Predicción del Valor de Cierre Diario del Índice DAX con RNN tipo LSTM junto a Exploración de Modelos Similares para S&P 500 y HSI: Reporte Técnico.

A. Monroy-Azpeitia, ESFM IPN/CIC IPN
amonroy.azpeitia@gmail.com
Abril 22, 2024.

***Resumen*—Utilizando datos del valor de cierre diario del índice bursátil alemán DAX obtenidos de Yahoo! Finanzas se entrenó una red neuronal recurrente tipo LSTM con datos de 2010-01-18 a 2019-11-07 y evaluando el modelo con datos de 2019-11-08 a 2024-01-18 obteniendo un MSE en los datos de prueba menor al 0.01 seleccionando el mejor modelo tras una búsqueda de rendija para caracterizar los hiperparámetros del modelo utilizando una normalización del valor de cierre *Mínimo-máximo*. También se exploró el uso de arquitecturas similares con sus respectivos entrenamientos y búsqueda de hiperparámetros en los índices S&P500 y HSI para periodos de tiempo similares obteniendo para los 3 índices un MSE menor a 0.02.**

I. INTRODUCCIÓN

El análisis y predicción de los mercados financieros resulta un desafío complejo debido a la naturaleza volátil, no lineal y caótica [1] de los datos bursátiles. En este contexto, los índices bursátiles como el DAX, S&P 500 y HSI juegan un papel crucial, ya que reflejan el desempeño económico de las principales empresas en sus respectivas regiones. La capacidad de predecir con precisión los valores de cierre diario de estos índices puede proporcionar ventajas significativas tanto para los inversores como para los analistas financieros.

En los últimos años, las Redes Neuronales Recurrentes (RNNs) han emergido como una herramienta poderosa para el análisis de series temporales debido a su capacidad para capturar dependencias a largo plazo en los datos. Entre las diversas arquitecturas de RNNs, las Long Short-Term Memory (LSTM) han demostrado ser particularmente efectivas para la predicción de series temporales complejas, gracias a su diseño que mitiga el problema del desvanecimiento del gradiente y permite la retención de información relevante a lo largo de extensas secuencias.

En este reporte, se presenta el uso de Redes Neuronales Recurrentes con arquitectura LSTM para la predicción del valor de cierre diario del índice DAX. Además, se exploran modelos similares aplicados a los índices S&P 500, HSI y Nikkei 225. Utilizando PyTorch, una biblioteca de aprendizaje profundo ampliamente adoptada, se desarrollaron y entrenaron los modelos LSTM para cada uno de estos índices. En todos los casos, los modelos lograron un error cuadrático medio (MSE) inferior a 0.02, lo cual destaca la eficacia de las LSTM para la predicción de datos financieros.

Este trabajo no solo proporciona una visión detallada sobre el uso de LSTM en la predicción de índices bursátiles, sino que contextualiza teóricamente este tipo de arquitectura.

II. Redes Neuronales Recurrentes

Dentro del campo del aprendizaje automático clásico, una rama de la inteligencia artificial que se centra en descubrir patrones automáticos en datos observados sin requerir programación explícita, el Aprendizaje Profundo, también conocido como Deep Learning (DL), se distingue por su capacidad para aprender representaciones de datos mediante múltiples capas de procesamiento [12]. Este enfoque se basa en algoritmos que utilizan una serie de pasos, o capas, de computación.En particular, las Redes Neuronales Recurrentes (RNN) son una familia de modelos de DL que se especializan en tratar datos secuenciales, como series de tiempo, texto o secuencias de acciones. Las RNN pueden entenderse conceptualmente como grafos computacionales, donde cada paso de tiempo en la secuencia representa un nodo en el grafo. Este enfoque permite capturar relaciones temporales y de dependencia a lo largo de la secuencia:

$$\mathcal{X} = \left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(\tau)}\right\} = \left\{\mathbf{x}^{(t)}\right\}_{t=1}^{\tau} \tag{1}$$

Para ilustrar este concepto, consideremos la forma clásica de un sistema dinámico expresado en términos de una ecuación recurrente, donde el estado del sistema en un paso de tiempo t depende del paso anterior y de la entrada al modelo. Este tipo de formulación es fundamental en el diseño y comprensión de las RNN, ya que refleja su capacidad para modelar la evolución temporal de

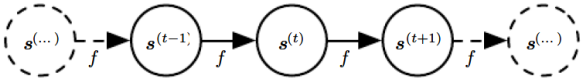


Figura 1. Sistema dinámico clásico descrito por (2) visualizado como n grafo computacional desplegado. Imagen tomada de [12].

datos secuenciales mediante iteraciones recursivas [12]. consideremos la forma clásica de un sistema dinámico:

$$\mathbf{s}^{(t)} = f(\mathbf{s}^{(t-1)}; \theta) \tag{2}$$

donde $\mathbf{s}^{(t)}$ indica el estado del sistema. Para un número finito de pasos τ el grafo puede ser desplegado $\tau - 1$ veces, por ejemplo la ecuación (2) con $\tau = 3$:

$$\begin{aligned} \mathbf{s}^{(3)} &= f(\mathbf{s}^{(2)}; \theta) \\ &= f(f(\mathbf{s}^{(1)}; \theta); \theta) \end{aligned} \tag{3}$$

al considerar el sistema dinámico impulsado por una señal externa $\mathbf{x}^{(t)}$:

$$\mathbf{s}^{(t)} = f(\mathbf{s}^{(t-1), \mathbf{x}^{(t)}, \theta}) \tag{4}$$

el estado contendrá toda la información del pasado. Muchas RNN definen sus estados ocultos de una forma silimar:

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}, \theta) \tag{5}$$

en particular para tareas que requieran realizar predicciones del futuro del pasado, se suele utilizar a $\mathbf{h}^{(t)}$ como un tipo tipo de resumen de pérdida de los aspectos del pasado relevantes para la tarea [12]. Se puede representar la recurrencia desplegada después de t pasas como:

$$\begin{aligned} \mathbf{h}^{(t)} &= g^{(t)}(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}, \cdots, \mathbf{x}^{(2)}, \mathbf{x}^{(1)}) \\ &= f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \theta) \end{aligned} \tag{6}$$

logrando así que la función g tome en consideración todo el pasado, teniendo siempre el mismo tamaño de entrada y la posibilidad de usar siempre la misma función de transición con los mismos parámetros en cada paso de tiempo. Con estas ideas es posible crear modelos capaces de realizar predicciones con base en datos secuenciales. Las RNN *estándar* padecen de un desvanecimiento de gradiente en su entrenamiento [17] que se traduce en dificultades para aprender dependencias a largo plazo en secuencias de datos.

