Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Ingeniería de Sistemas e Informática
Unidad de Posgrado

Series de tiempo y aprendizaje automático para el pronóstico de precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York

TESIS

Para optar el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software

AUTOR

Roberto Carlos EVARISTO BRONCANO

ASESOR

Dr. José Alfredo HERRERA QUISPE

Lima, Perú

2025



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Evaristo, R. (2025). *Series de tiempo y aprendizaje automático para el pronóstico de precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Roberto Carlos Evaristo Broncano
Tipo de documento de identidad	DNI
Número de documento de identidad	45829401
URL de ORCID	https://orcid.org/0000-0001-9010-3641
Datos de asesor	
Nombres y apellidos	José Alfredo Herrera Quispe
Tipo de documento de identidad	DNI
Número de documento de identidad	40362859
URL de ORCID	https://orcid.org/0000-0002-8207-9714
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Ciro Rodriguez Rodriguez
Tipo de documento	DNI
Número de documento de identidad	06020241
Miembro del jurado 1	
Nombres y apellidos	Ivan Carlo Petrlik Azabache
Tipo de documento	DNI
Número de documento de identidad	10140461
Miembro del jurado 2	
Nombres y apellidos	Rosa Menendez Mueras
Tipo de documento	DNI
Número de documento de identidad	10246770
Miembro del jurado 3	
Nombres y apellidos	
Tipo de documento	

Número de documento de identidad	
Datos de investigación	
Línea de investigación	C.0.3.22. Ingeniería de software
Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Edificio: Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Centro poblado: No aplica Urbanización: No aplica Manzana y lote: No aplica Calle: Avenida Carlos Germán Amézaga, número 375 Latitud: −12.0565 Longitud: −77.0844
Año o rango de años en que se realizó la investigación	2023 - 2024
URL de disciplinas OCDE	Otras ingenierías y tecnologías https://purl.org/pe-repo/ocde/ford#2.11.02

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Vicedecanato de Investigación y Posgrado
Unidad de Posgrado

SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO DE MAGÍSTER EN INGENIERÍA DE SISTEMAS E INFORMÁTICA CON MENCIÓN EN INGENIERÍA DE SOFTWARE

A los veintiséis (26) días del mes de septiembre de 2025, siendo las 12:00 pm., se reunieron en el Auditorio Alfredo Celso Alva Bravo, el Jurado de Tesis conformado por los siguientes docentes:

Dr. Ciro Rodriguez Rodriguez (Presidente)
Dr. Ivan Carlo Petrlik Azabache (Miembro)
Mg. Rosa Menendez Muera (Miembro)
Dr. José Alfredo Herrera Quispe (Miembro Asesor)

Se inició la Sustentación invitando al candidato a Magíster ***Roberto Carlos Evaristo Broncano,*** *para que realice la exposición oral de la tesis para optar el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software, siendo la Tesis intitulada:*

"SERIES DE TIEMPO Y APRENDIZAJE AUTOMÁTICO PARA EL PRONÓSTICO DE PRECIOS Y RENDIMIENTOS DE LAS ACCIONES DE UNA EMPRESA PERUANA EN LA BOLSA DE VALORES DE NUEVA YORK".

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, habiendo obtenido la siguiente calificación:

... Dieciséis ... (16) ... Bueno ...

Por tanto, el presidente del Jurado, de acuerdo con el Reglamento General de Estudios de Posgrado, otorga al Bachiller ***ROBERTO CARLOS EVARISTO BRONCANO*** *el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software.*

Siendo las 13:32 *horas, el presidente del Jurado de Tesis, da por concluido el acto académico de Sustentación de Tesis.*

Dr. Ciro Rodriguez Rodriguez
(Presidente)

Dr. Ivan Carlo Petrlik Azabache
(Miembro)

Mg. Rosa Menendez Mueras
(Miembro)

Dr. José Alfredo Herrera Quispe
(Miembro Asesor)

UNIVERSIDAD NACIONAL MAYOR DE
SAN MARCOS
Universidad del Perú, Decana de América

CERTIFICADO DE SIMILITUD

Yo, José Alfredo Herrera Quispe en mi condición de asesor acreditado con Dictamen N° 000063-2025-UPG-VDIP-FISI/UNMSM de la tesis, cuyo título es “Series de Tiempo y Aprendizaje Automático para el Pronóstico de Precios y Rendimientos de las Acciones de una Empresa Peruana en la Bolsa de Valores de Nueva York”

presentado por el egresado Roberto Carlos Evaristo Broncano para optar el grado de Magister en Ingeniería de Sistemas e Informática con Mención en Ingeniería de Software,

CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el 11 % de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional.**

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado/ título/ especialidad correspondiente.

Lima, 26 de FEBRERO de 2025



Firma del Asesor

Nombres y Apellidos del asesor: José Alfredo Herrera Quispe

DNI 40362859

DEDICATORIA

Este trabajo de tesis está dedicado a,
las personas más importantes en mi vida,
mis padres y Yuli.
Gracias a su apoyo y paciencia.
Hoy puedo alcanzar una nueva meta.

AGRADECIMIENTO

En primer lugar, agradezco a Dios por esta vida con muchas adversidades pero también con oportunidades que permitieron el crecimiento como persona y profesional.

En segundo lugar, agradezco a mis padres por su apoyo en mi formación académica.

Finalmente, agradezco al Dr. José Alfredo Herrera Quispe por su constante apoyo en el desarrollo de la presente tesis.

Índice general

DEDICATORIA I
AGRADECIMIENTO II
ÍNDICE GENERAL III
ÍNDICE DE TABLAS VI
ÍNDICE DE FIGURAS VI
RESUMEN IX
ABSTRACT X

I Introducción **1**
1.1 Situación Problemática 1
1.2 Formulación del Problema 3
1.2.1 *Problema General* 3
1.2.2 *Problemas Específicos* 3
1.3 Justificación Teórica de la Investigación 3
1.4 Justificación Práctica de la Investigación 4
1.5 Objetivos de la Investigación 5
1.5.1 *Objetivo General* 5
1.5.2 *Objetivo Específicos* 5

II Marco Teórico **6**
2.1 Antecedentes de Investigación 6
2.2 Bases Teóricas 10
2.2.1 *Pronóstico del Precio de una Acción* 10
2.2.2 *Técnicas de Series de Tiempo* 12
2.2.3 *Técnicas de Aprendizaje Automático* 19
2.2.4 Evaluación de la Capacidad Predictivas de las Técnicas 32
2.3 Marco Conceptual 34
2.3.1 Glosario de Términos Relacionados a las Técnicas 34

III Metodología **36**
3.1 Tipos de Investigación 36
3.1.1 *Nivel de Investigación* 36
3.1.2 *Diseño de Investigación* 36

3.1.3 *Método de Investigación* 36
3.2 Población y Muestra 37
3.2.1 *Población* 37
3.2.2 *Muestra* 37
3.2.3 *Muestreo* 37
3.3 Operacionalización de Variables 37
3.3.1 *Identificación de Variables* 37
3.3.2 *Tabla de Operacionalización de Variables* 39
3.3.3 *Descripción de Indicadores de las Variables* 40
3.4 Técnicas e Instrumentos de Recolección de Datos 42
3.4.1 *Técnicas* 42
3.4.2 *Instrumentos* 42
3.4.3 *Validación y Confiabilidad del Instrumento* 42
3.5 Análisis e Interpretación de la Información 43
3.5.1 *Entendimiento del Negocio y Definición de las Métricas de Rendimiento* 43
3.5.2 *Obtención de los Datos* 46
3.5.3 *Análisis Exploratorio de los Datos* 46
3.5.4 *Pre-procesamiento de los Datos e Ingeniería de Variables* . . 47
3.5.5 *Determinación de la Variable Dependiente* 49
3.5.6 *Construcción y Selección del Modelo* 50
3.5.7 *Siguientes fases* 51

IV Resultados y Discusión **52**
4.1 *Análisis Descriptivo de los Precios y Rendimientos* 52
4.2 *Determinación de la Ventana de Pronóstico* 52
4.3 Análisis de Resultados del Experimento 1 56
4.3.1 *Análisis Descriptivo de la Variable Dependiente Precio* 56
4.3.2 *Comparación de las Técnicas de Pronóstico de los Precios* . . 56
4.4 Análisis de Resultados del Experimento 2 57
4.4.1 *Análisis Descriptivo de la Variable Dependiente Rendimiento* . 58
4.4.2 *Comparación de las Técnicas de Pronóstico de los Rendimientos* 58
4.5 Análisis de Resultados del Experimento 3 59
4.5.1 *Análisis Descriptivo de la Variable Dependiente* 59
4.5.2 *Comparación de las Técnicas de Clasificación* 60
4.6 Pruebas de Hipótesis 61
4.7 Discusión de Resultados 62

V Conclusiones y Recomendaciones **64**
5.1 Conclusiones 64

5.2 Recomendaciones 65

Referencias Bibliográficas **73**

Anexos **74**

Anexo A. Hipótesis de Investigación 75
Anexo B. Matriz de Consistencia 76
Anexo C. Introducción a las Redes Neuronales Artificiales 77
Anexo D. Resultados de las Técnicas de Series de Tiempo 86
Anexo E. Arquitectura Tecnológica de Implementación 89

Índice de tablas

Tabla 1 Muestra de Cotización de Precios de una Acción 12
Tabla 2 Prueba de Hipótesis Dickey-Fuller 14
Tabla 3 Matriz de Datos para Aplicación de Modelos de Aprendizaje Automático 20
Tabla 4 Prueba de Hipótesis de Significancia de la Regresión 22
Tabla 5 Muestra de Datos Reales y Pronósticos 32

Tabla 6 Determinación del Precio en el Futuro Día T 41
Tabla 7 Determinación del Rendimiento en el Futuro Día T 41
Tabla 8 Estructura de Información de Cotizaciones 46
Tabla 9 Método de Cálculo de las Variables Dependientes 47

Tabla 10 Precisión de las Técnicas de Pronóstico del Precio 57
Tabla 11 Estabilidad del RMSE en las Técnicas de Pronóstico de los Precios 57
Tabla 12 Precisión de las Técnicas de Pronóstico del Rendimiento 58
Tabla 13 Estabilidad del RMSE en las Técnicas de Pronóstico de los Rendimientos 59
Tabla 14 Ajuste de las Técnicas de Clasificación de los Rendimientos . . 60
Tabla 15 Estabilidad del ACCURACY en las Técnicas de Clasificación . 61
Tabla 16 Verificación de la Hipótesis 61

Tabla 17 Unidades de activación 81
Tabla 18 Variantes para mejorar Overfitting 83
Tabla 19 Alternativas de Regularización 84

