**	**Impacto**
4	Media	Alto
Riesgo: Mal desempeño de los modelos en condiciones reales.		

Descripción: Existe la posibilidad de que el modelo de machine learning, que funciona bien en un entorno de testing, no tenga un desempeño óptimo en condiciones reales del mercado.
Indicador: Desviación entre los resultados esperados del modelo en el entorno de simulación y los resultados reales obtenidos en el mercado real.
Respuesta: Aceptar activamente
Contingencia: Utilizar estrategias de trading como porcentaje de stop loss o take profit en las órdenes para limitar las pérdidas y mejorar los resultados.

ID #	**Probabilidad**	**Impacto**
5	Media	Medio
Riesgo: Alta latencia y velocidad de ejecución		
Descripción: Existe el riesgo de que los modelos de machine learning no puedan generar decisiones y ejecutar las operaciones de manera rápida y eficiente en un entorno de trading en tiempo real.		
Indicador: Retraso en el tiempo de respuesta del modelo.		
Respuesta: Aceptar activamente.		
Contingencia: Optimizar aquellos componentes del modelo y algoritmos que afectan directamente la latencia y la velocidad de ejecución.		

ID #	**Probabilidad**	**Impacto**

6	Baja	Alto
Riesgo: Ausencia de datos relevantes sobre activos financieros.		
Descripción: Existe la posibilidad de que no se puedan obtener o acceder a los datos necesarios para abordar el problema específico de predicción de precios en activos financieros.		
Indicador: Número de datos recolectados.		
Respuesta: Aceptar activamente.		
Contingencia: Modificar el alcance del proyecto para implementar los módulos que son posibles y redirigir los objetivos de predicción hacia activos con los que existan datos accesibles.		

ID #	**Probabilidad**	**Impacto**
7	Media	Medio
Riesgo: Imposibilidad de obtener información de redes sociales en tiempo real.		
Descripción: El acceso a tweets y titulares de diarios puede estar restringido o tener un costo elevado. Esto impediría realizar análisis de sentimiento.		
Indicador: Limitada disponibilidad de datos de redes sociales.		
Respuesta: Transferir. Realizar una investigación para identificar servicios de terceros que proporcionen información del sentimiento del mercado en tiempo real.		

Contingencia: Modificar el alcance del proyecto y realizar las predicciones de precio sin un módulo de análisis en redes sociales.

Recursos

A continuación se listan los recursos disponibles para le proyecto

- Recursos humanos
    - Estudiante
    - Director
    - Asesor temático
- Bienes de capital
    - PC de escritorio
    - Servicios de almacenamiento
- Recursos de backtesting
    - Plataforma ReMarkets
- Recursos de conectividad e infraestructura
    - Servicio de internet.
    - Servicio eléctrico.

Presupuesto

En la siguiente tabla se detallan los distintos gastos asociados al proyecto, donde:

- VA: Valor actual
- VR: Valor residual
- VFP: Vida futura probable

Concepto	Detalle	VA	VR	VFP (meses)	Amortización mensual	Total (4 meses)
BIENES DE CAPITAL						

Computadora	MSI B450 Pro - 16GB Ram - Ryzen 1500x - Rx 580 8GB - SSD 480GB - HDD 1 TB - Monitor 24''	$400,000	$60,000	60	$5,667	$22,667
Concepto	Detalle				Gasto mensual	Total (4 meses)
VIAJES Y VIÁTICOS						
Merienda	$6000 mensuales				$6,000	$24,000
MATERIALES E INSUMOS						
Servicio de Electricidad	$2000 mensuales				$2,000	$8,000
Servicio de Internet	$3000 mensuales				$3,000	$12,000
Concepto	Perfil	Horas (mes)	Sueldo (hora)		Gasto mensual	Total (4 meses)
RECURSOS HUMANOS						
Alumno	Programador IA	91	$1,807		$164,437	$657,748
Director	Lider de Proyecto	30	$3,188		$95,640	$382,560
Asesor en Machine Learning	Consultor BI/IA	15	$3,400		$51,000	$204,000
Total						**$1,310,975**

La remuneración por hora fue estimada a partir de los salarios mensuales provistos por el Consejo Profesional de Ciencias Informáticas de la Provincia de Buenos Aires [10]. El proyecto ha sido planificado y estructurado basándose en una duración aproximada de 4 meses.

Matriz de comunicaciones

En la siguiente página se muestra una planificación de las diversas comunicaciones requeridas a lo largo del proyecto.

Matriz de Comunicaciones

Proyecto:	Desarrollo de un sistema inteligente para la toma de decisiones de inversión en el mercado financiero utilizando herramientas de machine learning