III. Redes Neuronales Recurrentes tipo Long-Short-Term-Memory

Una propuesta solución al desvanecimiento del gradiente en RNN se logra con la implementación de una Red Neuronal tipo Long-Short-Term-Memory (LSTM) donde en cada celda (bloque de la RNN) se le añade una celda de control (gate) que permite tener control de la cantidad de información que se conserva de los pasos anteriores [14]. Se consideran 4 celdas de control que

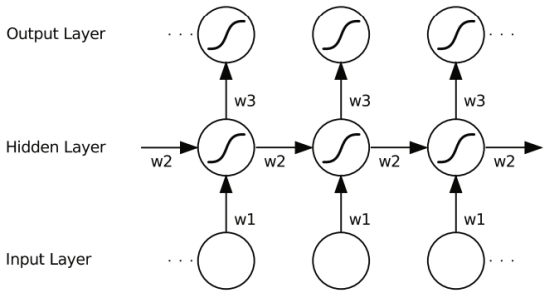


Figura 2. Ejemplo de RNN desplejaga donde cada nodo representa una capa de la red en un solo paso de tiempo. Los pesos se reutilizan en cada paso de tiempo. Imagen tomada de [17]

toman por entradas el estado oculto un paso atrás y la entrada actual del modelo:

- *Forget gate:* Controla que tanto olvidar o recordar del estado previo de la celda asociandolo a un valor entre 0 y 1:

$$f^{(t)} = \sigma(W_{if}\mathbf{x}^{(t)} + b_{if} + W_{hi}\mathbf{h}^{(t-1)} + b_{hf}) \tag{7}$$

- *Input gate:* Asociada a la cantidad de información de entrada que debe almacenar la celda:

$$i^{(t)} = \sigma(W_{ii}\mathbf{x}^{(t)} + b_{ii} + W_{hi}\mathbf{h}^{(t-1)} + b_{hi}) \tag{8}$$

- *Output gate*: Asociada a la cantidad de información de la memoria a largo plazo se debe utilizar para calcular la salida:

$$o^{(t)} = \sigma(W_{io}\mathbf{x}^{(t)} + b_{io} + W_{ho}\mathbf{h}^{(t-1)} + b_{ho}) \tag{9}$$

- *Candidate gate*: El candidato a salida de la celda:

$$c^{(t)} = \tanh(W_{ic}\mathbf{x}^{(t)} + b_{ic} + W_{hc}\mathbf{h}^{(t-1)} + b_{hc}) \tag{10}$$

con estos valores se calcula el **estado de la celda**:

$$\mathbf{C}^{(t)} = f^{(t)} \odot \mathbf{C}^{(t-1)} + i^{(t)} \odot c^{(t)} \tag{11}$$

donde se considera el producto de hadamard elemento a elemento, teniendo en el primer término de la ecuación que tanto hay que recordar del estado pasado y cuales elementos de dicho vector son relevantes, mientras el segundo término nos indica que tanto del candidato se pasará a la celda siguiente.
Para el **estado oculdo de la celda**:

$$\mathbf{h}^{(t)} = o^{(t)} \odot \tanh(\mathbf{C}^{(t)}) \tag{12}$$

donde el *output gate* pesa que tanta información se transmitirá a la siguiente celda.

IV. ÍNDICE BURSTIL DAX

El Índice DAX, administrado por la Bolsa de Fráncfort, sirve como el punto de referencia principal para el mercado bursátil alemán. Evalúa el rendimiento de las 40(30) empresas más grandes y más líquidas de Alemania, que representan aproximadamente el 80 % de

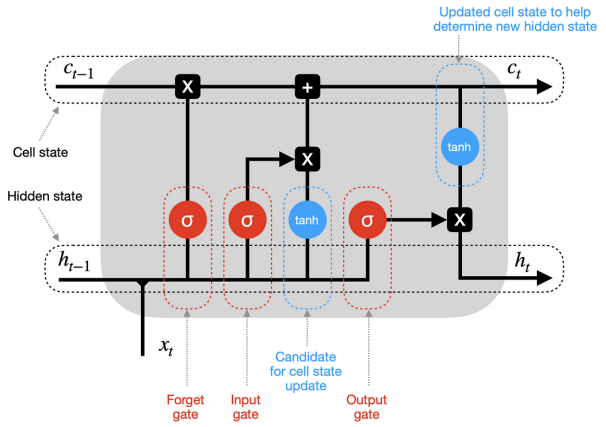


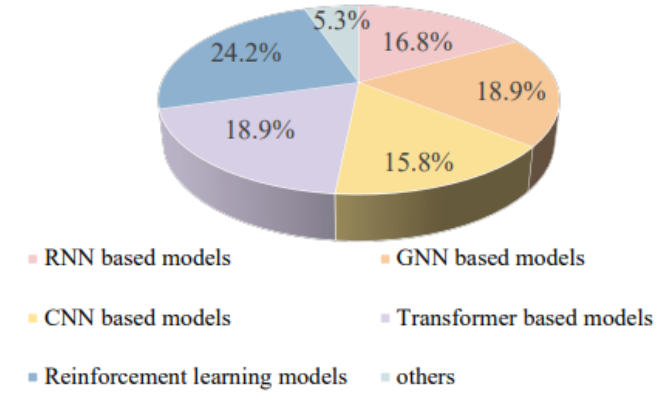
Figura 3. Celda de una RNN tipo LSTM

la capitalización de mercado de las empresas alemanas listadas. Ampliamente reconocido como un indicador clave, refleja el rendimiento general de la economía alemana [3]

La composición del índice se determina sobre la base de un conjunto claro y públicamente disponible de reglas. Los criterios básicos para incluir empresas en el DAX son los siguientes: una cotización existente en el Mercado Regulado de la FWB (Bolsa de Fráncfort), negociación continua en Xetra®, flotación libre mínima del 10 %, sede legal o sede operativa en Alemania, publicación oportuna de Informes Financieros Anuales auditados, Informes Financieros semestrales y Declaraciones Trimestrales.