Índice de figuras

Figura 1 Principales Bolsas de Valores del Mundo al Cierre de Agosto 2021, por Capitalización Bursátil de Empresas Cotizadas en Billones de Dólares Estadounidenses. 11
Figura 2 Evolutivo de los Precios de una Acción 12
Figura 3 Componentes de una Serie de Tiempo 13
Figura 4 Estacionariedad de una Serie de Tiempo en Media 14
Figura 5 Diferenciación de la Serie de Tiempo 15
Figura 6 Función de Autocorrelación de la Serie de Tiempo 16
Figura 7 Proceso de Modelamiento de la Serie de Tiempo 19
Figura 8 Partición de un Árbol 23
Figura 9 Proceso Random Forest 23
Figura 10 Ejemplo de Predicción para GBM 24
Figura 11 Comparación del Crecimiento de Árboles entre Ramas y Niveles 26
Figura 12 Estructura del Modelo Perceptrón Multicapas con Cuatro Capas Ocultas 27
Figura 13 Rede Neuronal Recurrente 28
Figura 14 Estructura del Modelo Redes Neuronales Recurrentes 28
Figura 15 Estructura del Modelo LSTM 30
Figura 16 Estructura del Modelo GRU 31
Figura 17 Validación Cruzada de una Serie Temporal 34

Figura 18 Operacionalización de las Variables. 40
Figura 19 Etapas para la Predicción de Series Temporales 44
Figura 20 Estructura General para el Mercado de Acciones 45

Figura 21 Evolutivo de los Precios de la Acción 53
Figura 22 Evolutivo de los Rendimiento de los Precios de la Acción . . . 54
Figura 23 Determinación de la Ventana de Predicción 55
Figura 24 Histograma de los Rendimientos de la Acción 58
Figura 25 Clasificación de los Días con Mayor Rendimiento 60

Figura 26 Matriz de Consistencia 76
Figura 27 Comparación de Neuronales del Cerebro y Neuronas Artificiales 78

Figura 28 Estructura de Redes Neuronales Artificiales 78
Figura 29 Resultados Obtenido de AUTOARIMA en el Pronóstico de Precios 86
Figura 30 Especificación de los Parámetros de ARIMA (2,1,0) para los Precios 87
Figura 31 Resultados Obtenido de AUTOARIMA en el Pronóstico de Rendimientos 87
Figura 32 Especificación de los Parámetros de ARIMA (2,1,0) para los Rendimientos 88
Figura 33 Arquitectura de Software de la Aplicación 89

RESUMEN

El presente trabajo de investigación tiene como objetivo general comparar las técnicas de Series de Tiempo y Aprendizaje Automático, en cuanto a su capacidad de pronóstico de los precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York, la cual se aborda desde tres enfoques: pronóstico de los precios, pronóstico de los rendimientos y la clasificación de los altos rendimientos. Para ello se utilizaron datos históricos de precios de acciones de la empresa entre enero de 2020 y enero de 2023, se generaron variables indicadoras derivadas de los precios históricos, tales como rendimientos históricos, medias móviles, ratios entre precios, índice de fuerza relativa, bandas de Bollinger y oscilador estocástico, entre otros. Se determinó una ventana futura óptima de predicción en 20 días, a través de un análisis histórico iterativo de los rendimientos, definiendo como variables objetivo de pronósticos: el precio al día 20, el rendimiento al día 20 y la clasificación de los rendimientos al día 20. Finalmente, se implementaron técnicas de series de tiempo como ARIMA (*Autoregressive Integrated Moving Average*) y de aprendizaje automático como regresión, árboles de decisión, random forest, XGBoost, lightGBM, redes neuronales de perceptrón multicapa y redes neuronales recurrentes con LSTM (*Long Short-Term Memory*) y GRU (*Gated Recurrent Units*). Los resultados mostraron que las técnicas de redes neuronales recurrentes LSTM y GRU, XGBoost y lightGBM lograron una mayor precisión tanto en el pronóstico de precios y rendimientos como en la clasificación de los altos rendimientos. En conclusión, la superioridad de las técnicas de aprendizaje automático es particularmente evidente en contextos de alta volatilidad y patrones no lineales complejos, lo que sugiere que su utilización puede proporcionar una ventaja competitiva para analistas financieros, inversores y gestores de portafolios.

Palabras clave: Pronóstico, Clasificación, Precios, Rendimientos, Aprendizaje automático, Regresión, Árboles de decisión, Redes neuronales, LSTM, GRU, XGBoost.

ABSTRACT

The present research aims to compare Time Series and Machine Learning techniques in terms of their forecasting ability for the prices and returns of a Peruvian company's stock listed on the New York Stock Exchange. This study is approached from three perspectives: price forecasting, return forecasting, and the classification of high returns. For this purpose, historical stock price data from the company between January 2020 and January 2023 were used. Indicator variables derived from historical prices were generated, such as historical returns, moving averages, price ratios, the relative strength index, Bollinger Bands, and the stochastic oscillator, among others. An optimal future prediction window of 20 days was determined through an iterative historical analysis of returns, defining the forecasting target variables as follows: the price on day 20, the return on day 20, and the classification of returns on day 20. Finally, time series techniques such as ARIMA (Autoregressive Integrated Moving Average) and machine learning techniques such as regression, decision trees, random forest, XGBoost, lightGBM, multilayer perceptron neural networks, and recurrent neural networks with LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Units) were implemented. The results showed that recurrent neural network techniques (LSTM and GRU), XGBoost, and lightGBM achieved higher accuracy in both price and return forecasting as well as in the classification of high returns. In conclusion, the superiority of machine learning techniques is particularly evident in contexts of high volatility and complex nonlinear patterns, suggesting that their use can provide a competitive advantage for financial analysts, investors, and portfolio managers.

Keywords: Forecasting, Classification, Prices, Returns, Machine Learning, Regression, Decision Trees, Neural Networks, LSTM, GRU, XGBoost.

Capítulo I

Introducción

1.1. Situación Problemática

En el sector de las finanzas existen mercados donde se realizan la compra o venta de activos financieros como acciones, bonos, etc. (Staszkiewicz y Staszkiewicz, 2015). El mercado de acciones generalmente es conocido como la bolsa de valores, donde diversas empresas emiten acciones en el mercado a un precio determinado y los inversores están dispuestos a adquirirlas (Amat, 2010). La Bolsa de Valores de Nueva York (NYSE) es la que tiene el mayor volumen de negociación a nivel mundial (ver Figura 1) y se caracteriza por ser dinámica y líquida, debido a las constantes transacciones que se realizan en cuestión de segundos, los cuales generan variaciones de los precios.

Las variaciones en los precios a corto plazo generan una gran parte de la incertidumbre en los mercados bursátiles, por lo que contar con pronósticos precisos de dichas oscilaciones supone una ventaja económica relevante para los inversionistas (Khare et al., 2017). Estas fluctuaciones representan un desafío, ya que los inversores desconocen con exactitud si los precios aumentarán o disminuirán en el futuro, lo cual repercute directamente en sus niveles de rentabilidad. Con el conocimiento de los precios futuros, los inversores podrían establecer sus estrategias de inversión para maximizar su rentabilidad y reducir el riesgo, por ejemplo, si un inversor sabe que mañana el precio de una acción aumentará, tomará la decisión de comprarla; en caso contrario, buscará venderla o esperar a que el precio se recupere y presente una tendencia alcista.

El pronóstico de precios de acciones ha sido un área de investigación clave en el ámbito de las finanzas, debido a la relevancia de estos pronósticos en la toma de decisiones financieras, tanto para inversores como para analistas de mercados. Tradicionalmente, las series de tiempo han sido utilizadas para modelar y predecir comportamientos financieros debido a su capacidad para capturar patrones temporales y estacionales. Técnicas como el modelo ARIMA han sido ampliamente utilizadas para modelar precios de activos debido a su simplicidad y efectividad en series de tiempo

estacionarias (Box et al., 2015). Sin embargo, los mercados financieros modernos, caracterizados por su alta volatilidad y la presencia de patrones no lineales, han desafiado la capacidad de estos modelos tradicionales para generar pronósticos precisos a largo plazo (Tsay, 2010).

Por otro lado, existen diferentes tipos de variables objetivo a los cuales se pueden aplicar los modelos de pronósticos de series de tiempo y de aprendizaje automático, como son: (i) el pronóstico de precios, (ii) el pronóstico de rendimientos y (iii) la clasificación de altos rendimientos. El pronóstico de precios se centra en estimar el valor futuro de un activo financiero, ayudando a los inversores a determinar los mejores momentos para comprar o vender. El pronóstico de rendimientos que permite determinar las ganancias o pérdidas esperadas sobre una inversión en un periodo determinado, proporcionando una visión más clara del potencial retorno de una acción. Además, se tiene a la clasificación de los altos rendimientos, entendiéndose como aquellos rendimiento que superan un umbral de rentabilidad mínimo, lo cual es útil para anticipar momentos de altos rendimientos en los mercados, lo cual permite ajustar las estrategias de inversión en función a la expectativas.

Para cada uno de estos tres enfoques existen ténicas que se pueden utilizar y ajustarse mejor a los tipos de datos, pero dentro del contexto de los modelos de aprendizaje supervisado (Shah et al., 2019). En el caso de los modelos de pronósticos de precios y de los rendimientos, al tener una variable respuesta de naturaleza continua, se pueden utilizar las técnicas de pronóstico como ARIMA en series de tiempo o las regresiones y las redes neuronales recurrentes como LSTM en aprendizaje automático. Respecto a la clasificación de altos rendimientos, es posible considerarlo como una variable dependiente categórica por lo que pueden utilizarse técnicas de clasificación.

Finalmente, este trabajo de investigación se enfoca en la analítica financiera mediante la aplicación de técnicas de regresión, pronósticos y de clasificación, en el enfoque de la estadística, aprendizaje automático y aprendizaje profundo. De acuerdo con García et al. (2013) la predicción de series de tiempo financieras tiene su origen en la década de 1980 donde se utilizaron las técnicas ARIMA para modelar la media de un activo, en la década de 2000 se tuvo mayor publicación de modelos híbridos Máquina Vector de Soporte, redes neuronales, algoritmos genéticos, entre otros. De acuerdo con Hu et al. (2021), desde el 2015 en adelante ha habido un aporte importante a nivel académico del uso de aprendizaje profundo como las redes neuronales recurrentes.

1.2. Formulación del Problema

1.2.1. *Problema General*

¿Qué técnicas, entre Series de Tiempo y Aprendizaje Automático, proporcionan la mayor capacidad de pronóstico de los precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York?

1.2.2. *Problemas Específicos*

- ¿Qué técnica, entre Series de Tiempo y Aprendizaje Automático, proporciona la mayor precisión en el pronóstico de los precios de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York?
- ¿Qué técnica, entre Series de Tiempo y Aprendizaje Automático, proporciona la mayor precisión en el pronóstico de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York?
- ¿Qué técnica de Aprendizaje Automático proporciona la mayor exactitud en la clasificación de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York?

1.3. Justificación Teórica de la Investigación

En los últimos años, los avances en el campo de aprendizaje automático han proporcionado nuevas herramientas para abordar el pronóstico de precios de acciones, superando las limitaciones de los enfoques tradicionales. Modelos como las redes neuronales recurrentes, y en particular las arquitecturas LSTM y GRU, han demostrado ser altamente efectivos en la modelización de secuencias temporales debido a su capacidad para aprender dependencias a largo plazo en datos secuenciales (Hochreiter, 1997). Estas redes tienen la ventaja de capturar patrones no lineales complejos, lo cual es esencial en mercados financieros donde los precios están influenciados por una amplia gama de factores, tanto económicos como psicológicos, que no siguen una relación lineal.

