ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERIA DE SISTEMAS

UNIDAD DE TITULACIÓN

Desarrollo de un modelo de predicción para los valores de las Acciones del Índice S&P500, sobre la base del estado del arte de la temática, usando un enfoque de análisis de datos, que incluya técnicas de minería de datos y de aprendizaje automático

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MAGISTER EN SISTEMAS DE INFORMACIÓN

JORGE ANDRÉS CÁRDENAS TORRES
jorge.cardenas@epn.edu.ec

Director: Carlos Montenegro Armas
carlos.montenegro@epn.edu.ec

2023

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación Desarrollo de un modelo de predicción para los valores de las Acciones del Índice S&P500, sobre la base del estado del arte de la temática, usando un enfoque de análisis de datos, que incluya técnicas de minería de datos y de aprendizaje automático desarrollado por Jorge Andrés Cárdenas Torres, estudiante de la maestría en sistemas de información, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

Carlos Montenegro

DIRECTOR

DECLARACIÓN DE AUTORÍA

Yo, Jorge Andrés Cárdenas Torres, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Firmado electrónicamente por:
JORGE ANDRES
CARDENAS TORRES

Jorge Andrés Cárdenas Torres

DEDICATORIA

A mi esposa Zamar Silva, quien ilumina mis días, me hace ser fuerte en cada circunstancia de la vida y con su amor me alienta a que me supere cada día.
A mi madre Teresita Torres, que siempre fue mi apoyo aquí en a tierra y ahora en el lugar de paz que esté me sigue bendiciendo.
A mi padre Jorge Cárdenas, que no importa el lugar y el tiempo, siempre tiene un corazón dispuesto, presto para animarme brindándome su cariño y consejo oportuno.
A mis hermanos, Danita y David por acompañarme en todo este camino, y ser mi motivación de superación.
A mi segunda madre, Jhennycita quien con su cariño está siempre presente en mis momentos más difíciles y felices.

A todos ustedes infinitas gracias por siempre impulsarme a salir adelante.

AGRADECIMIENTO

Quiero extender un cordial agradecimiento a mis profesores de Pre y Post grados por conferir los conocimientos necesarios para llevarlos a la práctica en mis labores diarias y en este presente trabajo.

Un agradecimiento especial a mi tutor Carlos Montenegro, por siempre estar pendiente del trabajo que realizaba.

ÍNDICE DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

RESUMEN

La predicción de precios en el mercado de valores es un desafío complejo debido a su naturaleza dinámica y volátil. Este trabajo propone un enfoque basado en redes neuronales LSTM (Long Short-Term Memory) para predecir las acciones del S&P500, uno de los índices más importantes del mercado de valores. El modelo se entrena utilizando un conjunto de datos históricos del S&P500 desde enero de 2015 a mayo de 2023, que incluye información sobre precios de apertura, cierre, máximo y mínimo, y volumen. Los resultados experimentales demuestran que el modelo propuesto supera a otros enfoques de aprendizaje y predicción. Los resultados de este trabajo pueden ser de gran utilidad para los inversores y analistas financieros en la toma de decisiones de inversión. Sin embargo, se sugiere que se realicen investigaciones adicionales para explorar otras variantes de modelos LSTM y mejorar aún más la precisión de la predicción.

Palabras clave: Índice S&P500. Mercados bursátiles. Predicción. Finanzas.

ABSTRACT

The work proposes an approach based on LSTM (Long Short-Term Memory) neural networks to predict the actions of the S&P500, one of the most important indexes of the stock market. Predicting prices in the stock market is a complex challenge due to its dynamic and volatile nature.

The model is trained using a historical data set of the S&P500 from January 2015 to May 2023, which includes information on open, close, high, and low prices, among others.

The experimental results show that the proposed model outperforms other traditional prediction approaches. The LSTM-based approach achieves higher accuracy in S&P500 price prediction, which can be of great use to investors and financial analysts in making investment decisions. However, it is suggested that further research be done to explore other variants of LSTM models and further improve the prediction accuracy.

Keywords: S&P 500. Index. Forecast. Finance.

1. INTRODUCCIÓN

En la actualidad el mundo produce más cantidad de datos y resultados, que siglos de ciencia (Serrano-Cobos, 2016). Uno de estos campos es el financiero ya que a diario se realizan millones de operaciones alrededor del mundo (Yahoo, 2023). Los mercados bursátiles son cambiantes en todo momento, por ejemplo, pueden verse afectados por crisis económicas, militares, opiniones de personas influyentes, nuevas tecnologías emergentes, etc.

En este sentido el tratar de predecir el valor de acciones en mercados cambiantes es un verdadero desafío, más aún el crear un modelo que tenga en cuenta todo lo que implica un sistema lleno de variables sensibles y dependientes unas de otras.

El presente trabajo se enfoca en crear un modelo que pueda adaptarse a estas variaciones, tomando una cantidad de datos que permitan lograr los resultados deseados.

Apoyado en la revisión sistemática de la lectura (Kitchenham, 2004) se dará énfasis en encontrar las mejores opciones que posteriormente coadyuven a crear, probar y proponer un modelo adecuado que como resultado permita realizar predicciones acertadas del valor del índice de las acciones S&P 500, las mismas que son importantes en el mundo bursátil, encaminados a la toma de decisiones para ejecutar o no inversiones.

1.1. Objetivo general

Desarrollar un modelo de predicción para los valores de las Acciones del Índice S&P500, sobre la base del estado del arte de la temática, usando un enfoque de análisis de datos, que incluya técnicas de minería de datos y de aprendizaje automático.

1.2. Objetivos específicos

a) Determinar el estado del arte de los modelos de Minería de Datos usados para la predicción de los valores del Índice S&P500.
b) Definir un modelo predictivo alternativo que incluya técnicas de minería de datos y de aprendizaje automático.
c) Validar el modelo.

1.3. Marco Teórico

Predecir el valor de los índices bursátiles es una tarea muy desafiante, principalmente ya que los datos son muy ruidosos, no lineales, complejos, dinámicos, no paramétricos y caóticos (Kumar, Jain, & Singh, 2020) (Yin & Si, 2013) (Molina & Montenegro, 2020). Sin embargo, es necesario tomar en cuenta que los índices bursátiles tienen poder predictivo como indicador adelantado de la economía (Kumar, Jain, & Singh, 2020) (Comincioli, 1996). Como es conocido, los datos del Índice S&P500 se refieren a los valores diarios de las acciones (Apertura, Máximo, Mínimo, Cierre) y al Volumen de las acciones del mercado de valores. Los datos día a día están disponibles en múltiples fuentes; por ejemplo, en (Yahoo, 2023) La predicción se la puede realizar sobre cualquiera de los valores; por ejemplo, el valor de apertura de cada día hábil. Teniendo en cuenta que, para los modelos de aprendizaje automático, es crucial tener tanto la cantidad como la diversidad de datos disponibles durante el entrenamiento, los datos del índice S&P500 deben procesarse previamente para incluir variables que los interpreten y mejoren los modelos. De esta manera, puede ser recomendable, inclusive, utilizar técnicas de aumento de datos para generar mejoras interesantes en la eficacia del modelo, como se informa en la literatura técnica (Le Guennec, Malinowski, & Tavenard, 2016) (Khushi & Mukherjee, 2021) (Yang, Zhang, Cui, & Cui, 2021).

Se utilizan varios enfoques para pronosticar el mercado de valores. El análisis fundamental utiliza conceptos económicos para predecir los precios de las acciones. Sin embargo, el comportamiento complejo del mercado financiero puede potencialmente tergiversar o a su vez no representar las características críticas de los datos subyacentes (Mostafa, Dillon, & Chang, 2017). Otra opción es el análisis técnico, basado en el estudio de estadísticas generadas por el propio mercado, que considera que la cotización ya comprende todos los fundamentos que la afectan y modela los comportamientos históricos en series de tiempo (Saleh, 2018). Modelos como ARMA, ES, ARIMA, ARCH y GARCH predicen los precios de las acciones en el futuro basándose en los precios de las acciones en el pasado (Box, Jenkins, Reinsel, & Ljung, 2015). Además, estos modelos se basan en el supuesto de que las series de tiempo financieras se generan a partir de procesos lineales.