Para ser incluida en el DAX [2], una empresa que aún no es un componente del índice debe cumplir con el siguiente requisito mínimo de liquidez de la FWB: volumen mínimo de libros de órdenes en los últimos 12 meses de 1 mil millones de euros o una tasa de rotación del 20 %. Una empresa que ya es un componente del índice debe tener un volumen mínimo de libros de órdenes de la FWB en los últimos 12 meses de al menos 0,8 mil millones de euros o mostrar una tasa de rotación del 10 %. Además, las empresas que no estén actualmente en el DAX al momento de la compilación de la lista de clasificación deben mostrar EBITDA positivo para los dos años fiscales más recientes. La selección de los componentes del índice se basa en la capitalización de mercado de flotación libre. La composición del índice se revisa trimestralmente según las reglas de Salida Rápida y Entrada Rápida y semestralmente según las reglas de Salida Regular y Entrada Regular.

Además, es importante destacar que el peso del índice de una acción individual está limitado al 10 %. Asimismo, el número de empresas en el DAX se incrementó de 30 a 40 a partir del 20 de septiembre de 2021.Por último, es relevante mencionar que el Grupo Deutsche Börse ha estado calculando el DAX desde el 1 de julio de 1988. En septiembre de 2019, STOXX Ltd. comenzó a administrar el DAX. Aunado a la predicción de valor de cierre del



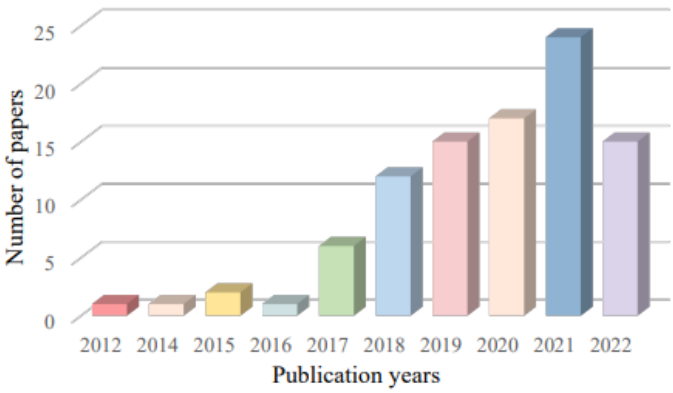


Figura 4. Publicaciones en revistas de prestigio* del uso de Deep Learning para la predicción de comportamiento de series de tiempo financieras [4]

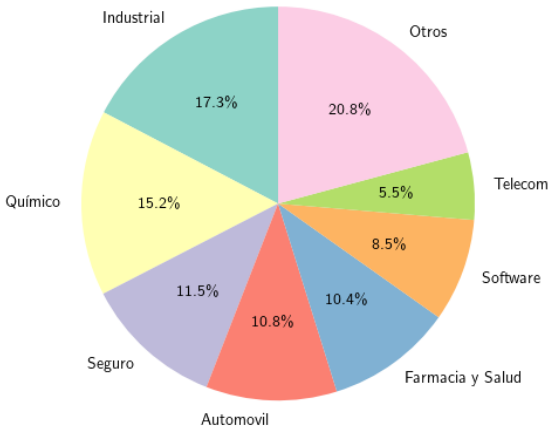


Figura 5. Distribución del índice DAX en 2022 [3]

Cuadro I
Datos descargados de Yahoo Finanzas e Investing para el índice DAX

	Yahoo Finanzas	Investing
Fecha inicio	2010-01-04	04.01.2010
Fecha final	2024-02-08	08.02.2024
Num. Datos	3585	3585
Días Faltantes	1563	1563
Columnas de los datos	Date,Open, High, Low Low,Close Adj_Close	Fecha, Cierre Apertura Max, Min Vol,% var

índice DAX, se entrenaron otros modelos LSTM para la predicción de otros índices. Los índices considerados fueron tomados de Yahoo Finanzas [10]:

- **Standard and Poor 500 (S&P 500):**
  Índice bursátil que mide el rendimiento de las 500 empresas más grandes que cotizan en las bolsas de valores de Estados Unidos. Es un indicador clave del desempeño del mercado de valores estadounidense y de la economía en general, ya que abarca una amplia variedad de sectores y refleja la salud financiera de las principales empresas del país.
- **Hang Seng Index (HSI)**
  Mide el rendimiento de las principales empresas que cotizan en la Bolsa de Valores de Hong Kong. Incluye las 50 empresas más grandes y representativas de la bolsa, abarcando una amplia gama de sectores. El HSI es un indicador clave de la salud económica y del mercado financiero de Hong Kong y se utiliza para evaluar el desempeño de las empresas más importantes de la región.