Además, técnicas de aprendizaje automático como XGBoost y lightGBM han demostrado ser útiles en tareas de clasificación o regresión, mostrando gran capacidad para manejar grandes volúmenes de datos y para adaptarse a diversas dinámicas del mercado (T. Chen y Guestrin, 2016). Estos métodos no sólo permiten mejorar la precisión en el pronóstico de precios y rendimientos, sino que también facilitan la identificación de patrones complejos en los datos, lo que los hace especialmente adecuados para el pronóstico en contextos de alta volatilidad y mercados impredecibles.

Este trabajo de investigación se justifica teóricamente en la medida en que busca comparar los enfoques tradicionales de series de tiempo con las modernas técnicas de aprendizaje automático, evaluando cuál de estos enfoques es más efectivo para el pronóstico de precios, rendimientos y la identificación de días con altos rendimientos en mercados altamente dinámicos. Con el desarrollo de este estudio, se pretende contribuir a la literatura existente al proporcionar un análisis detallado de las ventajas y limitaciones de ambos enfoques en el contexto específico de una empresa peruana que cotiza en la NYSE.

1.4. Justificación Práctica de la Investigación

El interés por la mejora de los modelos predictivos también se justifica por la necesidad de optimizar las estrategias de inversión y de toma de decisiones en un contexto de globalización financiera. La capacidad para prever con mayor exactitud los movimientos del mercado es un factor clave para los inversores que buscan maximizar sus rendimientos y mitigar los riesgos. En este contexto, el uso de aprendizaje automático y de técnicas de series de tiempo es relevante al permitir la mejora de los modelos de predicción utilizados en la gestión de carteras y en la toma de decisiones de inversión.

El estudio se lleva a cabo debido a la necesidad de los inversionistas individuales de anticipar el comportamiento de los precios de las acciones en el mercado bursátil, con el fin de disminuir la incertidumbre que enfrentan. Esto les permite optimizar sus niveles de rentabilidad esperada, especialmente en un contexto donde las plataformas digitales, como páginas *web* y aplicaciones móviles (*etoro*, *xtb trading*, entre otras), han facilitado el acceso a nuevos mercados financieros.

La Bolsa de Valores de Nueva York es la que tiene mayor volumen de capitalización a nivel mundial (STATISTA, 2021), con lo cual sustenta la elección como fuente de información para la aplicación de las técnicas diversas de pronóstico. Además, estudios diversos desde el enfoque financiero son aplicados a esta bolsa de valores. Por otro lado, una de las empresas peruanas listada en la NYSE es Credicorp Ltd. (desde 1995), holding financiero más grande de Perú y tiene una fuerte presencia en la banca, seguros y gestión de fondos (educatyba, 2024). Su importancia dentro del país es significativo y su quiebra implicaría resultados negativos para la situación económica del país. En ese sentido, pronosticar el precio futuro de la acción de esta empresa representa una actividad importante no sólo para beneficiar a los inversionistas peruanos y extranjeros, sino que también representa un signo de bienestar económico del país.

1.5. Objetivos de la Investigación

1.5.1. *Objetivo General*

Comparar las técnicas de Series de Tiempo y Aprendizaje Automático en cuanto a su capacidad de pronóstico de los precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.

1.5.2. *Objetivo Específicos*

- Comparar las técnicas de Series de Tiempo y Aprendizaje Automático en cuanto a su precisión en el pronóstico de los precios de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.
- Comparar las técnicas de Series de Tiempo y Aprendizaje Automático en cuanto a su precisión en el pronóstico de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.
- Comparar las técnicas de Aprendizaje Automático en cuanto a la exactitud en la clasificación de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.

Capítulo II

Marco Teórico

2.1. Antecedentes de Investigación

Arjomandi-Nezhad et al. (2022), en su publicación "*Pandemic-Aware Day-Ahead Demand Forecasting Using Ensemble Learning*", planteó la aplicación de los modelos ensamblados para el pronóstico del consumo de energía durante la situación de la pandemia utilizando técnicas como *Bagging*, *Random Forest*, *Ada Boost*, *Gradient Boosting*, XGBoost, CatBoost y LGBoost. Para ello se utilizó una fuente de internet que corresponde al caso de la predicción de la carga a nivel nacional de Alemania (Series, 2019), se buscó pronosticar la demanda del día siguiente y se evaluó el performance del modelo a través de los indicadores $RMSE$, $MAPE$, MAE y R^2. Se concluyó que los modelos *boosting* y *bagging-boosting* son los más precisos y robustos, además, que el modelo es extensible para futuras pandemias.

Da Silva et al. (2022), en su publicación "*Performance evaluation of LSTM neural networks for consumption prediction*", planteó la aplicación de técnicas alternativas de *deep learning* y *machine learning* para el pronóstico del consumo de energía como son la red neuronal recurrente LSTM y las técnicas *Random Forest* y XGBOOST. Para ello se utilizaron tres fuentes diferentes de información donde se evaluó el rendimiento de las técnicas, para ello aplicaron cross validation para series temporales que les permitió determinar 10 mediciones de precisión del modelo de pronóstico para concluir estadísticamente si un modelo es mejor que otro. Finalmente, concluyeron estadísticamente mediante la prueba de hipótesis no paramétrica Kruskal Wallis que la red neuronal recurrente LSTM presentó mejor precisión que las otras dos técnicas, además, las LSTM alcanzaron una generalización de capacidad sobresaliente con datos que no fueron utilizados en la fase de entrenamiento.

Ensafi et al. (2022), en su publicación "*Time-series forecasting of seasonal items sales using machine learning – A comparative analysis*", planteó la aplicación de técnicas avanzadas de redes neuronales para el pronóstico de las ventas en una tienda

retail a través de técnicas como Prophet, CNN y LSTM cuyos niveles de *performance* son comparados con técnicas estadísticas como SARIMA ya que la serie de tiempo presenta la componente estacional. Para ello se utilizó una fuente de internet *Superstore* como referencia para la aplicación de las técnicas ARMA, ARIMA, SARIMA, Prophet, Vanilla LSTM, stacked LSTM, Bidireccional LSTM, LSTM y CNN, para evaluar el performance del pronóstico se utilizaron indicadores MSE, RMSE y MAPE. Finalmente, se concluyó que los métodos de redes neuronales alcanzaron mejores resultados que los métodos tradicionales, siendo stacked LSTM el mejor método, seguido por el método CNN.

Al Mamun et al. (2020), en su publicación "*A Comprehensive Review of the Load Forecasting Techniques Using Single and Hybrid Predictive Models*", tiene por objetivo el estudio exhaustivo de diversos métodos relacionados con el pronóstico de series de tiempo y se realiza un estado del arte para construir modelos híbridos. Se asume que la aplicación de los modelos híbridos en el pronóstico del consumo permite encontrar un menor porcentaje de error y con ello ahorrar millones de dólares a las empresas de servicios públicos. Los algoritmos considerados se enfocan en Deep Learning, métodos basados en reglas, los métodos metaheurísticos y SVM, de las cuales dos o más se fusionan para diseñar un modelo híbrido. Además, se desarrolla una comparación entre diversos indicadores de evaluación de rendimiento en términos de error cuadrático medio (RMSE), error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE). En la publicación se determina que SVM, ANN son los modelos relevantes y que la publicación representa una base teórica para los lectores.

En su trabajo titulado "Stock Price Prediction Using News Sentiment Analysis", Mohan et al. (2019) destacó la existencia de una fuerte relación entre los artículos de noticias y la evolución de los precios bursátiles, lo que permite mejorar la capacidad de predicción. Este enfoque, conocido como análisis de sentimientos, se combina con diferentes algoritmos como Máquinas de Vectores de Soporte (SVM), Regresión Bayesiana en Redes, técnicas de Deep Learning y modelos de series temporales como ARIMA, RNN y Facebook Prophet. Para el entrenamiento de dichos modelos se emplearon precios diarios de las acciones incluidas en el índice SP500 a lo largo de cinco años, junto con más de $265,000$ artículos de noticias financieras. Los hallazgos mostraron que las RNN ofrecieron el mejor desempeño y confirmaron la correlación entre el contenido textual y la tendencia de los precios de las acciones; sin embargo, se observó que los modelos tuvieron bajo rendimiento en escenarios de precios reducidos o con alta volatilidad.

Z. Chen et al. (2020) realizaron la investigación titulada: "*Bitcoin price prediction using machine learning: An approach to sample dimension engineering*". El objetivo de la investigación es la aplicación de diferentes técnicas de modelado a muestras con diferentes estructuras de datos y características dimensionales. En el aspecto me-

todológico se utilizaron los métodos estadísticos que incluyen la regresión logística y el análisis discriminante lineal para la predicción diaria de precios de Bitcoin con características de alta dimensión logran una precisión del $66\,\%$, superando a los algoritmos de aprendizaje automático más complicados, que tiene $65.3\,\%$. Por otro lado, los modelos de aprendizaje automático que incluyen *Random Forest*, XGBoost, Análisis discriminante cuadrático, Máquina de vectores de soporte y Memoria a corto plazo a largo plazo para la predicción de precios de intervalos de cinco minutos de Bitcoin son superiores a los métodos estadísticos, con una precisión que alcanza el $67.2\,\%$.

Z. Peng et al. (2019) en su publicación "*Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm*", estudió el precio de la vivienda de casas de segunda mano, donde analizó y estudió $35,417$ datos capturados por la red Chengdu HOME LINK. Utilizó modelos de regresión lineal múltiple, árbol de decisión y XGboost para ajustar la curva de puntaje del precio de la vivienda y seleccionó el modelo de predicción óptimo. Los resultados experimentales mostraron que la precisión de la predicción XGboost es la más alta, y el puntaje de precisión de la predicción alcanzó un valor de 0.9251, seguido por los árboles de decisión y finalmente la regresión lineal.

Moukalled (2019), en su publicación "*Automated Stock Price Prediction Using Machine Learning*", propuso un sistema de trading automatizado que integra funciones matemáticas, aprendizaje automático y otros drivers externos, como los sentimientos de las noticias, con el fin de lograr una mejor precisión de predicción de acciones y emitir operaciones rentables. Para ello, determina el precio o la tendencia de un determinado stock para el próximo final del día, teniendo en cuenta las primeras horas de cotización del día. Para lograrlo, entrena algoritmos tradicionales de aprendizaje automático, múltiples modelos de aprendizaje profundo teniendo en cuenta la importancia de las noticias relevantes. Se realizaron varios experimentos pero la precisión más alta ($82.91\,\%$) se logró usando SVM para acciones de *Apple Inc*.

C. Peng et al. (2019) en su publicación "*Stock Price Prediction based on Recurrent Neural Network with Long Short-Term Memory Units*", propone dos métodos de predicción integrados con un marco de aprendizaje profundo que utilizan unidades de memoria a corto y largo plazo (modelo RNN y LSTM con dos esquemas de predicción de optimización), proporcionando predicciones para horizontes tanto a corto como a largo plazo. Se han llevado a cabo pruebas de comparación entre los dos métodos propuestos, y los resultados del experimento han demostrado que los nuevos métodos tienen un rendimiento significativo al capturar tanto la tendencia como el valor exacto de los datos del stock. Ambos esquemas funcionan razonablemente bien en la predicción de precios de acciones a largo plazo. Además, en general, el método híbrido supera al método recursivo tanto en predicción general como en intervalos seleccionados.