Por otro lado, los enfoques de computación blanda reportan una ventaja sobre los datos económicos y las series de tiempo (Beyaz, 2019). Al momento, las opciones incluyen Red neuronal artificial (RNA), Máquinas de vectores de soporte (SMV) y Algoritmos genéticos (GA). Hay características deseables de RNA, lo que las convierte en técnicas adecuadas para pronosticar los valores comerciales de las acciones. Los resultados reportados muestran la ventaja de RNA sobre los modelos estadísticos, entre otras razones, porque

son naturalmente aplicables a las no linealidades de los datos (Mostafa, Dillon, & Chang, 2017) (Qian, 2018) (Kang & Lee, 2020). Las arquitecturas generalizadas de aprendizaje profundo, incluido el perceptrón multicapa (MLP) (Carta, Corriga, Ferreira, & Podda, 2021) (Hajiagha & Farahani, 2021) (Wang, 2020), LSTM (Onibonoje, Djoussa, & Roantree, 2020) y las redes neuronales convolucionales (CNN) (Onibonoje, Djoussa, & Roantree, 2020), son algunos de los modelos no lineales que se han mostrado prometedores en este dominio.

Los modelos de minería de datos son herramientas poderosas para predecir valores de acciones debido a su capacidad para descubrir patrones y tendencias ocultas en los datos (Lifeng, Tao, Jingjun, & Haiyang, 2011).

Estos modelos pueden procesar grandes conjuntos de datos históricos de precios de acciones, así como datos relacionados con variables económicas, políticas y sociales que pueden afectar los mercados. Esto les permite capturar y analizar una amplia gama de información que puede influir en el comportamiento de los precios de las acciones. Así mismo pueden detectar patrones y tendencias en los datos históricos de precios de acciones. Pueden analizar múltiples variables y relaciones complejas entre ellas para identificar correlaciones significativas. Por ejemplo, un modelo puede descubrir que ciertos indicadores económicos o eventos geopolíticos tienen un impacto en los precios de las acciones.

2. METODOLOGÍA

A continuación, se revisará el apartado de metodología la cual será la guía a lo largo del presente trabajo, tomando en cuenta que metodología se define como: “Conjunto de métodos que se siguen en una investigación científica o en una exposición doctrinal.” (ESPAÑOLA, 2023). Por la particularidad de los datos y la naturaleza académica del proyecto, no se usará una metodología tradicional. En su lugar, de acuerdo con el marco teórico se tratará este trabajo con una metodología específica según las fases a continuación:

- **FASE I**: Identificación y análisis de los trabajos de relevancia acerca de la predicción de los valores de las acciones de la Bolsa y parámetros relacionados (estado del arte).
- **FASE II**: Definición de un procedimiento de predicción basado en análisis de datos, que incluya técnicas de minería de datos y de aprendizaje automático.
- **FASE III**: Colección y preprocesamiento de datos del Índice S&P500. Corrida y validación técnica del procedimiento para representar los valores históricos del Índice y predecir sus valores futuros.

Así también se puede apreciar el esquema por fases en la Figura 1.

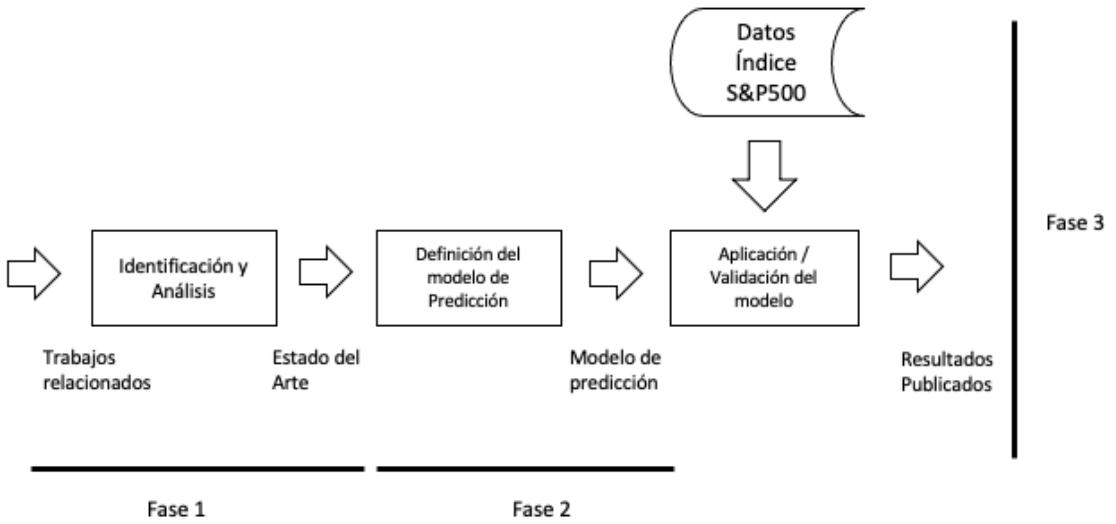

Figura 1 – Fases de la metodología del trabajo de investigación

Tras la descripción general y de alto nivel de cómo se va a estructurar la metodología, ahora, se realizará una descripción más detallada del trabajo en cada fase.

2.1. Fase I

En una primera fase se realiza la revisión sistemática de la literatura. Conforme a lo explicado por (Tebes, Peppino, Becker, & Olsina, 2019), la revisión sistemática de la literatura es un proceso destinado a obtener información de diversos artículos científicos almacenados en bibliotecas o repositorios digitales (Tebes, Peppino, Becker, & Olsina, 2019). Dicho proceso debe ser reproducible, sistemático y auditable. En otras palabras, la revisión sistema debe ser clara para cualquier investigador que quiera reproducir y/o revisar los resultados obtenidos en este trabajo, por ejemplo. Para iniciar este proceso el primer paso es el diseño de la revisión sistemática de la literatura (RSL).

2.1.1. Diseño de la revisión sistemática de la literatura

El paso de inicio de una RSL es el diseño. De acuerdo las sugerencias de (Kitchenham, 2004), el diseño se compone de:

- Especificar preguntas de investigación
- Especificar cadena de búsqueda
- Identificar Metadatos para Búsqueda
- Seleccionar Bibliotecas Digitales
- Definir criterios de Selección y Calidad
- Diseñar formulario de Extracción de datos
- Validación del diseño

2.1.2. Ejecución de la revisión sistemática de la literatura

Es recomendable en este punto realizar varias pruebas en varios buscadores de bibliotecas digitales fin de ajustar la cadena de búsqueda que se va a emplear. Esta actividad debe ser iterativa hasta que llegar a la cantidad y calidad de documentos deseados, una vez esto sea satisfactorio se puede proceder a la extracción de datos, para su posterior análisis. Como último paso se debe informar de los resultados obtenidos en el caso de este artículo, se lo vera plasmado en la sección Resultados. El diseño de la RSL se lo describe en la tabla 1:

Tabla 1 – Ficha para el diseño de la Revisión sistemática de la literatura

Preguntas de investigación	
PI_1: ¿Existen trabajos que hablen sobre la predicción de índices bursátiles? PI_2: ¿Existen trabajos que hablen sobre la predicción del índice S&P 500?	
Protocolo de búsqueda	
Cadena de búsqueda	"S&P500 INDEX" AND ("Forecast" OR "Predict")
Metadatos para búsqueda	Título; Resumen; Palabras Clave
Bibliotecas digitales seleccionadas	
Criterios de selección y de calidad	
Criterios de inclusión	1) El trabajo deberá estar publicado en los últimos 5 años; 2) 2) El trabajo debe pertenecer al área de ML, IA, Data mining.
Criterios de inclusión	1) El trabajo debe estar escrito en Inglés o Español.
Criterios de calidad	1) El trabajo debe describir claramente los resultados, 2) El trabajo debe contar con técnicas de regresión
Extracción de trabajos relevantes	
• Autor y titulo • Técnicas utilizadas • Rango de datos • Medidas de rendimiento	

2.2. Fase II

Basándose en la información del conjunto preliminar de artículos, se recopilarán las mejores técnicas y modelos empleados hasta el momento lograr predecir el valor de las acciones del Índice S&P500.