V. ANN PARA PREDICCIÓN DE ÍNDICES FINANCIEROS

Incorporando la información de las obras referenciadas [4], [5] y [6], es evidente que la fluctuación de los precios de las acciones está influenciada por una compleja variedad de factores, lo que contribuye a la incertidumbre inherente y la variabilidad del mercado de valores. Esta complejidad motiva el uso de técnicas de aprendizaje automático para diversas tareas de predicción del mercado de valores, incluida la predicción del movimiento de las acciones, la predicción de los precios de las acciones, el manejo de portafolios y las estrategias de negociación. Estos métodos aprovechan tanto las teorías económicas como los algoritmos computacionales para predecir el comportamiento futuro del mercado caótico.
Es importante señalar las limitaciones de los métodos estadísticos tradicionales, que dependen en gran medida de suposiciones iniciales y pueden carecer de adaptabilidad a las dinámicas cambiantes del mercado. Por el contrario, los enfoques de aprendizaje automático , en particular de deep learning que han adquirido relevancia en años recientes, ver Figura 2, enfrentan su propio conjunto de desafíos como la interpretabilidad limitada, la dependencia del rendimiento en la selección manual de características, susceptibilidad al sobreajuste, la selección de hiperparámetros y el entrenamiento del modelo, particularmente el uso de técnicas de aprendizaje profundo, deep learning, ofrecen la capacidad de descubrir patrones intrincados dentro de datos de mercado de valores altamente no estructurados.
Trabajos académicos a destacar que implementen Redes Neuronales Recurrentes tipo LSTM destacan:

- ***Comparison of machine learning methods for financial time series forecasting at the examples of over 10 years of daily and hourly data of DAX 30*, Artículo de investigación [7]**
  En este trabajo:
  - Utilizan únicamente los precios el índice DAX y S&P500 junto a valores que pueden obtener de la misma serie de tiempo para la predicción de la dirección de la serie de tiempo
  - Consideran datos con un periodo de más de 10 años (02/01/2004-06/03/2015).
  - Comparan KNN vs SVR vs LSTM vs Bagging

  Llegando a las conclusiones:
  - Los 3 modelos y sus variantes tiene buenos resultados.
  - Usar KPCA en SVR y KNN mejora los modelos.
  - KNN es el que tiene el mejor performance.
- ***Machine Learning for Financial Market Forecasting*, Tesis de Maestria [8]**
  En dicho trabajo:
  - Analizan el índice S&P500 y artículos de noticias para el análisis de sentimientos.
  - Utilizan datos de 12 y 6 meses.
  - Comparan BERTvsFinBERT (análisis de sentimientos), LSTM vs Regresión logística vs SVM.

  Llegando a las conclusiones:
  - Utilizando FinBERT+SVM, se obtiene un mejor Accuracy,Precision y Recall.
  - LSTM fue mejor en predecir acciones que índices. BERT y FinBERT mantuvieron la precisión del modelo.
  - LSTM favorece el performance un 20
  - Literatura limitada para optimización de LSTM.
  - Modelos LSTM tienen mejor rendimiento con menos variables características.
  - Modelos LSTM requieren de un extenso tuneo de hiperparámetros y composición del modelo

OBTENCIÓN DE DATOS

A lo largo del estudio se consideran 2 conjuntos de datos en un formato de archivo CSV, donde se ha realizado una limpieza de datos remplazando valores nulos con fecha registradas en la serie de tiempo ignorando las Fechas no registradas.

	DAX	S&P 500	HSI
Epocas	500	320	500
Capas	1	1	1
Tasa de Aprendizaje	0.00045	0.0002	0.00045
Tamaño del Estado Oculto	4	10	4
Ventana Temporal	10	15	10
Escalamiento	Min-Max	Min-Max	Min-Max

Cuadro II
HIPERPARÁMETROS DE LOS MEJORES MODELOS

Yahoo Finanzas: Yahoo Finance ofrece cotizaciones de acciones gratuitas, noticias actualizadas, recursos de gestión de carteras, datos de mercados internacionales, interacción social y tasas hipotecarias [10]

Investing: Plataforma de mercados financieros, ofrece tiempo real, cotizaciones, gráficos y herramientas financieras en 44 ediciones internacionales.Es una de las tres principales webs financieras del mundo según SimilarWeb y Alexa. Da acceso a más de 300.000 instrumentos financieros incluyendo bursátiles, materias primas, criptomonedas, índices, divisas, bonos, entre otros. [11]
En ambas plataformas se descargó un archivo CSV y se realizó un análisis y limpieza simple de ellos con la libreria *Pandas* de Python, generando los CSV finales con las caracteristicas mostradas en I. Para el índice S&P500 se consideraron datos desde 2010-01-26 hasta 2023-12-29 y para el índice HSI 2010-01-18 hasta 2023-12-29.

VI. Experimentos Computacionales y Resultados

Utilizando el framework PyTorch [18], se desarrollaron tres clases especializadas para el modelo. La clase *TimeSeries_Dataset* hereda de *torch.Dataset*, *LSTM* hereda de *torch.nn.Module*, y se creó una clase específica para el entrenamiento del modelo llamada *LSTM_Model_Training*. En todos los casos, se realizó una búsqueda en rejilla para la selección de hiperparámetros como la tasa de aprendizaje, el tipo de normalización de los datos, las épocas, las capas de la red LSTM y la ventana temporal, utilizando un conjunto de entrenamiento que empleaba el valor de precio de cierre del índice para predecirlo. La función de costo utilizada fue el error cuadrático medio (MSE), y en el mejor modelo (escalamiento mínimo-máximo), el MSE en todos los datos de entrenamiento para los 3 índices fue menor a 0.01.
Los mejores hiperparámetros de cada modelo asociado a los 3 índices se muestra en la Tabla II. Se muestra la función de costo a lo largo de las epocas para cada modelo, junto a su desempeño en los datos de entrenamiento y prueba en las Figuras 7 a 14.

VII. Discusión y Conclusión

El uso de Redes Neuronales Recurrentes para la predicción de los índices DAX, S&P500 y HSI para periodos de tiempo tan largos en su precio de cierre diario resulta complicado y demandante tanto en la selección de hiperparámetros como en el entrenamiento. Es importante resaltar que los mejores modelos aquí mostrados contrastan demasiado con los demás considerados en los demás hiperparámetros mostrando el efecto de la normalización en los datos y en número de epocas, mostrando resultados y comportamientos abruptos en la evaluación en los datos de entrenamiento con diferencia de tan solo 10 épocas a pesar de tener una tasa de aprendizaje baja en todos los casos, para más detaller ver los resultados de los entrenamientos para distintos hiperparámetros en [19]. En este trabajo se logró predecir con resultados aceptables verificables al comparar las predicciones y valores reales en la serie de tiempo en los datos de prueba para los 3 índices aquí estudiados teniendo un MSE menor a 0.01 en todos los casos aquí considerados.