En su artículo "Study of Stock Return Predictions Using Recurrent Neural Net-

works with LSTM", Naik y Mohan (2019) analizaron el uso de Redes Neuronales Recurrentes (RNN) con unidades de Memoria a Largo Plazo (LSTM) para estimar los rendimientos futuros de las acciones. Para el experimento, se emplearon datos provenientes de la Bolsa Nacional de Valores de India (NSE), a partir de los cuales se calcularon los rendimientos en función de los precios de cierre. Estos valores se utilizaron como entrada del modelo RNN, cuyo objetivo principal fue reducir al mínimo el error de predicción. En conclusión, el enfoque basado en RNN con LSTM logró un mejor desempeño frente a una Red Neuronal Artificial de tipo feedforward.

Liu et al. (2019), en su publicación "*Stock Prices Prediction using Deep Learning Models*", se enfocó en la predicción de los precios de las acciones por medio de los modelos de aprendizaje profundo. Este trabajo utilizó autoencoders dispersos con redes convolucionales residuales unidimensionales (1-D), que es un modelo de aprendizaje profundo, para alterar los datos. La memoria a corto y largo plazo (LSTM) se usaron para predecir el precio de las acciones. Los resultados del experimento muestran que las redes convolucionales residuales (1-D) pueden eliminar el ruido de los datos y extraer características profundas mejor que un modelo que combina transformaciones wavelet (WT) y autoencoders apilados (SAE). Además, al comparar el performance de los modelos pronosticando el precio absoluto de las acciones y la tasa de cambio del precio, se concluye que es mejor predecir la tasa de cambio del precio.

Sachdeva et al. (2019), en su publicación "*An Effective Time Series Analysis for Equity Market Prediction Using Deep Learning Model*", desarrollaron un modelo de deep learning utilizando la Red Neuronal Recurrente (RNN, por sus sigla en inglés) para pronosticar los precios de las acciones de INFOSYS LTD y el índice NIFTY50 de la Bolsa de Valores de India. A partir de los resultados obtenidos, concluyeron que los modelos de aprendizaje profundo son capaces de pronosticar el mercado de valores de la India con un ajuste de $97.64\,\%$ usando un optmizador RmsProp.

Qian y Gao (2017), en su publicación "*Financial series prediction: Comparison between precision of time series models and machine learning methods*", se compara la precisión de los modelos de series temporales tradicionales como ARIMA y GARCH con técnicas de aprendizaje automático, que han demostrado ser mejores en ámbitos como la regresión y la clasificación. Entre modelos de aprendizaje automático se consideraron las siguientes técnicas: Regresión Logística (RL), Multilayer Perceptrón (MLP), Máquina Vectorial de Soporte (SVM) y SVM con Denoising Autoencoders (DAE-SVM). En su aplicación, se han utilizado los índices SP 500, Down30 y Nasdaq entre las fechas 02-01-2012 y 26-12-2016. En cada índice se midió la precisión de la predicción mediante la medida hitratio y se encontró que el mejor modelo es SVM y su variante DAE-SVM seguidos por la RL, mientras que ARIMA sólo está por arriba de MLP (sólo en el índice Down30 resulta arriba que RL).

Khare et al. (2017), en su publicación "*Short term stock price prediction using deep learning*", se enfoca en la creación de un modelo algorítmico predictivo utilizando el aprendizaje automático, específicamente las redes neuronales profundas. Para la predicción de precios a corto plazo de las acciones consideró 10 acciones registradas en la Bolsa de Valores de Nueva York. La revisión se centró en la predicción de estos precios a corto plazo aprovechando el poder del análisis técnico. El análisis técnico guía el marco para comprender los patrones de los precios históricos que se introdujeron en él, e intenta pronosticar probabilísticamente los precios futuros fugaces de la acción bajo revisión. El artículo analiza dos tipos distintos de redes neuronales artificiales, las redes neuronales de avance y las redes neuronales recurrentes. La revisión revela que el perceptrón multicapa Feed Forwards tiene un rendimiento superior al de la memoria a corto plazo a la hora de predecir los precios a corto plazo de una acción.

2.2. Bases Teóricas

2.2.1. *Pronóstico del Precio de una Acción*

Las acciones son instrumentos financieros que representan una parte del capital social de una empresa. Al comprar acciones, un inversor adquiere una participación en la propiedad de la empresa y, por lo tanto, tiene derecho a una parte de sus beneficios y activos (Jimenez, 2015).

La bolsa de valores es un mercado organizado donde se compran y venden acciones, así como otros instrumentos financieros como bonos y derivados. La Bolsa de Valores de Nueva York (NYSE) es uno de los mercados más grandes del mundo para valores y otras inversiones (Britannica, 2016). Según (STATISTA, 2021), a fines de mayo de 2019, NYSE se clasificó como el mayor operador bursátil con una capitalización de mercado de 23.21 billones de dólares estadounidenses, tres veces más que el segundo operador más grande, NASDAQ (Ver Figura 1).

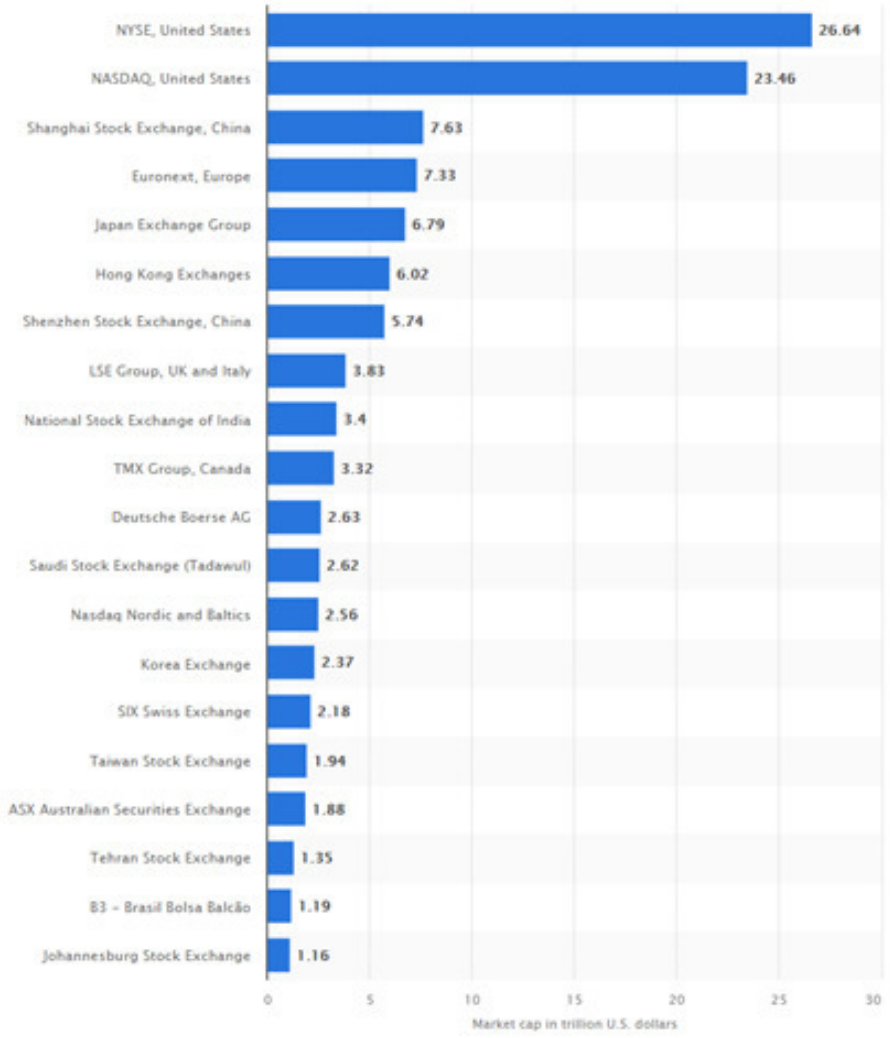
Por otro lado, el pronóstico del precio de una acción en la bolsa de valores es la expectativa del valor futuro de los precios diarios que se caracteriza por ser incierto, dinámico y no lineal, por esa razón los inversores han buscado formas de adquirir conocimientos sobre las empresas que cotizan en el mercado para mejorar el rendimiento de sus inversiones.

Según Peixeiro (2022), pronosticar es predecir el futuro utilizando datos históricos y el conocimiento de eventos futuros. Se considera un arte porque se necesita de un razonamiento y capacidad de seleccionar la información relevante para analizar un evento en particular, asimismo, se considera como ciencia porque se cuantifica y modela el comportamiento del mercado (Tjung, 2010).

Para Fama (1970), de acuerdo con su hipótesis de mercados eficientes que sugi-

Figura 1

Principales Bolsas de Valores del Mundo al Cierre de Agosto 2021, por Capitalización Bursátil de Empresas Cotizadas en Billones de Dólares Estadounidenses.



Nota. Tomado de (STATISTA, 2021).

rió, es imposible predecir los precios porque los precios del mercado ya incorporan y reflejan toda la información disponible. Sin embargo, es una hipótesis que no necesariamente se cumple porque cada inversor reacciona al mercado de diversas formas en diferentes instantes del tiempo de acuerdo a la información que considera relevante y a las conclusiones que pueda tener de su análisis, expertise y el perfil de inversionista.

En la literatura de la analítica financiera, se han formulado diversas técnicas de pronóstico a lo largo del tiempo, iniciando por las técnicas tradicionales como las series de tiempo y últimamente por la aplicación de técnicas de aprendizaje automático. A continuación, se propone las bases teóricas de las técnicas seleccionadas en el presente trabajo de investigación.

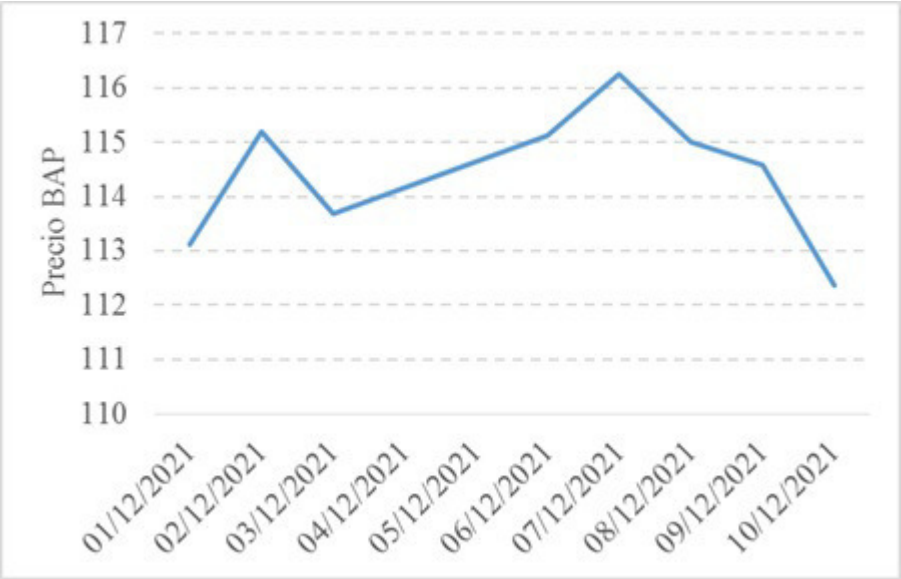
2.2.2. *Técnicas de Series de Tiempo*

Definición de las Series de Tiempo. Las series de tiempo son datos que miden el cambio de una variable en estudio en función del tiempo (estructura primaria que puede utilizarse para ordenar los datos) (Lazzeri, 2020). En la Tabla 1 y en la Figura 2 se presenta una muestra de serie de tiempo que hace referencia al evolutivo de precios de una acción.