Para realizar predicciones se debe usar técnicas de machine learning de regresión, debido a que al contrario de las técnicas de clasificación estas trabajan con números infinitos tratando en este caso de predecir cual sería el valor monetario de una acción en un futuro. Los modelos de clasificación, en cambio tratan de etiquetar según un conjunto finito de valores el resultado del análisis que están haciendo (Yamamoto, Yoshii, Kinoshita, & Touyama, 2020).

Existen diferentes tipos de redes neuronales tales como: Convolucionales, LSTM, RNN, etc.

Las cuales realizan sus procesos en capas ocultas. Dichas capas contienen un número "n" de neuronas, que se ocupan para procesar las características de los datos ingresados.

2.3. Fase III

Tras el análisis y diseño del modelo para realizar las predicciones de los valores de las acciones del S&P500, serán ingresará datos reales los cuales van desde enero de 2015

hasta mayo de 2023. Si el resultado de la evaluación del modelo cumple con parámetros aceptables, se publicará los resultados.

En relación con las medidas de rendimiento se puede observar que los artículos que han empleado dicha medida están enfocados en realizar predicciones de mercados bursátiles y en muchos casos en específico de las acciones del índice S&P 500.

En este punto diferentes medidas de rendimiento aparecen para tratar de explicar los resultados obtenidos luego de los análisis.

- Error cuadrático medio (MSE)
- Raíz de Error cuadrático medio (RMSE)
- Error absoluto medio (MAE)
- R al cuadrado (R^2)

El Error cuadrático medio (MSE) es una métrica simple y de uso común para regresiones. Mide el error cuadrado promedio de los resultados del modelo. Por cada valor realiza un cálculo de la diferencia cuadrada entre las predicciones y el resultado que se esperaba, para después promediar esos valores. Realizando un análisis basado en esta métrica, se tiene que a más alto resultado el modelo es peor. Este resultado jamás será de valores negativos.

La raíz del error cuadrático medio se expresa sacando la raíz cuadrada del MSE. Es una medida de precisión para realizar la comparación de errores de predicción de diversos modelos para un conjunto de datos en particular. Esta magnitud nunca podrá ser negativa.

El resultado de R al cuadrado (R^2) oscila entre 0 y 1. Cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar. De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será.

3. RESULTADOS Y DISCUSIÓN

3.1. Resultados Fase I:

Para construir la cadena de búsqueda se realizaron varias pruebas, se utilizó “S&P index”, obteniendo un resultado de más de 170 000 estudios. Después se utilizó “S&P Forecasting”, obteniendo alrededor de 200 000 artículos. El siguiente paso fue probar con “S&P 500 index forecast” y la búsqueda arrojo 2 000 resultados. Finalmente se utilizó la cadena “S&P 500 index forecast” OR “prediction” se refino también mediante la búsqueda avanzada a solo artículos de software, y el período de tiempo a partir de 2017, obteniendo un resultado de alrededor de 200 estudios por biblioteca. El resultado de este refinamiento se lo puede ver en la figura 2.

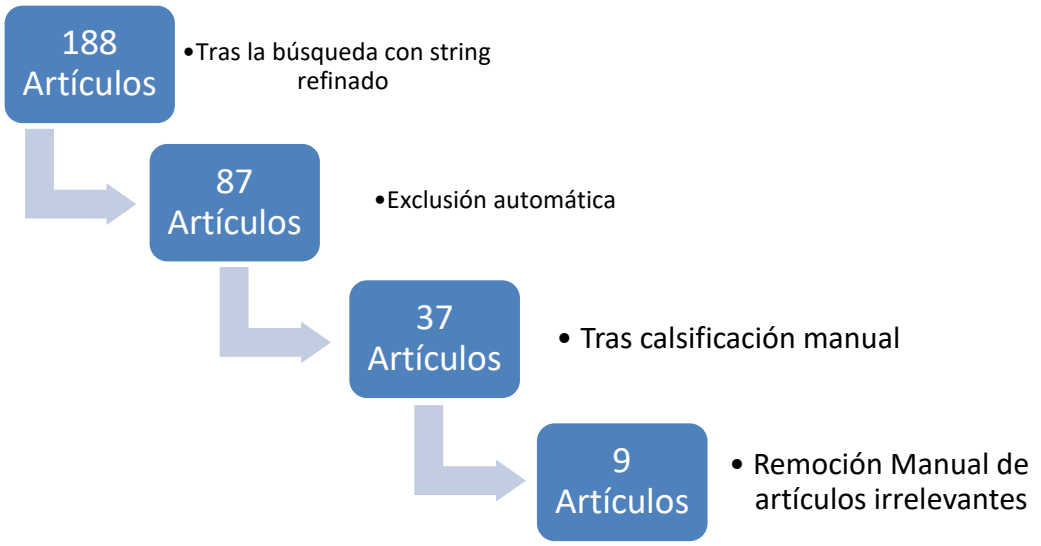

Figura 2 – Proceso de clasificación de artículos

En el transcurso de la revisión sistemática se debe definir los criterios de inclusión y exclusión, con el fin de evitar sesgos. En la tabla 2 se puede ver ciertos tipos de criterios elegidos para esta revisión.

Tabla 2 – Criterios de inclusión, exclusión

Tipo de Criterio	Descripción	Criterio
Periodo de tiempo	Para este caso los artículos serán seleccionados de 2017 en adelante, debido a la rápida evolución de las TICS	Inclusión: De 2017 a 2023 Exclusión Antes de 2017
Lenguaje	Se excluyen artículos de acuerdo con el lenguaje escrito	Exclusión: Lenguajes diferentes al inglés y español
Tipo de fuente	Dependiendo del origen, por ejemplo: conferencia o revistas científicas	Inclusión: Artículos de conferencias y revistas

3.1.1. Trabajos relacionados con predicción de acciones bursátiles

Después de realizar la clasificación de todos los trabajos correspondientes con el tema de estudio fueron agregados a la tabla 4.

Tabla 3 – Trabajos relacionados, predicción S&P 500

Autor, Título	Técnicas Usadas	Rango de datos	Medidas de Rendimiento
(Montenegro & Molina, 2019)	DNN	Junio 2013 a Junio 2018, por día	R^2 Otras
(Pan & Li, 2022)	LSTM GRU	Diciembre 2017 a Junio 2018, por día	Movement direction accuracy (MDA) (MSE)
(Mohan & Durairaj, 2022)	Chaos Theory CNN LSTM	Enero 1990 a Enero 2021, por mes	(MSE) (MAPE) (Dstat)
(Zhu, Shen, & Angelova, 2022)	Logistic Weighted Dynamic Time Warping (LWDTW) Weighted Dynamic Time Warping (WDTW)	Enero de 2005 a Noviembre de 2017, a diario	(MAPE), (MAE),(MSE), (RMSE), (R^2).
(Gowthul & Baulkani, 2019)	Multi-kernel support vector machine (MKSVM) Fruit fly optimization (FFO)	Enero 2001 a Octubre 2009, Diario	Absolute Percentage error (APE), (RMSE) Relative Mean Absolute Error (ReIMAE)

(Chacon, Kesici, & Najafirad, 2020)	LTSM (EMD)	Enero 2018 a Abril 2020, a Diario	Accuracy MAPE
(Hajiagha & Farahani, 2021)	ARIMA-ANN	Noviembre 2013 a Julio de 2018, a Diario	MSE
(Huynh, Dang, & Duong, 2017)	RNN GRU BGRU LSTM	Octubre 2006 a Noviembre 2013, Diario	Accurancy

El uso del error medio cuadrático (EMC) como función de pérdida es común en muchos problemas de aprendizaje automático, especialmente en problemas de regresión. Hay varias razones por las que se prefiere el EMC como medida de pérdida en estos casos:

- Interpretación geométrica: El EMC tiene una interpretación geométrica intuitiva. Representa la distancia cuadrática promedio entre las predicciones del modelo y los valores reales. Al elevar al cuadrado los errores, se enfatizan los errores grandes, lo que hace que el modelo sea más sensible a las predicciones incorrectas. Al minimizar el EMC, el modelo tiende a ajustarse mejor a los datos de entrenamiento.
- Diferenciabilidad: El EMC es una función diferenciable, lo que facilita su uso en algoritmos de optimización. Al minimizar el EMC, se pueden aplicar métodos de optimización como el descenso de gradiente para encontrar los parámetros del modelo que minimizan la pérdida.
- Sensibilidad a los errores grandes: El uso del EMC como función de pérdida penaliza de manera más significativa los errores grandes en comparación con otras funciones de pérdida, como el error absoluto medio. Esto puede ser beneficioso en problemas en los que los errores grandes son especialmente problemáticos o costosos.