Referencias

[1] Jeffrey D. Scargle. An introduction to chaotic and random time series analysis. Imaging Systems and Technology Volume1. (1989)
[2] Stoxx, DAX The German equity benchmark – for over 30 years,https://qontigo.com/index/dax/
[3] Deutsche Börse , The DAX index universe index universe.
[4] J. Zou, Q. Zhao, Y. Jiao, H. Cao, Y. Liu, Q. Yan, E. Abbasnejad, L. Liu, J. Qinfeng Shi (2023), *Stock Market Prediction via Deep Learning Techniques: A Survey* , arXiv:2212.12717.
[5] A. Thakkar, K. Chaudhar (2021). *A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions*, Expert Systems with Applications,Volume 177 114800
[6] W. Jiang (2021), Applications of deep learning in stock market prediction: Recent progress . Expert Systems With Applications 184, 11553
[7] .Ersan, C. Nishioka & A. Scherp (2019), Comparison of machine learning methods for financial time series forecasting at the examples of over 10 years of daily and hourly data of DAX 30 and SP 500, J Comput Soc Sc 3, 103–133
[8] J. Johnson (2023), Machine Learning for Financial Market Forecasting. Master's thesis, Harvard University Division of Continuing Educatio
[9] D.Ersan, C. Nishioka & A. Scherp (2019), *Comparison of machine learning methods for financial time series forecasting at the examples of over 10 years of daily and hourly data of DAX 30 and S&P 500*, J Comput Soc Sc 3, 103–133.
[10] Yahoo Finance, About; https://www.linkedin.com/company/yahoo-finance/about/
[11] Investing, About us; https://mx.investing.com/about-us/
[12] I. Goodfellow, Y. Bengio & A. Courville, *Deep Learning*, MIT Press (2016)
[13] D.V. Godoy, *Deep Learning with Pytorch, Volumen 1: Fundamentals* (2021)
[14] P. Cantoral, *LSTM: Todo lo que necesitas saber* [Video],Youtube https://www.youtube.com/watch?v=f6PaCo-NfJA&t=735s (2023)
[15] M. Nielsen, *Neural Networks and Deep Learning* [Libro],http://neuralnetworksanddeeplearning.com (2019)
[16] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York (2006)
[17] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer-Verlag Berlin Heidelberg (2012)
[18] Paszke, Adam and Gross, Sam and Chintala, Soumith and Chanan, Gregory and Yang, Edward and DeVito, Zachary and Lin, Zeming and Desmaison, Alban and Antiga, Luca and Lerer,*Automatic differentiation in PyTorch* (2017)
[19] https://github.com/AzpMon/DAX_Index_Forecasting-with-ANN

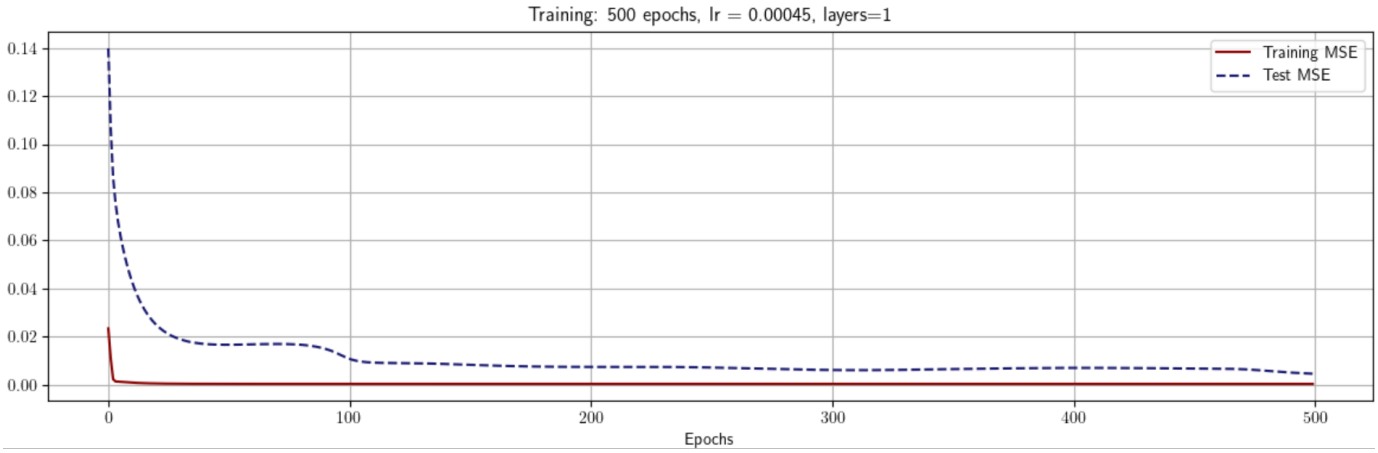


Figura 6. Error cuadrático medio del mejor modelo LSTM para la predicción del índice bursátil DAX