Tabla 1
Muestra de Cotización de Precios de una Acción

Fecha	Precio
01-12-2021	113.12
02-12-2021	115.19
03-12-2021	113.69
06-12-2021	115.11
07-12-2021	116.26
08-12-2021	114.99
09-12-2021	114.57
10-12-2021	112.37

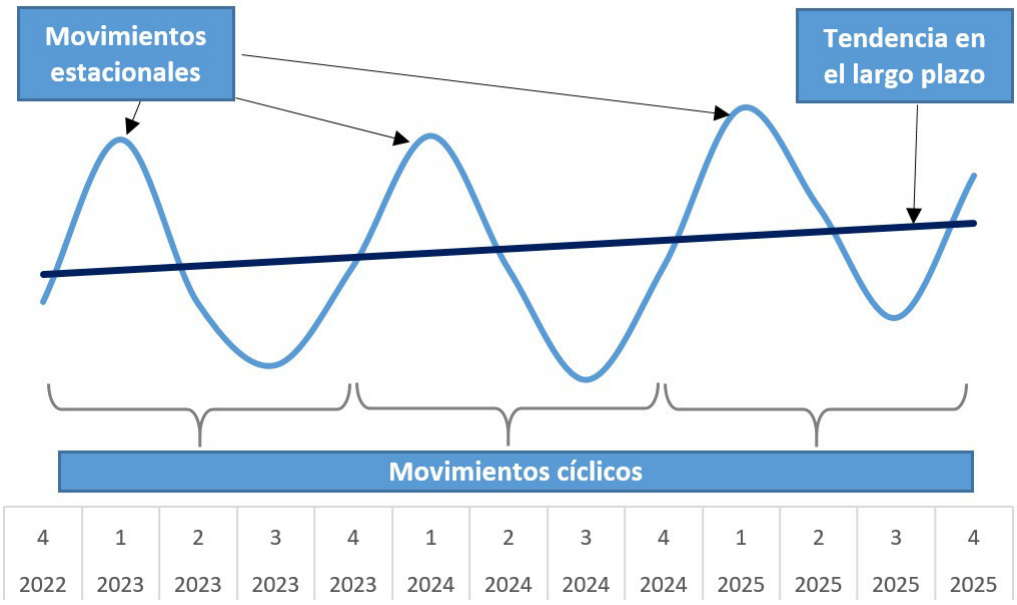
Figura 2
Evolutivo de los Precios de una Acción



En el análisis de las series de tiempo se tienen cuatro componentes que se deben evaluar (ver Figura 3):

- **Tendencia en el largo plazo:** se refiere al movimiento general de los valores de series de tiempo para aumentar o disminuir durante un intervalo de tiempo prolongado.
- **Movimientos estacionales a corto plazo:** son fluctuaciones temporales periódicas que muestran la misma variación y generalmente se repiten durante un período fijo y conocido, menor a un año.
- **Movimientos cíclicos a corto plazo:** Las variaciones cíclicas son patrones recurrentes que existen cuando los datos exhiben subidas y bajadas que no son de un período fijo. Un período completo es un ciclo, pero un ciclo no tendrá un período de tiempo predeterminado específico, incluso si la duración de estas fluctuaciones temporales suele ser superior a un año.
- **Fluctuaciones aleatorias:** Representa el comportamiento de la serie que no puede ser explicado por las componentes anteriores, conocido como ruido blanco.

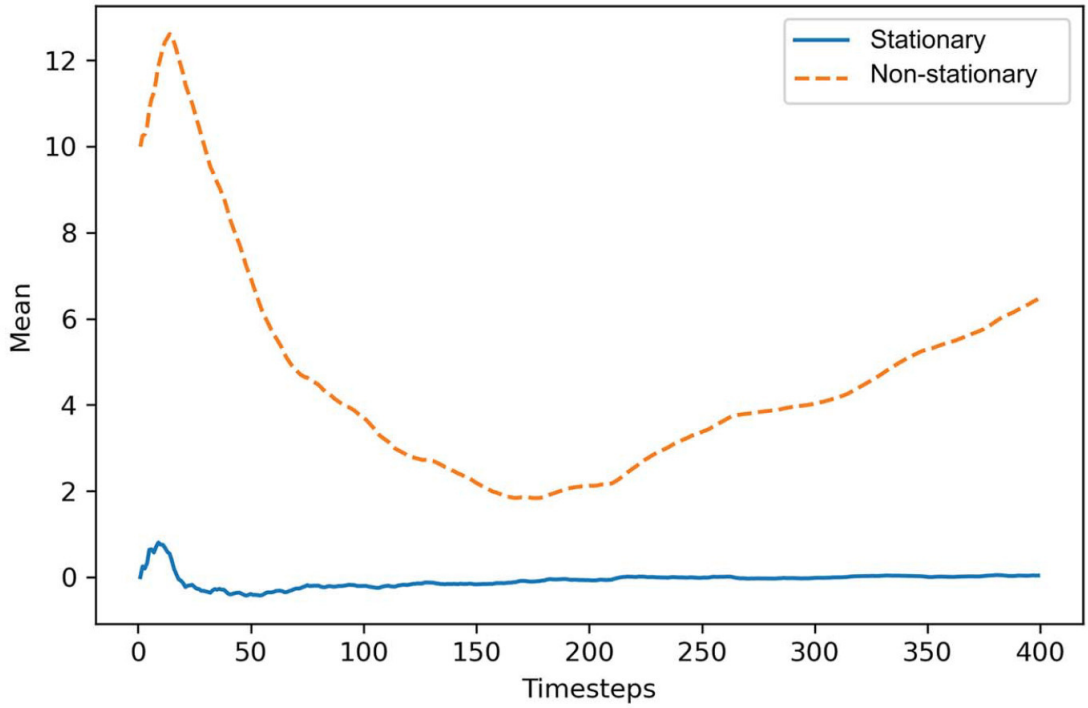
Figura 3
Componentes de una Serie de Tiempo



Nota. Adaptado de "Machine learning for time series forecasting with Python", Lazzeri (2020). Pag.7.

Estacionariedad de las Series de Tiempo. Una serie de tiempo se considera estacionaria cuando sus propiedades estadísticas de la serie no cambian en el tiempo, es decir, la media, varianza y autocorrelación de la serie permanece constantes y son independientes del tiempo (Peixeiro, 2022). En la Figura 4 se muestra unos ejemplos de series estacionarias y no estacionarias.

Figura 4
Estacionariedad de una Serie de Tiempo en Media



Nota. Tomado de "Time Series Forecasting in Python", Peixeiro (2022). Pag.40.

La importancia del concepto de estacionariedad radica en que muchos de los modelos de series de tiempo asumen estacionariedad y para poder verificar su cumplimiento se ha definido la prueba de hipótesis Dickey-Fuller, donde se considera como hipótesis nula que la serie tiene raiz unitaria lo que significa que la serie es no estacionaria; por tanto, se debe verificar previamente que se rechace la prueba de hipótesis Dickey-Fuller considerando como nivel de significancia al 5 % (ver Tabla 2).

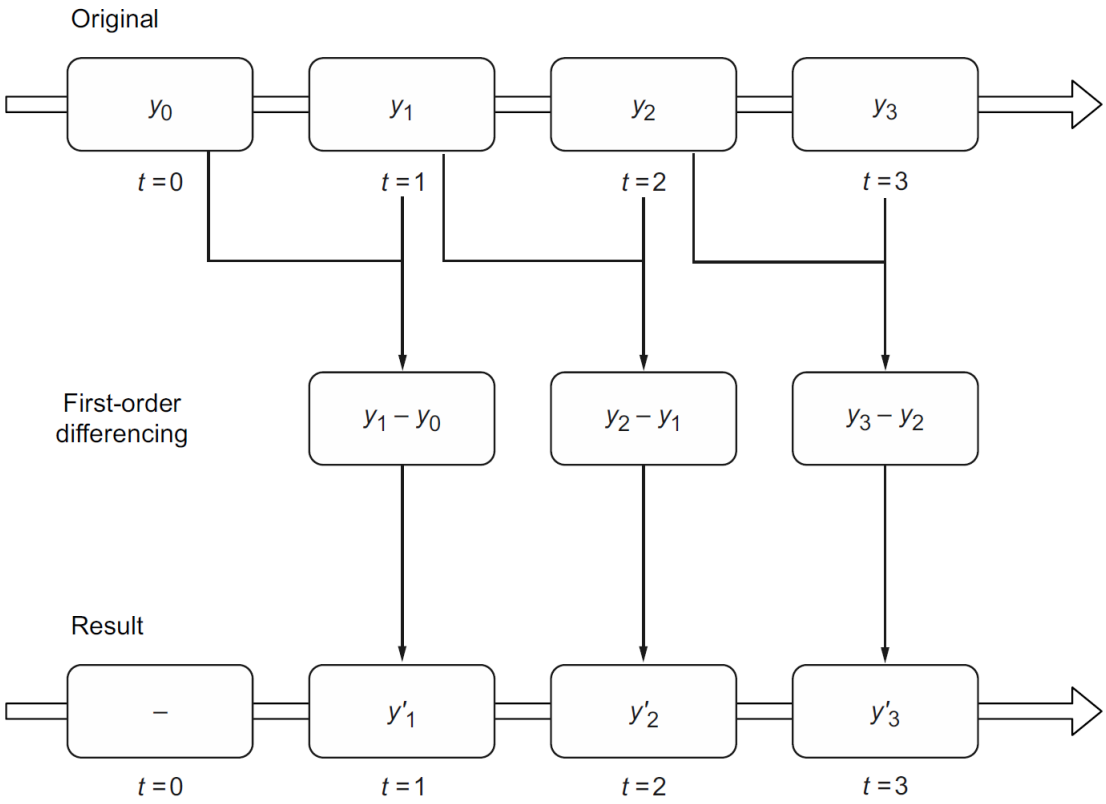
Tabla 2
Prueba de Hipótesis Dickey-Fuller

H0: $\phi = 1$. La serie tiene raiz unitaria (no estacionaria)
H1: $\phi < 1$. La serie no tiene raiz unitaria (estacionaria)
Considerando que
$\triangle X_t = (\phi - 1)X_{t-1} + \epsilon_t$
Al aplicar una regresión se obtiene el estadístico de prueba T sobre la variable dependiente rezagada y también se puede estimar el valor del coeficiente de regresión de la variable X_{t-1} y por tanto el valor de ϕ.

En el caso de que la serie no sea estacionaria existe una alternativa de transformación que permite alcanzar el cumplimiento del supuesto, a través de la diferenciación de la serie (ver Figura 5).