3.2. Resultados Fase II:

3.2.1. Arquitectura del modelo

Para la predicción de series de tiempo, mercados bursátiles, se tiene varios tipos de redes neuronales a escoger, por ejemplo, redes convolucionales, recurrentes o redes LSTM.

Las redes convolucionales, presentan una desventaja al momento de trabajar con series temporales. Al no tener un estado de memoria, los pesos para el aprendizaje van tendiendo a cero, lo cual produce estancamiento y la red neuronal deja de aprender. Una red neuronal recurrente ataca este problema mediante “memoria”, es decir guarda el estado anterior, la variante mejor adaptada para trabajar con series de tiempo es la LSTM (Menacho, 2014)

La principal fortaleza de las redes LSTM radica en su capacidad para manejar el problema del desvanecimiento del gradiente (Kumar, Gupta, & Sehgal, 2014), que ocurre cuando se propagan los gradientes a través de múltiples pasos de tiempo en las RNN convencionales. Este problema puede hacer que los gradientes se vuelvan extremadamente pequeños y dificultar el aprendizaje de relaciones a largo plazo.

La arquitectura de una red LSTM se compone de unidades de memoria llamadas "celdas LSTM". Cada celda LSTM tiene tres componentes principales: una celda de memoria, una puerta de entrada y una puerta de salida.

- **Celda de memoria:** La celda de memoria es responsable de almacenar y mantener la información a largo plazo. Actúa como una especie de "cinta transportadora" que puede agregar o eliminar información a medida que se procesa la secuencia de entrada. La celda de memoria se actualiza en cada paso de tiempo y se basa en la entrada actual y en la información almacenada previamente en la celda.
- **Puerta de entrada (Input gate):** La puerta de entrada determina cuánta información nueva debe agregarse a la celda de memoria en cada paso de tiempo. Se basa en la entrada actual y en la información de la iteración anterior de la red.
- **Puerta de olvido (Forget gate):** La puerta de olvido controla cuánta información anterior debe descartarse de la celda de memoria en cada paso de tiempo. Al igual que la puerta de entrada, se basa en la entrada actual y en la información de la iteración anterior.
- **Puerta de salida (Output gate):** La puerta de salida determina cuánta información de la celda de memoria debe transmitirse a la salida en cada paso de tiempo. La salida se calcula en función de la información actual de la celda de memoria y de la entrada actual.

Los componentes anteriormente descritos se los puede ver en la figura 3.

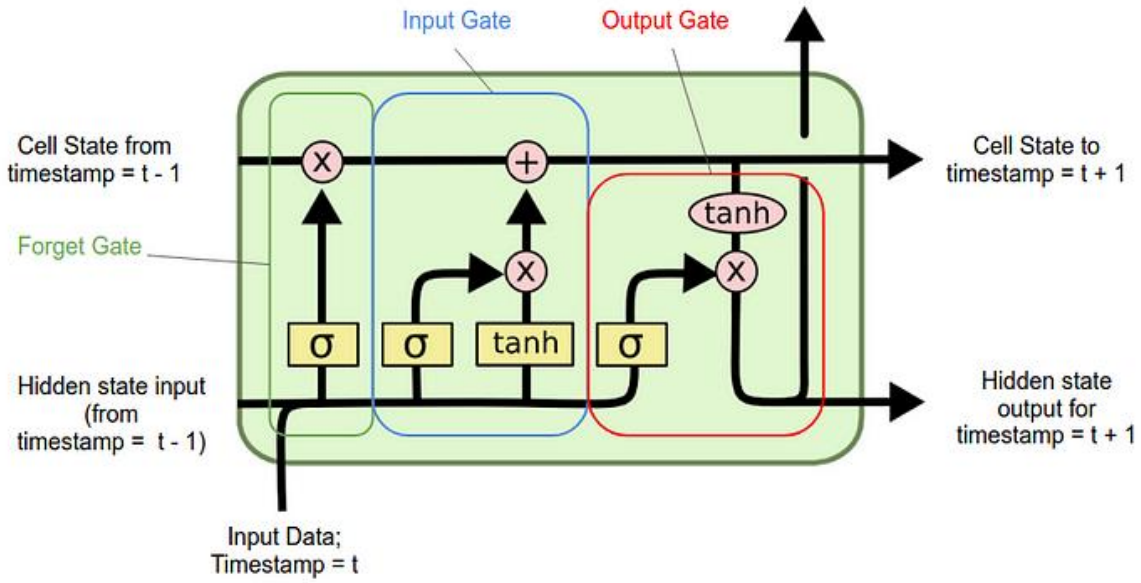

Figura 3 – Unidad recurrente LSTM
(Ryan, 2022)

En cuanto a la configuración de capas LSTM en una red neuronal, generalmente se colocan en secuencia, una tras otra, donde cada capa LSTM toma como entrada la salida de la capa anterior. Esto permite que la red aprenda representaciones de mayor nivel a medida que se profundiza en las capas.

El diseño del modelo tras la revisión de los trabajos comprende la siguiente estructura, una neurona de entrada, la cual corresponde a los valores más altos de las acciones por día. Seguida de una capa oculta con 40 neuronas, tras esto se utilizó una capa oculta de 20 neuronas, para a continuación colocar una capa oculta de 5 neuronas. Las capas anteriormente mencionadas fueron configuradas con una función de activación “tanh”, la razón de usar esta función es debido a que, en varios ensayos, la función de activación “relu” ≈ 0.1029 tuvo mayor índice de perdida que la función “tanh” ≈ 7.6409e-05. Hay que mencionar que la estructura de capas descritas anteriormente se basó en el trabajo de (Montenegro & Molina, 2019), autores que usaron esta estructura con buenos resultados.
En la capa de salida, se utilizó una neurona la cual al ser densa tiene una función de activación por defecto “relu”. La cantidad de épocas para el entrenamiento fue de mil, debido a que el error en la validación no fue alto, y al revisar modelos propuestos este fue el numero en general de épocas utilizadas. Se debe tomar en cuenta que el modelo propuesto es alimentado hacia adelante (Feedforward), ya que las conexiones de la unidad

no viajan en un bucle, sino en una única ruta dirigida. La arquitectura del modelo lo podemos ver plasmado en la figura 4.