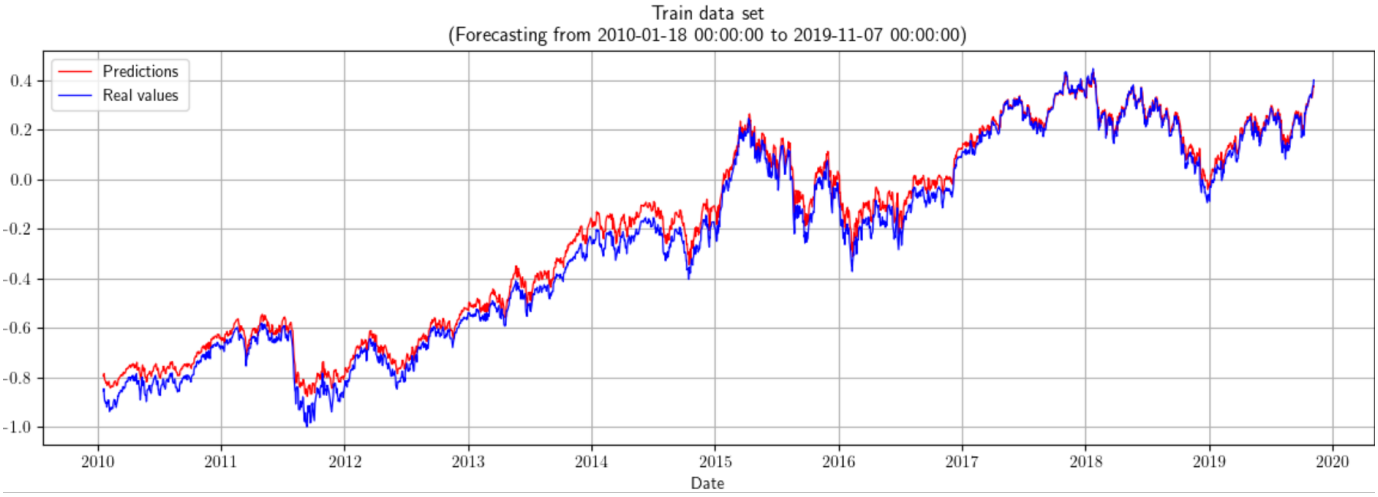


Figura 7. Predicciones del modelo LSTM en los datos de entrenamiento.

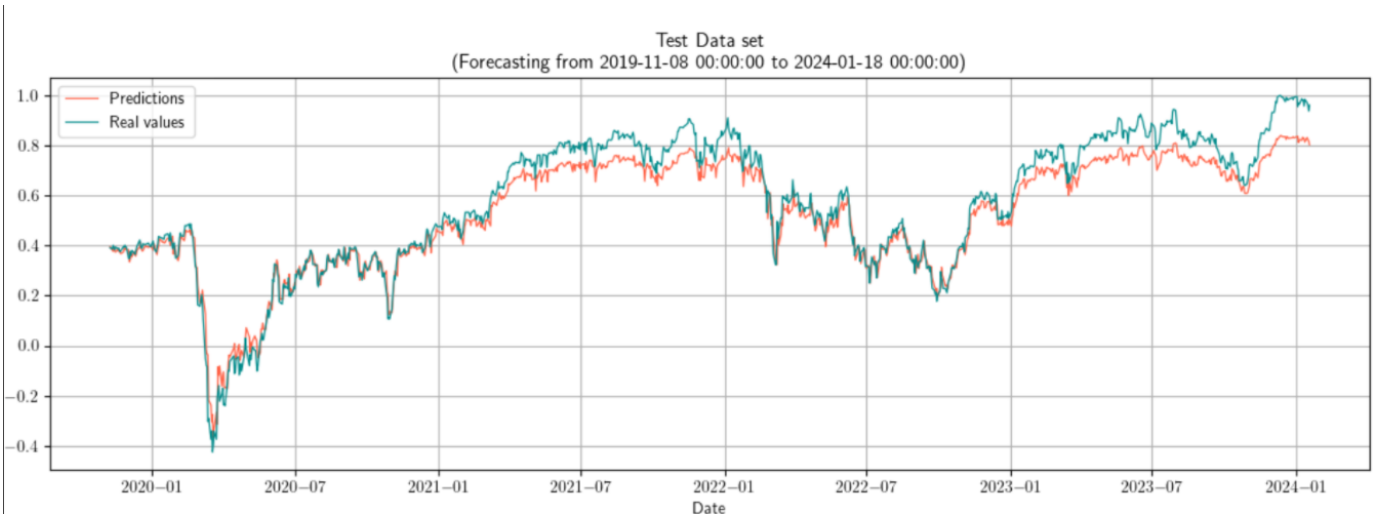


Figura 8. Predicciones del modelo LSTM en los datos de prueba del índice DAX.

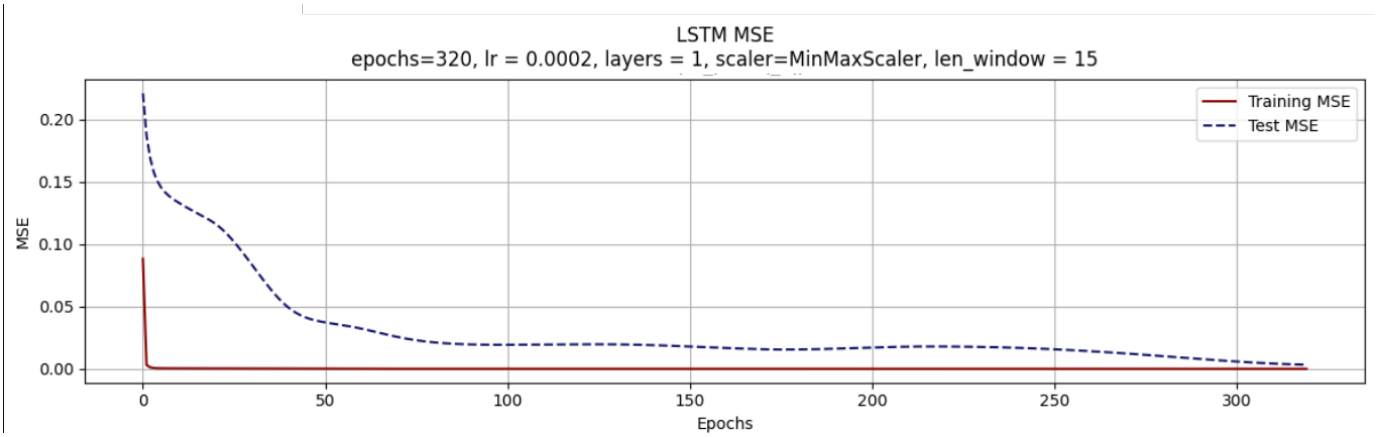


Figura 9. Error cuadrático medio del mejor modelo LSTM para la predicción del índice bursátil S&P500