Figura 5

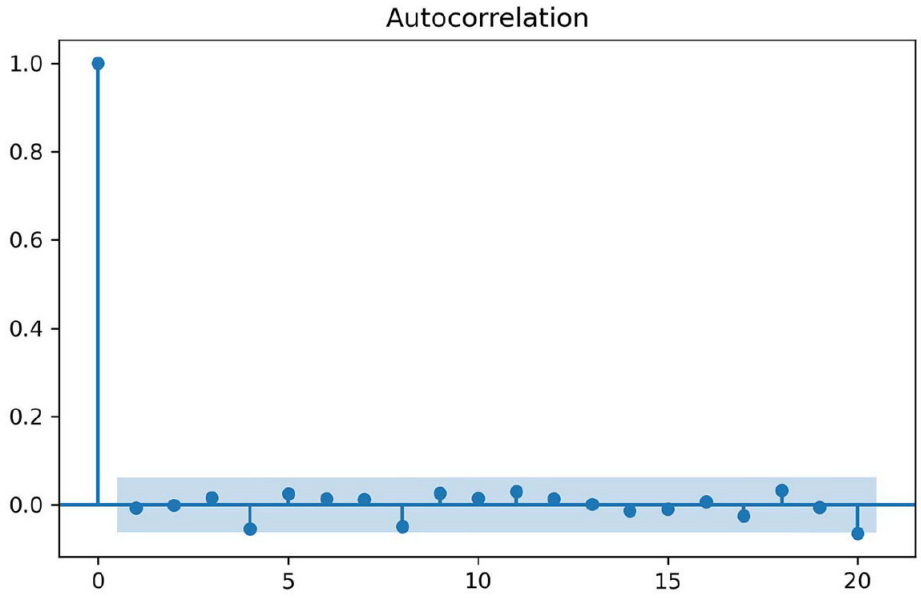
Diferenciación de la Serie de Tiempo



Nota. Tomado de "Time Series Forecasting in Python", Peixeiro (2022). Pag.37.

Función de Autocorrelación. Luego de verificar el cumplimiento del supuesto de estacionariedad de la serie, es importante evaluar la función de autocorrelación (ACF) para comprender el tipo de proceso de la serie, con ello se mide la relación lineal entre valores rezagados de una serie de tiempo, por ejemplo, se puede calcular el coeficiente de correlación lineal entre Y_t y $Y_{t\text{-}1}$ considerando como rezago igual a 1, también se puede determinar la correlación lineal entre Y_t y $Y_{t\text{-}2}$ teniendo un rezago 2.

Figura 6
Función de Autocorrelación de la Serie de Tiempo



Nota. Tomado de "Time Series Forecasting in Python", Peixeiro (2022). Pag.45.

En la Figura 6 el eje X es el número de rezagos y el eje Y es el coeficiente de correlación lineal. En presencia de una tendencia, un gráfico de ACF mostrará que los coeficientes son altos para retrasos cortos y disminuirán linealmente a medida que aumenta el retraso; si los datos son estacionales, la gráfica ACF también mostrará patrones cíclicos.

Función de Autocorrelación Parcial. La autocorrelación parcial mide la correlación entre los valores rezagados en una serie de tiempo cuando se elimina la influencia de los valores rezagados correlacionados en el medio. Se puede graficar la función de autocorrelación parcial para determinar el orden de un proceso $AR(p)$ estacionario. Los coeficientes serán no significativos después del desfase p (Peixeiro, 2022).

Técnicas de Series de Tiempo. Los modelos estadísticos aplicados al modelamiento de series temporales que asumen estacionariedad son los modelos de media móvil ($MA(q)$), modelos autoregresivos ($AR(p)$), modelos $ARMA(p, q)$. Los modelos $ARIMA(p, d, q)$ son aplicados a series no estacionarias, $SARIMA(p, d, q)(P, D, Q)_m$ para series de tiempo estacionales y modelos $SARIMAX$ para incluir variables externas en el pronóstico.

Modelo Media Móvil (MA, por sus siglas en inglés). El modelo de media móvil considera que el valor actual depende linealmente de los términos de error actuales y pasados, donde se supone que los términos de error son mutuamente independientes

y normalmente distribuidos, al igual que el ruido blanco (Peixeiro, 2022). La media móvil aprovecha los errores de pronóstico anteriores en un enfoque de regresión para pronosticar las observaciones futuras en las próximas marcas de tiempo, es decir, cada observación futura puede considerarse como un promedio móvil ponderado de los errores de pronóstico anteriores (Lazzeri, 2020).

Una media móvil de orden q, $MA(q)$, establece que el valor presente es una combinación lineal del promedio de la serie μ, el término de error presente ϵ_t y término de error pasado ϵ_{t-q}, que se fomula como:

$$Y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q}$$

De acuerdo con la ecuación anterior, se puede concluir que cuanto más grande es q, más términos de error pasados afectarán al valor presente. Entonces, para determinar el orden del proceso de promedio móvil se utiliza el gráfico de la función de autocorrelación (ver Figura 6)

Modelo Autorregresivo (AR). El modelo autorregresivo constituye un método de pronóstico para series temporales que se basa exclusivamente en los valores pasados de la propia serie. Este enfoque parte del supuesto de que las observaciones futuras en un instante determinado están vinculadas de manera lineal con las observaciones registradas en momentos anteriores. Dicho de otro modo, en un modelo AR, cada valor de la serie se explica a partir de los valores previos de la misma (Lazzeri, 2020).

En una autorregresión, el valor de salida en la marca de tiempo anterior se convierte en el valor de entrada para predecir el siguiente valor de marca de tiempo, y los errores siguen las suposiciones habituales sobre errores en un modelo de regresión lineal simple. En las autorregresiones, el número de valores de entrada anteriores en la serie de tiempo que se utilizan para predecir el próximo valor de sello de tiempo se llama orden (a menudo nos referimos al orden con la letra p).

$$Y_t = \Theta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t$$

Para determinar el número de rezagos p se analiza el gráfico de Función de Autocorrelación Parcial. Cuando el gráfico de Función de Autocorrelación presenta un decaímiento exponencial es probable que haya un efecto de proceso autoregresivo por lo que se debe evaluar la Función de Autocorrelación Parcial (Peixeiro, 2022).

Modelo Media Móvil Auto Regresivo (ARMA). Un ARMA consta de un $AR(p)$ y un $MA(q)$. Los modelos ARMA proporcionan el modelo lineal más eficiente de series de tiempo estacionarias, ya que son capaces de modelar el proceso desconocido con un número mínimo de parámetros. El modelo ARMA se representa mediante la siguiente

función:

$$Y_t = \Theta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$

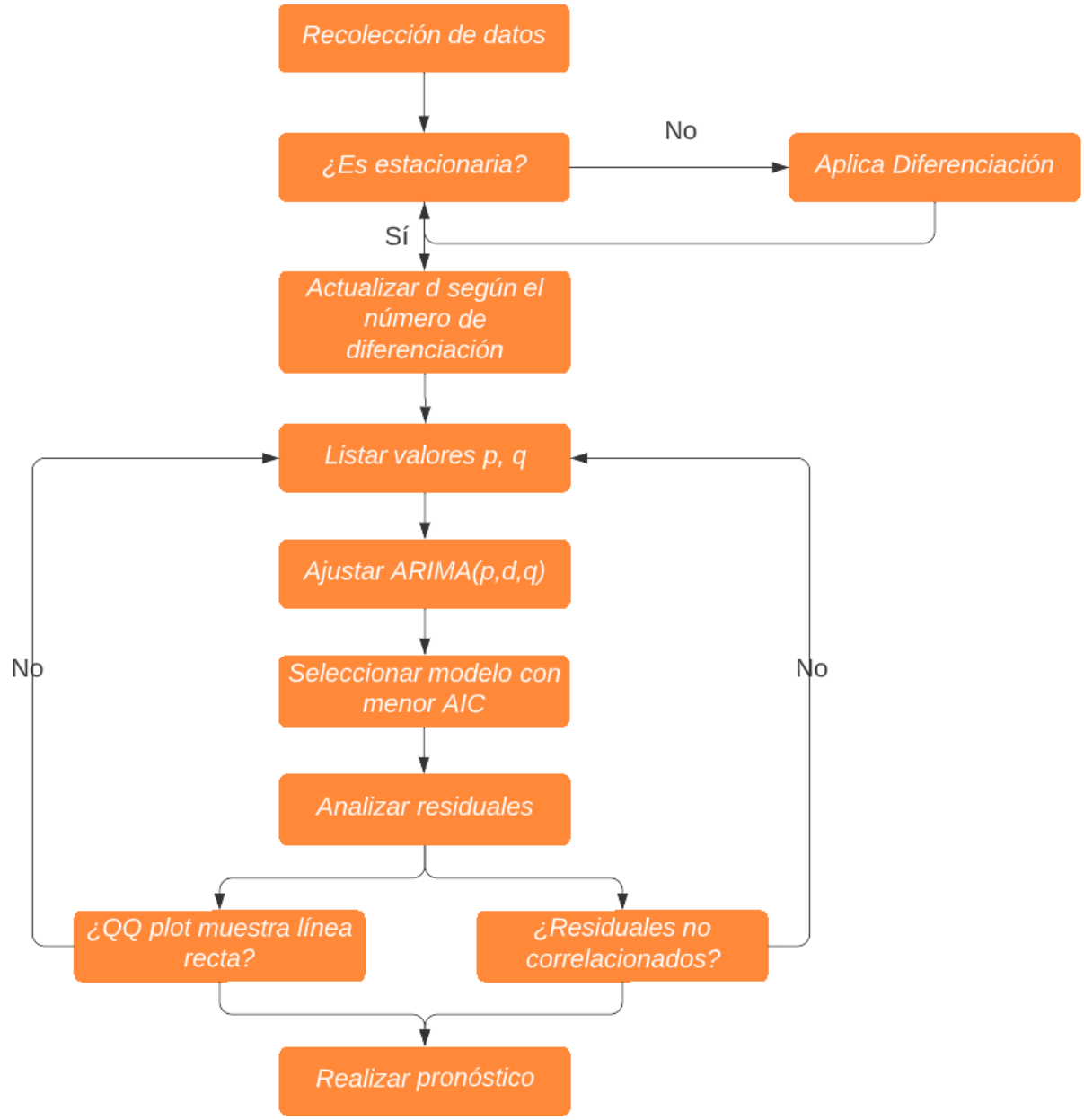
Modelo Media Móvil Auto Regresivo Integrado (ARIMA). Es la combinación de los modelos $AR(p)$ y $MA(q)$, pero en términos de las series diferenciadas cuyo número de rezagos está denotado por d. Se denota por $ARIMA(p, d, q)$. A continuación, se define el modelo:

$$Y'_t = \Theta + \phi_1 Y'_{t-1} + \phi_2 Y'_{t-2} + ... + \phi_p Y'_{t-p} + \epsilon'_t + \theta_1 \epsilon'_{t-1} + \theta_2 \epsilon'_{t-2} + ... + \theta_q \epsilon'_{t-q}$$

donde, Y'_t representa la serie diferenciada y que puede ser diferenciada d veces.