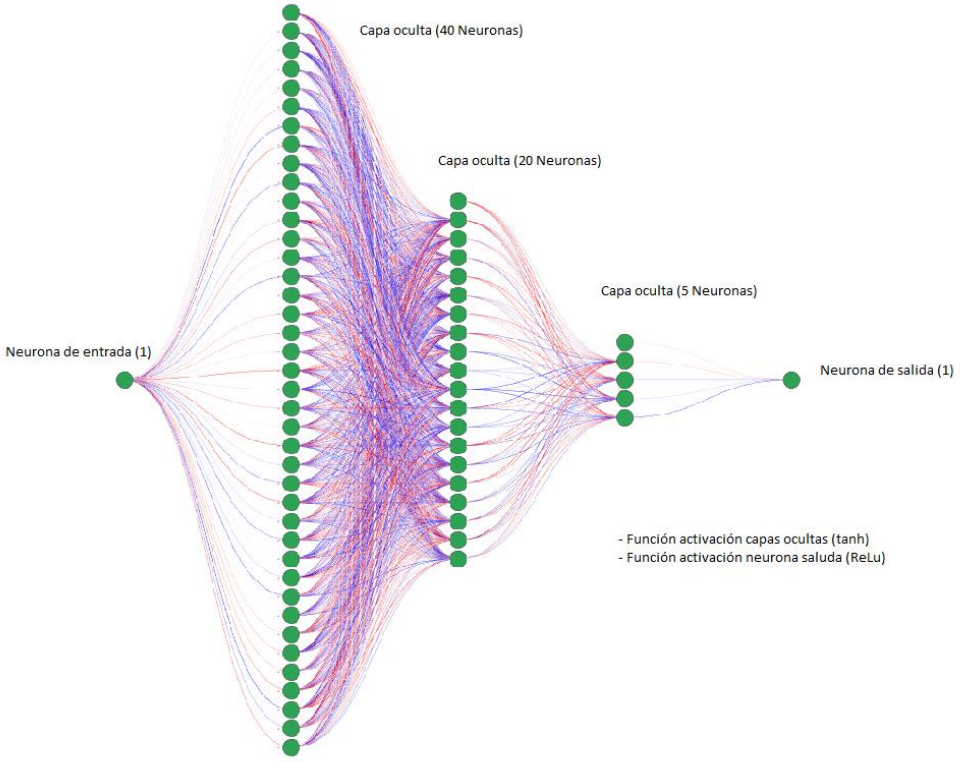

Figura 4 – Arquitectura del modelo

Para realizar la compilación del modelo se utilizó el algoritmo optimizador “adam”,
El algoritmo ADAM utiliza tanto el gradiente de los parámetros del modelo como los momentos de primer y segundo orden para adaptar la tasa de aprendizaje durante el entrenamiento. Esto le permite converger más rápido y lidiar mejor con problemas de gradiente disperso o funciones de pérdida no lineales (Diederik & Lei Ba, 2015).

3.3. Resultados Fase III:

Para que el modelo procese los datos y realice predicciones de los datos descargados del S&P500 se trabajó con la columna correspondiente a “High” es decir los valores diarios más altos de las acciones. En la fase de entrenamiento, sobre el 80% del total de los datos, se obtuvo $R^2 \approx 0.9983$. También se observa que el error medio cuadrático es de 26.54, lo cual es un valor más bajo de los resultados de referencia. En la figura 5 se muestran los resultados.

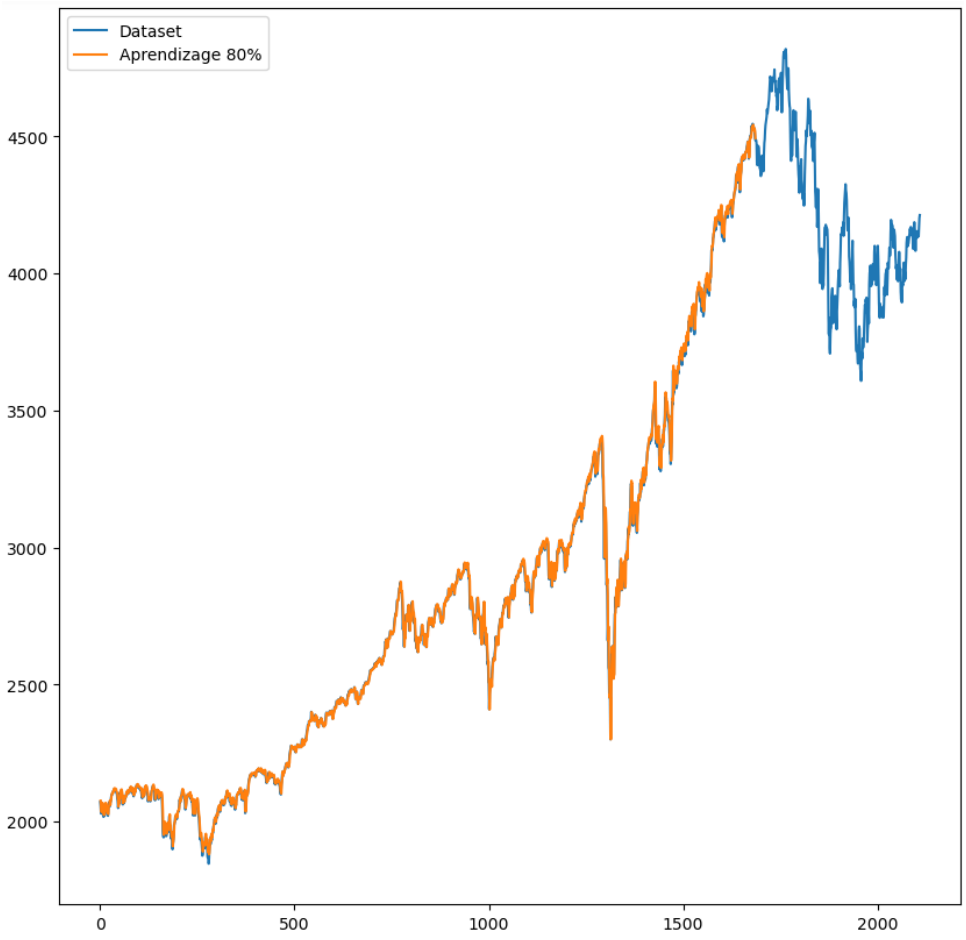

Figura 5 – Proceso de entrenamiento del modelo

En la fase de validación, se usa el 20% de los datos. Para este caso se observa que el modelo obtiene un $R^2 \approx 0.9557$ (Figura 6).