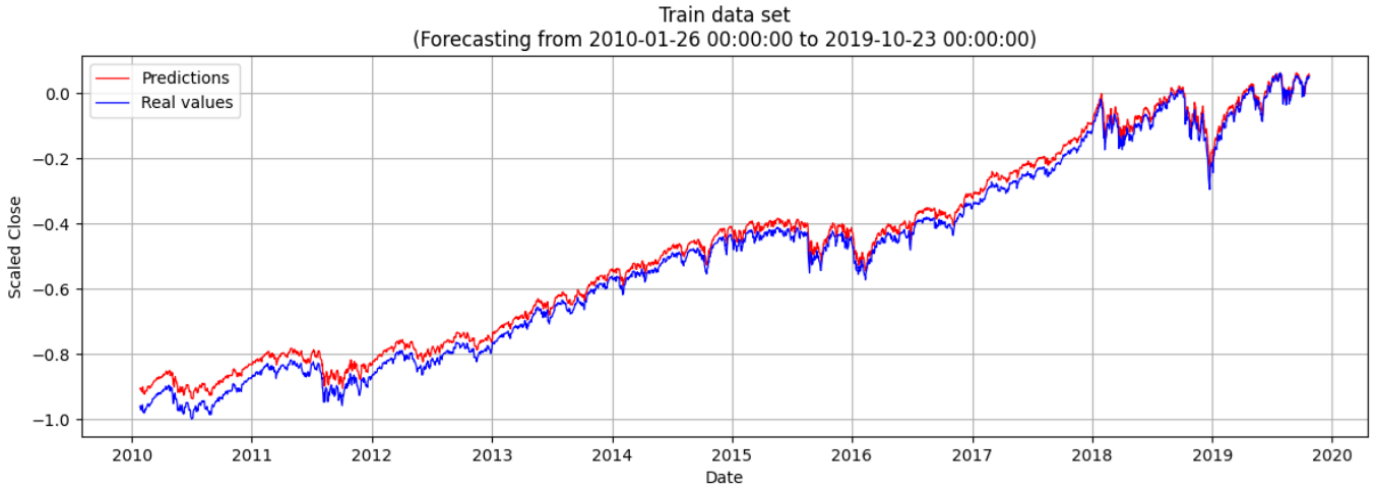


Figura 10. Predicciones del modelo LSTM en los datos de prueba del índice S&P500.

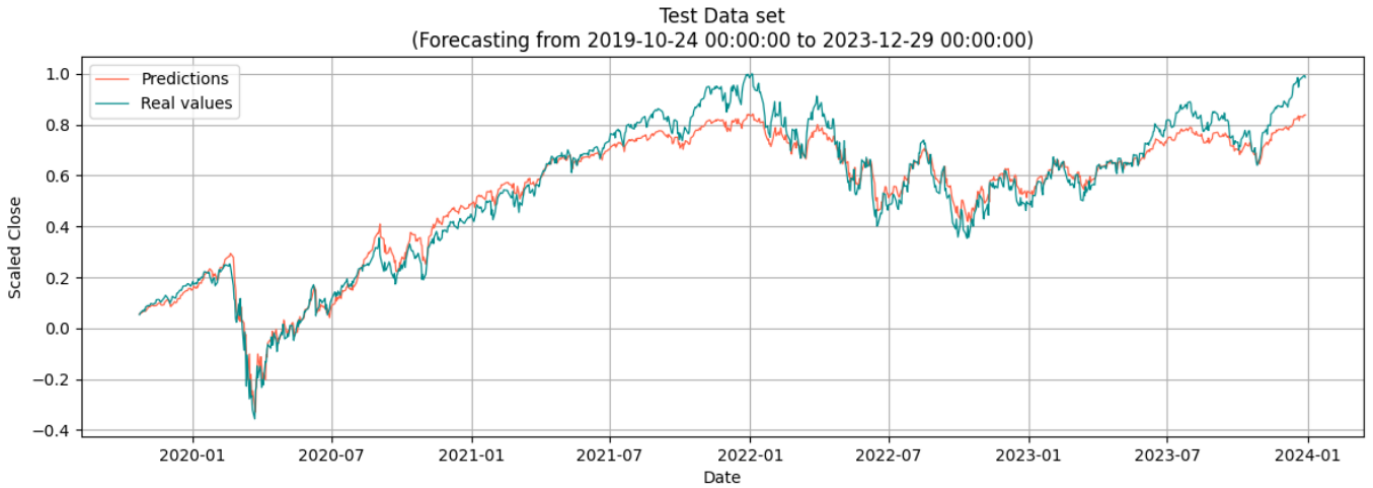


Figura 11. Predicciones del modelo LSTM en los datos de prueba del índice S&P500.