Los modelos ARMA y ARIMA comparten gran parte de sus características, pues ambos se construyen a partir de una estructura autorregresiva general $AR(p)$ y de una media móvil general $MA(q)$. La principal diferencia radica en la incorporación de los conceptos de integración y diferenciación. En la Figura 7 se ilustra el procedimiento de modelado estadístico de series temporales, mostrando de forma resumida todas las etapas descritas previamente.

Figura 7
Proceso de Modelamiento de la Serie de Tiempo



Nota. Adaptado de "Time Series Forecasting in Python", Peixeiro (2022). Pag.146.

2.2.3. *Técnicas de Aprendizaje Automático*

Definición de Aprendizaje Automático. El aprendizaje automático desempeña un rol fundamental en el análisis y la obtención de insights a partir de grandes volúmenes de datos. Se trata de una disciplina de la informática enfocada en el desarrollo de teorías, algoritmos y aplicaciones que permiten a los sistemas aprender de manera similar a los seres humanos (Sugiyama, 2015).

El aprendizaje automático está enfocado en responder cómo una computadora que puede "aprender" tareas específicas como reconocer caracteres, apoyar el diagnóstico de personas con enfermedades graves, clasificar tipos de vino, separar algún material según su calidad.

De manera general, el aprendizaje automático se divide en dos categorías principales: (a) aprendizaje supervisado, en el cual el modelo se entrena con datos de entrada previamente etiquetados, con el propósito de aproximar el mejor clasificador posible $f : X \to Y$ y así poder predecir las etiquetas de nuevos datos; y (b) aprendizaje no supervisado, que se distingue por la ausencia de etiquetas predefinidas, construyendo los modelos a partir del análisis de similitudes entre los datos de entrada con el fin de identificar y formar grupos relevantes y diferenciados (Fernandes y Antonelli, 2018).

En el contexto del aprendizaje supervisado se tienen dos tipos de variables respuesta (cuantitativo y cualitativo). En el caso de las técnicas con variable respuesta cuantitativa son denominadas como técnicas de regresión y las técnicas con variable respuesta cualitativa son denominadas como técnicas de clasificación (Hastie et al., 2009).

Para las técnicas de regresión y clasificación que se aplicarán en el presente trabajo de investigación se debe tener en cuenta una matriz de datos donde se tiene un conjunto de variables útiles para la predicción $X = (X_1, X_2, \ldots, X_p)$ y también una variable dependiente Y para un conjunto de n observaciones (ver Tabla 3).

Tabla 3
Matriz de Datos para Aplicación de Modelos de Aprendizaje Automático

Observación	Respuesta	Regresores			
i	Y	X_1	X_2	...	X_p
1	Y_1	X_{11}	X_{12}	...	X_{1p}
2	Y_2	X_{21}	X_{22}	...	X_{2p}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
n	Y_n	X_{n1}	X_{n2}	...	X_{np}

Regresión Lineal Múltiple (RLM). Sean p variables independientes $X_1, ..., X_p$ y la variable dependiente Y_i (caracterizada por ser de naturaleza continua) que describen a n individuos, entonces la formulación del modelo de RLM para la i-ésima observación se da como sigue:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i$$

Donde los errores $\epsilon_1, ..., \epsilon_n$ son independientes e idénticamente distribuidos con $E(\epsilon_i) = 0$ y $V(\epsilon_i) = \sigma^2$, normalmente, se asume adicionalmente que $\epsilon_i \sim N(0, \sigma^2)$.

La ecuación anterior se puede representar de manera matricial, considerando un vector Y de orden n x 1, la matriz de diseño X de orden n x $(p+1)$ que tiene todas las variables y también un vector de $1's$ al inicio de la matriz que representa al intercepto, tambien se tiene el vector de los coeficientes de regresión β de orden $(p+1)$ x 1 y el término de error ϵ de orden n x 1.

$$Y = X\beta + \epsilon$$

En el modelo de Regresión Lineal Múltiple (RLM) resulta fundamental estimar los coeficientes de regresión β, para lo cual se emplea habitualmente el método de mínimos cuadrados. Este procedimiento consiste en determinar los valores de β que minimicen la suma de los cuadrados de los errores residuales $SCE(\beta)$.

$$SCE(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$$

Derivando la $SCE(\beta)$ con respecto a β e igualando a 0.

$$\hat{\beta} = (X'X)^{-1}X'Y$$

La estimación es válida si se cumple con el supuesto de insesgadez, que se entiende como $\hat{\beta}$ es un estimador insesgado de β ($E(\hat{\beta}) = \beta$) con una matriz varianza covarianza $Cov(E(\hat{\beta})) = \sigma^2(X'X)^{-1}$ (Montgomery et al., 2006). Además, para evaluar la significancia del modelo de regresión se tiene la prueba de hipótesis (ver Tabla 4)

Árboles de Regresión o Clasificación. Según Hastie et al. (2009), los métodos basados en árboles dividen el espacio de características en un conjunto de rectángulos y luego ajustan un modelo simple (como una constante) en cada uno, por ejemplo, sea un problema de regresión con respuesta continua Y e inputs $X1$ y $X2$, primero, se divide el espacio en dos regiones y se modela la respuesta por el promedio de Y en cada región; luego, una o ambas de estas regiones se dividen en dos regiones más, y este proceso continua, hasta que se aplica alguna regla de parada (ver Figura 8).

En la Figura 8, se hace una primera división primero en $X_1 = t_1$. Entonces la región $X_1 \leq t_1$ se divide en $X_2 = t_2$ y la región $X_1 > t_1$ se divide en $X_1 = t_3$. Finalmente, la región $X_1 > t_3$ se divide en $X_2 = t_4$. El resultado de este proceso es una partición en las cinco regiones R_1, R_2, R_3, R_4, R_5; por lo que el modelo de

regresión correspondiente predice Y con una c_m constante en la región R_m.

Tabla 4
Prueba de Hipótesis de Significancia de la Regresión

H0: $\beta_1 = \beta_2 = ... = \beta_p = 0$.
H1: $\beta_j \neq 0$ para un j

$$F_0 = \frac{SCR/p}{SCRes/(n-p-1)} = \frac{MSR}{MSE} \sim F_{p,n-p-1}$$

donde $F_{p,n-p-1}$ es una distribución Fisher con grados de libertad p y $n-p-1$ y sean $SCRes$, SCT y SCR

$$SCRes = Y'Y - \hat{\beta}'X'Y$$

$$SCT = Y'Y - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

$$SCR = SCT - SCRes = \hat{\beta}'X'Y - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

Para rechazar la $H0$ se debe cumplir que $F_0 > F_{p,n-p-1}$.

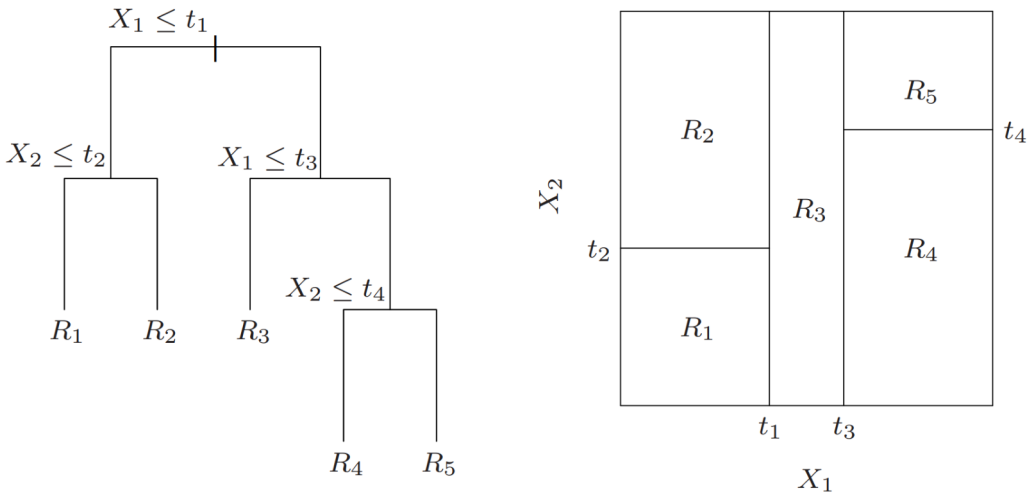
Nota. Adaptado de Ïntroducción al Análisis de Regresión", Montgomery et al. (2006)

Random Forest. Fue creado por Breiman (2001), se enfoca en muestrear tanto las observaciones como las variables de los datos de entrenamiento para desarrollar árboles de decisión independientes y tomar la votación mayoritaria para promediar los problemas de regresión. De esta forma se crean árboles que dependen unos de otros, por lo que se penalizará la precisión.

Según Kamalov (2020), *Random Forest* se basa en agregar el resultado de una colección de árboles de decisión mediante la técnica de bootstrap, con lo que se reduce el sobreajuste característico de los árboles de decisión individuales. Cada árbol de decisión se construye dividiendo de forma recursiva los datos en diferentes valores de las características de entrada.

Figura 8

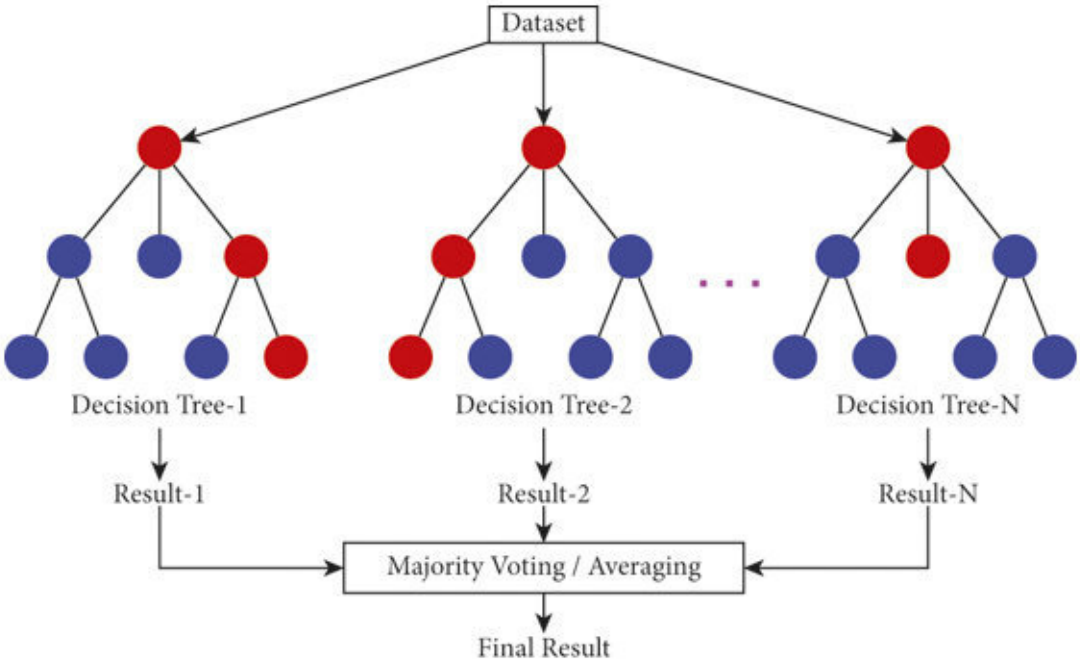
Partición de un Árbol



Nota. Tomado de "The elements of statistical learning: data mining, inference, and prediction", Hastie et al. (2009). Pag.306.