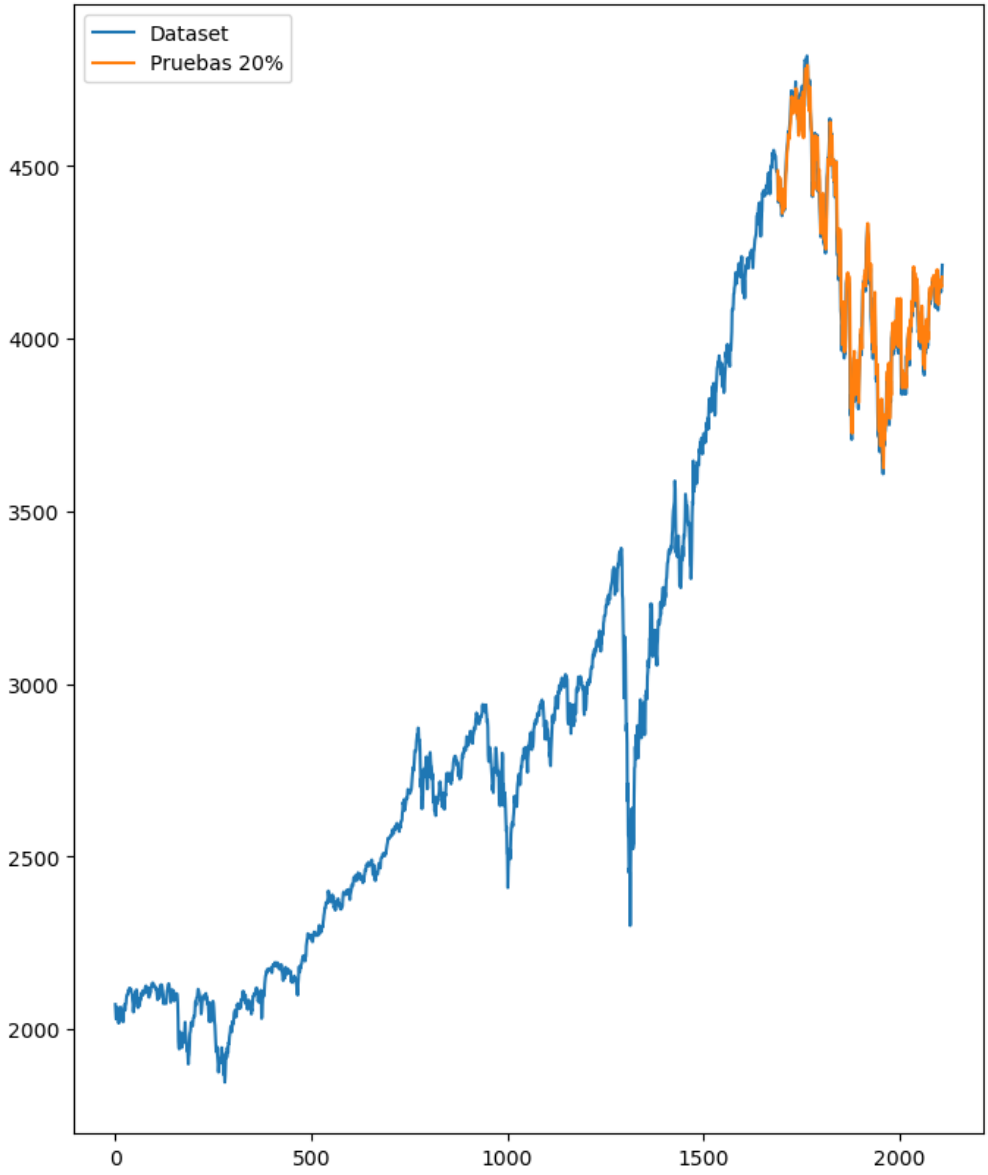

Figura 6 – Proceso de prueba del modelo

La validación del modelo sobre el 10% de los datos correspondiente a los 211 primeros. Como se puede observar en la figura 7, el modelo se ajusta bien a los datos de validación.

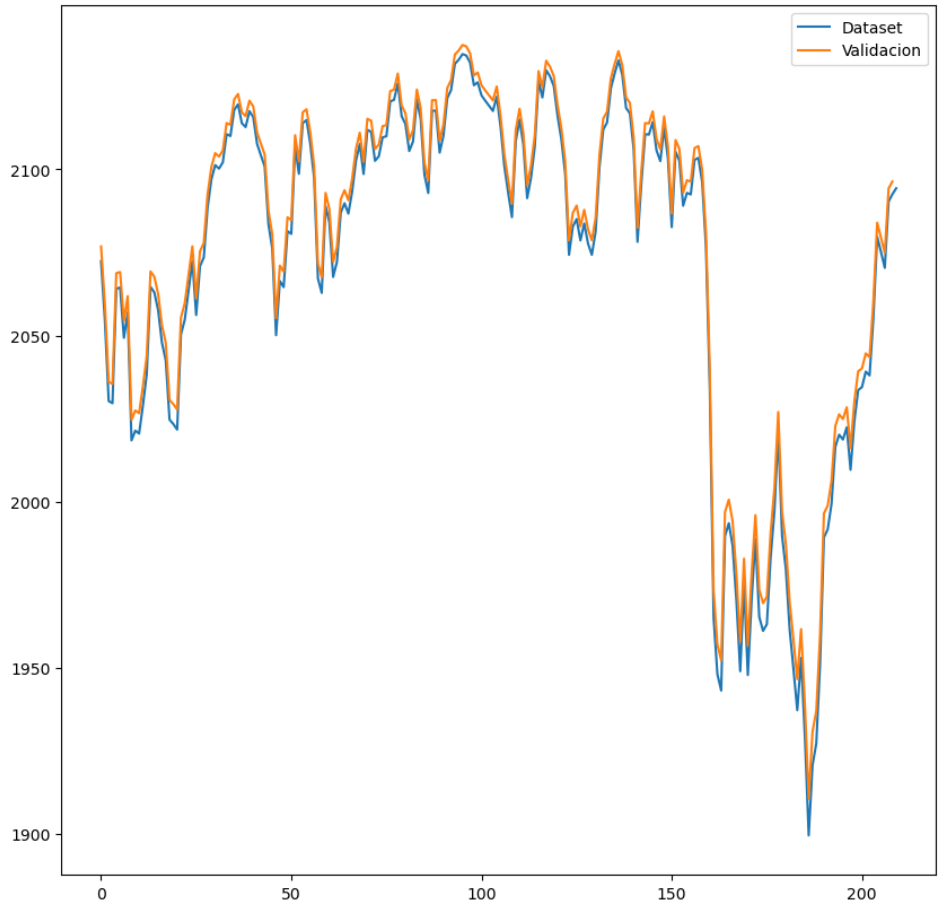

Figura 7 – Proceso de validación del modelo

Para la predicción del modelo se usa un set de datos de 36 valores correspondientes a los vales entre 02 al 31 de mayo de 2023, descargados de (Yahoo, 2023). Como se puede observar en la figura 8, el modelo presenta un $R^2 \approx 0.9231$.

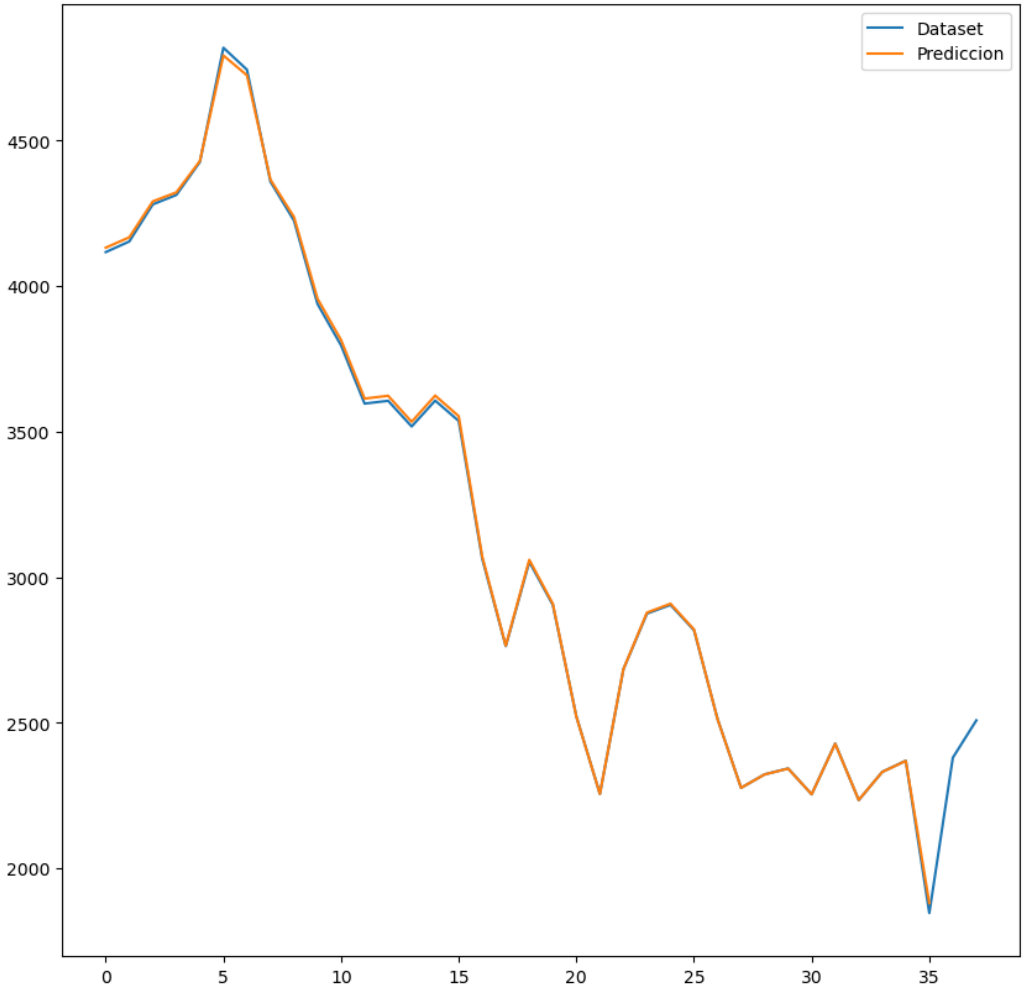

Figura 8 – Proceso de predicción del modelo

Las medidas de rendimiento del modelo propuesto (columnas con ´), se comparan en la tabla 4 con los modelos de la literatura. En este caso, se han seleccionado tres estudios para la comparación, debido a que utilizan modelos basados en neuronas LSTM, o en su defecto realizan predicciones del valor de las acciones de S&P500.