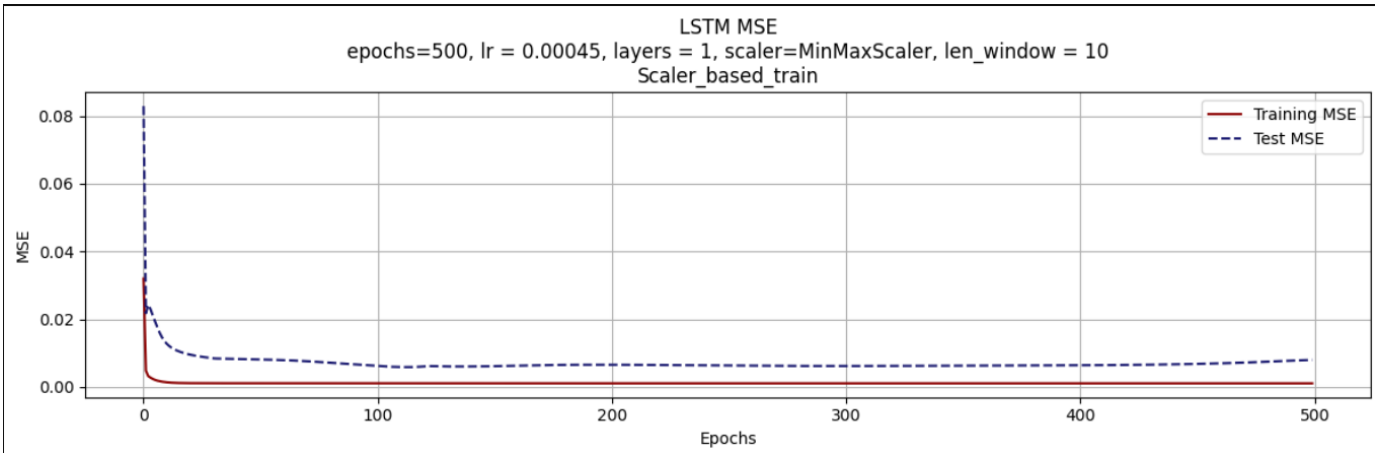


Figura 12. Error cuadrático medio del mejor modelo LSTM para la predicción del índice bursátil HSI

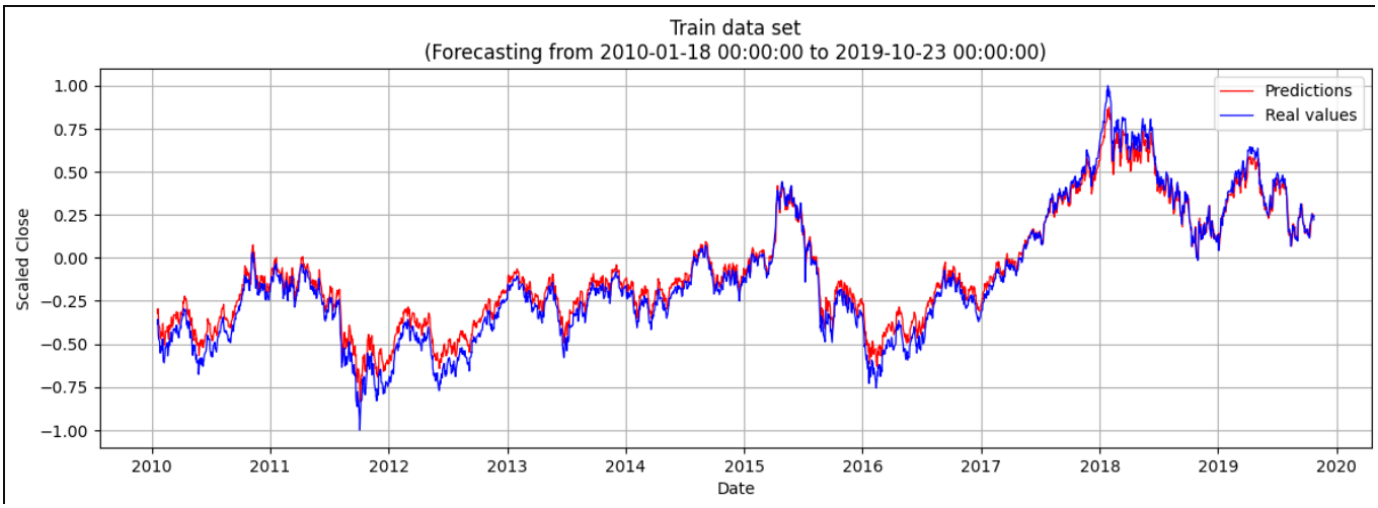


Figura 13. Predicciones del modelo LSTM en los datos de prueba del índice HSI

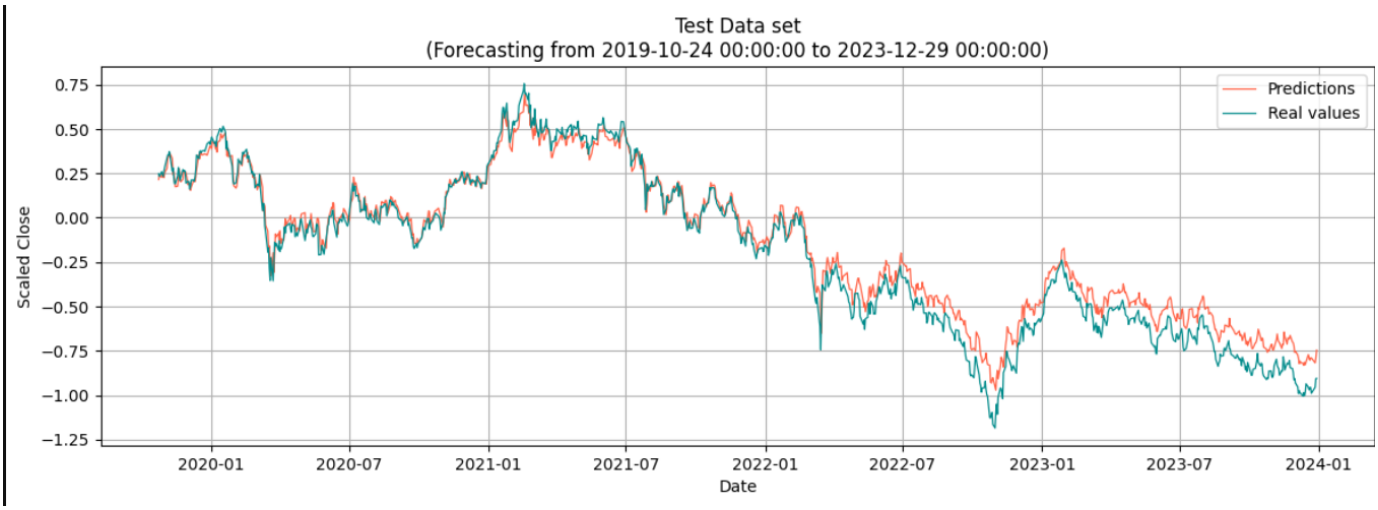


Figura 14. Predicciones del modelo LSTM en los datos de prueba del índice HSI