ETAPA	Elemento de la EDT	¿Qué comunicamos?	¿Por qué?	Destinatarios (stakeholders)	Método de Comunicación	Responsables	Fecha de inicio	Frecuencia	Fecha de finalización
1. Análisis del Proyecto.	1.3 Definición del alcance del proyecto	Entrega de documento de justificación, objetivos y alcance.	Fecha estipulada de entrega del documento de justificación, objetivos y alcance.	Director de Proyecto	Entorno FICH	Programador IA	22/08/23	Una vez	
2. Recopilación de Datos	2.1 Búsqueda de datos relevantes para el análisis bursátil	Consulta sobre disponibilidad de datos.	Recibir asesoriamiento sobre los datos disponibles en la plataforma de simulacion de trading y consultar sobre otras fuentes.	Soporte ReMarkets	Mail	Programador IA	23/08/23	Una vez	
2. Recopilación de Datos	2.1 Búsqueda de datos relevantes para el análisis bursátil	Entrega de informe que describa los datos a procesar para entrenar los modelos, la fuente y de qué forma serán consumidos.	Primer Informe de avance solicitado por el Equipo Directivo.	Equipo Directivo	Entorno FICH	Programador IA	31/08/23	Una vez	
2. Recopilación de Datos	2.3 Organización y almacenamiento de los datos	Obtener asesoramiento sobre el tipo de base datos a utilizar para el almacenamiento de datos.	Es necesario decidir sobre las diferentes alternativas de base de datos.	Director de Proyecto	Mail	Programador IA	01/09/23	Una vez	
3. Procesamiento de datos	Hito: Datos recolectados y organizados en una base de datos.	Informar sobre los datos obtenidos y solicitar apoyo en el procesamiento, análisis y normalizácion de datos.	El procesamiento de datos es una etapa critica del proyecto, lo que requiere apoyo por parte de un experto en ML.	Asesor en ML	Mail	Programador IA	12/09/23	Una vez	
3. Procesamiento de datos	Hito: Datos recolectados y organizados en una base de datos.	Reunión de seguimiento sobre la etapa de procesamiento de datos.	Informar sobre los avances y obtener un feedback.	Asesor en ML	Videoconferencia/ Presencial	Programador IA	13/09/23	Una vez por semana	3/10/23
4. Seleccion de modelos	Hito: Set de datos limpios con características más relevantes para el problema.	Informar sobre posibles arquitecturas de ML para cada modelo.	Es necesario realizar una investigación y evaluación de las arquitecturas candidatas a ser implementadas.	Programador IA	Mail	Asesor en ML	04/10/23	Una vez	
3. Procesamiento de datos	Hito: Set de datos limpios con características más relevantes para el problema.	Entregar de conjunto de datos limpios y procesados. También un informe de análisis de datos donde se detallen los resultados del análisis exploratorio y la selección de características y variables relevantes.	Segundo Informe de avance solicitado por el Equipo Directivo.	Equipo Directivo	Entorno FICH	Programador IA	05/10/23	Una vez	
5. Entrenamiento y ajuste de modelos	Hito: Selección final de los modelos de ML que se utilizarán en el proyecto.	Reunión de seguimiento sobre la etapa de entrenamiento y ajuste de modelos.	Informar sobre los avances y obtener un feedback.	Asesor en ML	Videoconferencia/ Presencial	Programador IA	20/10/23	Una vez por semana	16/11/23
5. Entrenamiento y ajuste de modelos	Hito: Modelos entrenados y ajustados con un rendimiento aceptable	Código con los modelos implementados. Informe que describa los modelos utilizados, métricas para evaluar el desempeño y resultados obtenidos hasta la fecha.	Tercer Informe de avance solicitado por el Equipo Directivo.	Equipo Directivo	Entorno FICH	Programador IA	9/11/23	Una vez	
6. Puesta en producción	6.1 Estudio del entorno de producción	Consultas de dudas relacionadas al entorno de simulación ReMarkets.	Pueden ocurrir errores al operar en el entorno mediante la API.	Soporte ReMarkets	Mail	Programador IA	21/11/23	Una vez por semana	6/12/23
6. Puesta en producción	Hito: Modelo en producción y operativo en el sistema de trading.	Entrega final del proyecto.	Se realiza la entrega del proyecto finalizado en la fecha planificada.	Equipo Directivo	Entorno FICH	Programador IA	21/12/23	Una vez	

Referencias

[1] Saad, L., & Jones J. (2022). What Percentage of Americans Own Stock?. Gallup. www.news.gallup.com/poll/266807/percentage-americans-owns-stock.aspx

[2] Santander (2022). What is happening with global inflation?. www.santander.com/en/press-room/press-releases/2022/02/what-is-happening-with-global-inflation

[3] Joubert, T. (2022). ¿Cómo afecta la inflación a la bolsa de valores?. Ig. www.ig.com/es/estrategias-de-trading/-como-afecta-la-inflacion-a-la-bolsa-de-valores--211008

[4] Chatterjee, S. (2019). How to train your machine: JPMorgan FX algos learn to trade better. Reuters. www.reuters.com/article/us-jpm-trading-machines/how-to-train-your-machine-jpmorgan-fx-algos-learn-to-trade-better-idUSKCN1S61JG

[5] Segal, T. (2022). Fundamental Analysis: Principles, Types, and How to Use It. Investopedia. www.investopedia.com/terms/f/fundamentalanalysis.asp

[6] Hayes, A. (2022). Technical Analysis: What It Is and How to Use It in Investing. Investopedia. www.investopedia.com/terms/t/technicalanalysis.asp

[7] Dotras, E. (2014). La nueva era de los mercados financieros y su globalización. Oikonomics. www.oikonomics.uoc.edu/divulgacio/oikonomics/es/numero02/dossier/eruiz.html

[8] Korolov, M. (2021). What is sentiment analysis? Using NLP and ML to extract meaning. CIO. www.cio.com/article/189218/what-is-sentiment-analysis-using-nlp-and-ml-to-extract-meaning.html

[9] Altexsoft. (2022). What is an API: Definition, Types, Specifications, Documentation. www.altexsoft.com/blog/engineering/what-is-api-definition-types-specifications-documentation/

[10] CPCIBA. (2023). Tabla de referencia de honorarios - Actualización marzo 2023. https://www.cpciba.org.ar/honorarios