Tabla 4 – Comparación de resultados variables de modelos de la literatura vs propuestos

Autor y Titulo	RSME	MSE	R^2	RSME´	SME´	$R^{2´}$
(Montenegro & Molina, 2019)	N/A	N/A	0.8689	26.54	704.44	0.97701
(Mohan & Durairaj, 2022)	N/A	841.16				
(Pan & Li, 2022)	N/A	438.94				

3.4. Discusión.

De acuerdo con los resultados presentados en este estudio, se puede evidenciar una mejora en los modelos anteriores respecto a la predicción del valor de las acciones. Es importante notar que fue necesario trabajar con datos escalados para los cálculos, esto debido a la cantidad de datos con la que se trabajó, lo cual podía causar errores de cálculo al tener desviaciones muy grandes.

La manipulación de los hiperparámetros muestra mejorías en el modelo. En el caso de las funciones de activación en las capas previas a la salida es notable el uso la función “tanh”, sobre la función “relu”, ya que se comporta como se puede observar en la figura 9, formando una S, pero tomando valores entre -1 y 1.

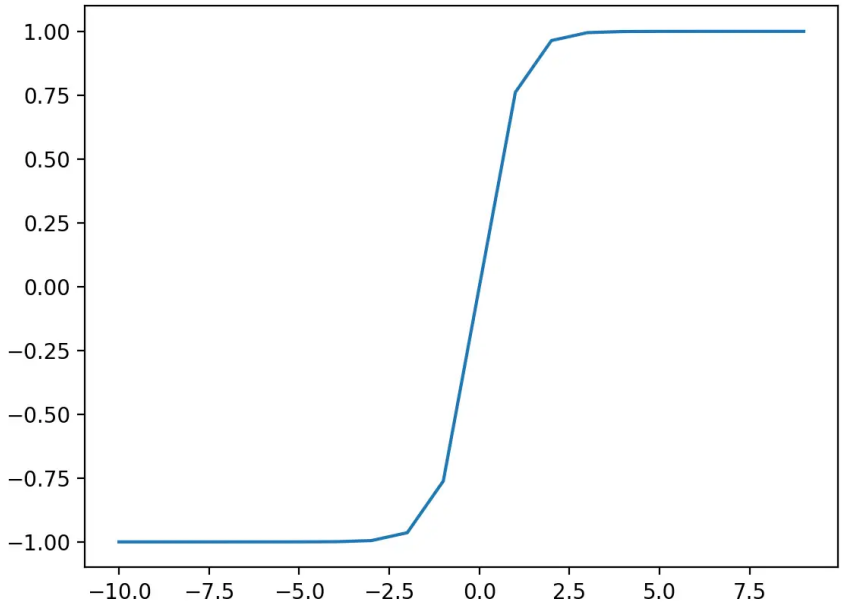

Figura 9 – Función de activación “tanh”

En el caso de la neurona de salida se utiliza una neurona de tipo denso, la cual está completamente conectada a la capa anterior y utiliza por defecto la función de activación “relu”, misma que se comporta como se puede ver en la figura 10

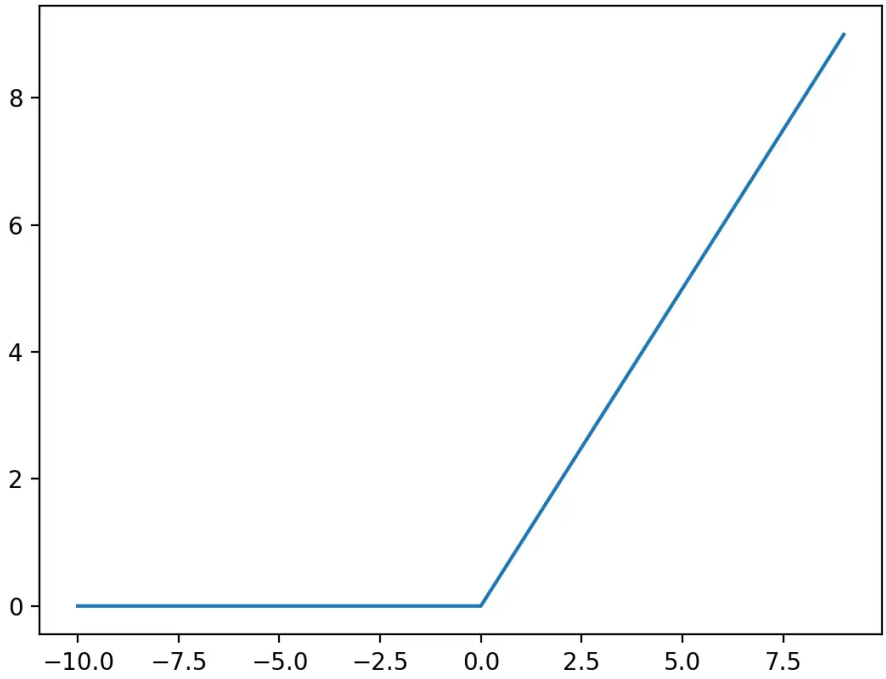

Figura 10 – Función de activación “relu”

La función “relu”, es menos susceptible a la desaparición de gradientes, problema que dificulta el entrenamiento de modelos profundos. Esto significa que si el valor de entrada (x) es negativo, se devuelve un valor 0.0; de lo contrario, se devuelve el valor.

4. CONCLUSIONES

Como se comenta en este trabajo, los valores de las acciones S&P500, y en general de los mercados bursátiles, son complejas puesto que dependen de múltiples factores internos y externos.

Para la predicción de los valores de las acciones S&P500 ha generado diferentes modelos utilizando técnicas de minería de datos, como se evidencia en los resultados de la revisión sistemática de la literatura. La mayoría de estos modelos de referencia utilizan técnicas de clasificación, y la minoría de regresión. Los modelos de regresión son importantes puesto que permiten realizar predicciones numéricas y, en consecuencia, calcular posibles ganancias/perdidas en un periodo de tiempo de posibles inversiones.

Es muy común también que los modelos de regresión se limiten a predecir los valores de entrenamiento y no extrapolan. En el caso de este trabajo se utilizan datos fuera de aquellos de entrenamiento, encontrando que el modelo realiza predicciones adecuadas.

Es importante anotar que el modelo utilizado se construye sobre la base de neuronas tipo LSTM alimentadas hacia adelante, que simulan una estructura tipo convolucional, que va

eliminando complejidades de los datos, y genera una salida con una neurona lo más simple posible.

Finalmente, se sugiere que se realicen investigaciones adicionales para explorar otras variantes de modelos LSTM y mejorar aún más la precisión de la predicción. Por ejemplo, estudiando la optimización del número de capas escondidas; igualmente, se podrían lograr mejores resultados, modificando el número de neuronas y de capas, así como afinando los hiperparámetros del modelo.

REFERENCIAS BIBLIOGRÁFICAS

Beyaz, E. (2019). Effective Stock Price Forecasting Using Machine Learning Techniques Whilst Accounting for the State of The Market. Manchester: School of Computer Science .

Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2015). Time Series Analysis: Forecasting and Control, 5th Edition. *Journal of Time Series Analysis Volume 37 Issue 5*, 709-711.

Carta, S., Corriga, A., Ferreira, A., & Podda, A. (2021). A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning. *Applied Intelligence 51*, 889–905.

Chacon, H., Kesici, E., & Najafirad, P. (2020). Improving Financial Time Series Prediction Accuracy Using Ensemble Empirical Mode Decomposition and Recurrent Neural Networks. *IEEE Access Volume: 8*, 117133 - 117145.

Comincioli, B. (1996). The Stock Market As A Leading Indicator: An Application Of Granger Causality. *University Avenue Undergraduate Journal of Economics: Vol. 1: Iss. 1, Article 1*, 5-6.

Diederik, K., & Lei Ba, J. (2015). ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. *ICLR 2015.* Cornell University.

ESPAÑOLA, R. A. (16 de May de 2023). *REAL ACADEMIA ESPAÑOLA: Diccionario de la lengua española, 23.ª ed., [versión 23.6 en línea]*. Obtenido de https://dle.rae.es

Gowthul, A., & Baulkani, S. (2019, August 1). Local and global characteristics-based kernel hybridization to increase optimal support vector machine performance for stock market prediction. *Knowledge and Information Systems*, p. 566.

Hajiagha, S., & Farahani, M. (01 de July de 2021). Forecasting stock price using integrated artificial neural network and metaheuristic algorithms compared to time series models. *Soft Computing - A Fusion of Foundations, Methodologies and ApplicationsVol. 25, No. 13*, págs. 8483–8513.

Huynh, H., Dang, L., & Duong, D. (2017). A New Model for Stock Price Movements Prediction Using Deep Neural Network. *SoICT '17: Proceedings of the 8th International Symposium on Information and Communication Technology* (págs. 57–62). New York: ACM.

Kang, J., & Lee, J. (2020). Effectively training neural networks for stock index prediction: Predicting the S&P 500 index without using its index data. *PLoS ONE 15(4): e0230635*.

Khushi, M., & Mukherjee, M. (2021). SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. *ASI Volume 4 Issue 1* , 4-18.

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews .* Keele University.

Kumar, D., Gupta, S., & Sehgal, P. (2014). Comparing gradient based learning methods for optimizing predictive neural networks. *2014 Recent Advances in Engineering and Computational Sciences (RAECS).* Chandigarh: IEEE.

Kumar, G., Jain, S., & Singh, U. (2020). Stock Market Forecasting Using Computational Intelligence: A Survey. *Archives of Computational Methods in Engineering*, pages1069–1101.

Le Guennec, A., Malinowski, S., & Tavenard, R. (2016). Data Augmentation for Time Series Classification using Convolutional Neural Networks. *Hal Science*, 1-9.

Lifeng, H., Tao, L., Jingjun, S., & Haiyang, Z. (2011). Research on the data warehouse and data mining techniques applying to decision assistant system. *2011 International Conference on Computer Science and Service System (CSSS).* Nanjing: IEEE.

Menacho, C. (2014). Comparación de los métodos de series de tiempo y redes neuronales. *Comparación de los métodos de series de tiempo y redes neuronales*.

Mohan, D., & Durairaj, M. (2022). A convolutional neural network based approach to financial time series prediction. *Special Issue on Deep Learning for Time Series Data*, 13319–13337.

Molina, M., & Montenegro, C. (2020). Improving the Criteria of the Investment on Stock Market Using Data Mining. *International Journal of Machine Learning and Computing, Vol. 10, No. 2*, 1-7.

Montenegro, C., & Molina, M. (2019). A DNN Approach to Improving the Short-Term Investment Criteria for S&P500 Index Stock Market. *ICEEG '19: Proceedings of the 3rd International Conference on E-commerce, E-Business and E-Government* (págs. 100–104). Lyon: Association for Computing Machinery.

Mostafa, F., Dillon, T., & Chang, E. (2017). *Computational Intelligence Applications to Option Pricing, Volatility Forecasting and Value at Risk.* Springer International Publishing.

Onibonoje, O., Djoussa, K., & Roantree, M. (2020). Analysis of Machine Learning Methods for Predicting Stock Prices. *CEUR, Vol 2771*, 1-12.

Pan, Y., & Li, Y. (2022). A novel ensemble deep learning model for stock prediction based on stock prices and news. *International Journal of Data Science and Analytics*, 139–149.

Qian, X. (2018, December 8). *Financial Series Prediction: Comparison Between Precision of Time Series Models and Machine Learning Methods.* Retrieved from https://arxiv.org/: https://arxiv.org/abs/1706.00948

Ryan, T. (02 de Sep de 2022). *Analytics Vidhya*. Obtenido de medium.com: https://medium.com/analytics-vidhya/lstms-explained-a-complete-technically-accurate-conceptual-guide-with-keras-2a650327e8f2

Saleh, A. (2018). *Forecasting Stock Index using Deep Learning and how it can be applied in the financial sector.* Retrieved from https://www.diva-portal.org/: https://www.diva-portal.org/smash/get/diva2:1272871/FULLTEXT01.pdf

Serrano-Cobos, J. (05 de 05 de 2016). Tendencias tecnologicas en internet: Hacia un cambio de paradigma. *Profesional de la informacion*, pág. 843.

Tebes, G., Peppino, D., Becker, P., & Olsina, L. (2019). Especificación del Modelo de Proceso para una Revisión Sistemática de Literatura. *XXII Conferencia Iberoamericana en Software Engineering (CIbSE'19)* (págs. 391-404). La Habana, Cuba: Curran Associates Publisher.

Wang, F. (2020). Predicting S&P 500 Market Price by Deep Neural Network and Enemle Model. *2020 International Conference on Energy Big Data and Low-carbon Development Management (EBLDM 2020).* GREENVILLE, NC: E3S Web of Conferences.

Yahoo. (10 de 05 de 2023). *Yahoo Finance*. Obtenido de Yahoo Finance: https://es-us.finanzas.yahoo.com/

Yamamoto, K., Yoshii, M., Kinoshita, F., & Touyama, H. (2020). Classification vs Regression by CNN for Handwashing Skills Evaluations in Nursing Education. *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC).* Fukuoka: IEEE.

Yang, X., Zhang, Z., Cui, X., & Cui, R. (2021). A Time Series Data Augmentation Method Based on Dynamic Time Warping. *2021 International Conference on Computer Communication and Artificial Intelligence (CCAI).* Guangzhou: IEEE.

Yin, J., & Si, Y. (2013). OBST-based segmentation approach to financial time series. *Engineering Applications of Artificial Intelligence*, 2581-2596.

Zhu, M., Shen, Y., & Angelova, M. (14 de October de 2022). *Clustering-enhanced stock price prediction using deep learning.* Obtenido de https://link.springer.com/article/10.1007/s11280-021-01003-0#citeas

ANEXOS

Anexo I – Estructura del modelo de LSTM Python

```python
#Modelo LSTM

# Definir una función de activación personalizada para la puerta de olvido
def custom_forget_gate_activation(x):
    return tf.keras.activations.sigmoid(x + 1.0)  # Ajustar el sesgo para mod


units = 40  # Neuronas de la primera capa oculta
activation = 'tanh'  # Función de activación

# Crear el modelo
model = Sequential()
model.add(LSTM(units, activation=activation, input_shape=(1,look_back), retur
               recurrent_activation=custom_forget_gate_activation))
model.add(LSTM(20, activation=activation, return_sequences=True))
model.add(LSTM(5, activation=activation))
model.add(Dense(1))  # Capa de salida

# Compilar el modelo
model.compile(optimizer='adam', loss='mean_squared_error')

model.fit(trainX,trainY,epochs=1000, batch_size=32, verbose=2)

# Imprimir un resumen del modelo
model.summary()
```

```
53/53 - 0s - loss: 9.4772e-05 - 64ms/epoch - 1ms/step
Epoch 59/1000
53/53 - 0s - loss: 9.5475e-05 - 78ms/epoch - 1ms/step
Epoch 60/1000
53/53 - 0s - loss: 9.8067e-05 - 63ms/epoch - 1ms/step
Epoch 61/1000
53/53 - 0s - loss: 9.5799e-05 - 78ms/epoch - 1ms/step
Epoch 62/1000
53/53 - 0s - loss: 9.0771e-05 - 63ms/epoch - 1ms/step
Epoch 63/1000
53/53 - 0s - loss: 9.0007e-05 - 63ms/epoch - 1ms/step
Epoch 64/1000
53/53 - 0s - loss: 8.7869e-05 - 80ms/epoch - 2ms/step
Epoch 65/1000
53/53 - 0s - loss: 8.5763e-05 - 92ms/epoch - 2ms/step
Epoch 66/1000
53/53 - 0s - loss: 9.1903e-05 - 79ms/epoch - 1ms/step
Epoch 67/1000
53/53 - 0s - loss: 8.5809e-05 - 63ms/epoch - 1ms/step
Epoch 68/1000
```