Figura 9

Proceso Random Forest



Nota. Tomado de “Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques”, Khan et al. (2021). Pag.7.

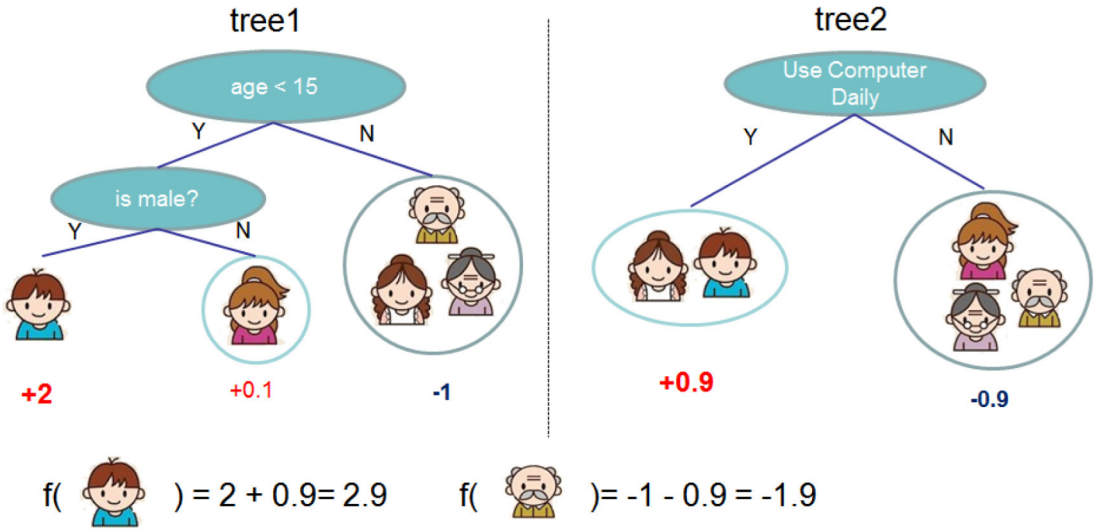
Xtreme Gradient Boosting (XGBoost). Propuesto por T. Chen y Guestrin (2016), este método corresponde a una biblioteca de gradiente reforzado distribuido, optimizada para ofrecer alta eficiencia, flexibilidad y portabilidad (Zhang y Zhan, 2017), incorporando además mejoras sustanciales en los mecanismos de regularización (Bentéjac et al., 2019). Su principio consiste en combinar, de manera iterativa, cientos de árboles de decisión simples con baja precisión para conformar un modelo más robusto, mientras que la función de pérdida se adapta con variaciones específicas que permiten controlar la complejidad de los árboles.

Dado un conjunto de datos con N casos y k variables independientes, $D = \{X_i, Y_i\}_1^N$, el objetivo es encontrar una aproximación de Y, $\hat{Y}$, a través de M funciones aditivas:

$$\hat{Y}_i = \phi(X_i) = \sum_{m=1}^{M} f_m(X_i) \tag{1}$$

donde f es un arbol de decisión que genera una puntuación W. En la Figura 10 se muestra un ejemplo de determinación de f:

Figura 10
Ejemplo de Predicción para GBM



Nota. Tomado de "XGBoost: A Scalable Tree Boosting System", T. Chen y Guestrin (2016).

A continuación, se presenta la Función de Pérdida Regularizada:

$$L_{xgb} = \sum_{i=1}^{N} L(Y_i, \hat{Y}_i) + \sum_{m=1}^{M} \Omega(f_m) \tag{2}$$

donde

- L es la función de pérdida para cada observación, como el error cuadrático para

regresión o la entropía cruzada para clasificación.

- $\Omega(f)$ es el término de regularización definido por:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

tal que γ controla la penalización de la cantidad de hojas, T es el número de hojas del árbol, λ es el parámetro de regularización $L2$ y w son las puntuaciones de salida de las hojas.

Esta función de pérdida se puede integrar en el criterio de división de los árboles de decisión que conducen a una estrategia previa a la poda. Los valores más altos de λ dan como resultado árboles más simples.

Para hacer el ajuste del modelo más eficiente, XGBoost utiliza una expansión de segundo orden de la función de pérdida. La función de pérdida aproximada en el $t-$ésimo paso es:

$$L_{xgb}^{t} = \sum_{i=1}^{N} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

donde:

- $g_i = \frac{\partial L(Y_i, \hat{Y}_i)}{\partial \hat{Y}_i)}$ es el gradiente de la función de pérdida con respecto a la predicción actual.
- $g_i = \frac{\partial^2 L(Y_i, \hat{Y}_i)}{\partial \hat{Y}_i^2)}$ es el hessiano (segunda derivada) de la función de pérdida.

Luego, sea $I_j = \{i | q(x_i) = j\}$ es el conjunto de índices de las observaciones que caen en la hoja j, se puede determinar la siguiente fómula:

$$L_{xgb}^{t} = \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T \tag{4}$$

El valor óptimo de w_j que minimiza la función de pérdida es:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

Por otro lado, XGBoost mejora el rendimiento mediante la creación de histogramas de características. Cada valor de la característica se agrupa en bins, lo que reduce el costo computacional al evaluar posibles divisiones en cada nodo del árbol.

Finalmente, para mayor detalle del algoritmo se puede revisar a T. Chen y Guestrin (2016) y Zhang y Zhan (2017)

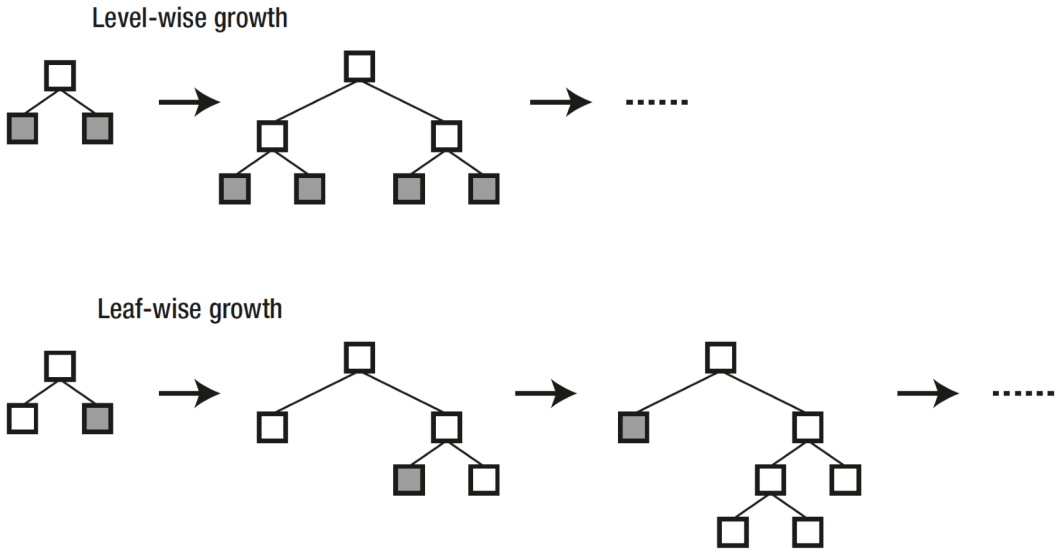
LightGBM. Es un algoritmo de *gradient boosting* basada en árboles, similar a XGBoost, que se lanzó el 2016 como parte del proyecto *Distributed Machine Learning Toolkit* (DMTK) de Microsoft. Diseñado para ser rápido y distribuido, lo que resultó en una mayor velocidad de entrenamiento y un bajo uso de memoria; admite GPU, aprendizaje paralelo y maneja grandes conjuntos de datos (Quinto, 2020).

Ke et al. (2017) propusieron dos técnicas novedosas: *Gradient-Based One-Side Sampling* (GOSS) y *Exclusive Featur Bundling* (EFB), los cuales permitieron mejorar técnicas boosting como XGBOOST evidenciando en experimentos que LightGBM aceleró el proceso de entrenamiento 20 veces más manteniendo el nivel de precisión. Con GOSS se excluye una proporción significativa de instancias de datos con gradientes pequeños debido a que probaron con un tamaño de datos más pequeño pero con gradientes más grandes se puede tener una mayor ganancia de información. Con EFB se agrupan las funciones mutuamente excluyentes sin afectar la precisión.

LightGBM utiliza histogramas para agrupar características continuas en contenedores discretos, que permite mejorar el rendimiento sobre XGBoost, como un uso reducido de la memoria, un costo reducido de cálculo de la ganancia para cada división y un costo de comunicación reducido para el aprendizaje paralelo y generalmente supera a XGBoost en términos de precisión al hacer crecer los árboles en forma de hojas (Quinto, 2020) (ver Figura 11).

Figura 11

Comparación del Crecimiento de Árboles entre Ramas y Niveles



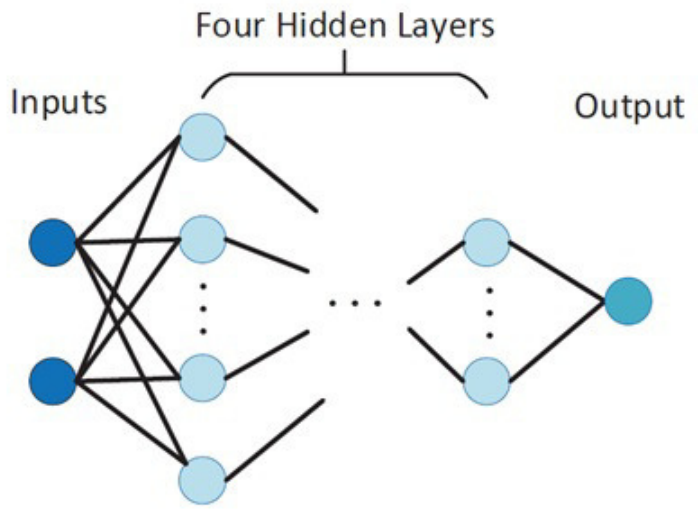
Nota. Tomado de "Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More", Quinto (2020). Pag.146.

Perceptrón Multi Capa (MLP en inglés). Es una técnica de las redes neuronales

artificiales (ver el Anexo B). Kamalov (2020) considera que el perceptrón multicapa es un tipo básico de red neuronal de retroalimentación que consta de una capa de entrada, capa (s) oculta (s) y una capa de salida (Ver Figura 12). Cada capa consta de varios nodos que están interconectados mediante pesos y durante la etapa de entrenamiento del modelo el algoritmo ajusta los pesos de la red para aumentar la precisión.

Figura 12

Estructura del Modelo Perceptrón Multicapas con Cuatro Capas Ocultas



Nota. Tomado de "The Application of Stock Index Price Prediction with Neural Network", Gao et al. (2020). Pag.5.

En la mayoría de aplicaciones comerciales donde se utilizan las redes neuronales se emplean redes con múltiples capas donde las neuronas de cada capa son independientes entre sí y prácticamente se ejecutan en paralelo, además, una neurona de la capa i es input en la capa $i + 1$ pero no se tiene una retroalimentación (no tiene "memoria"), como en las redes recurrentes (Berzal, 2018).

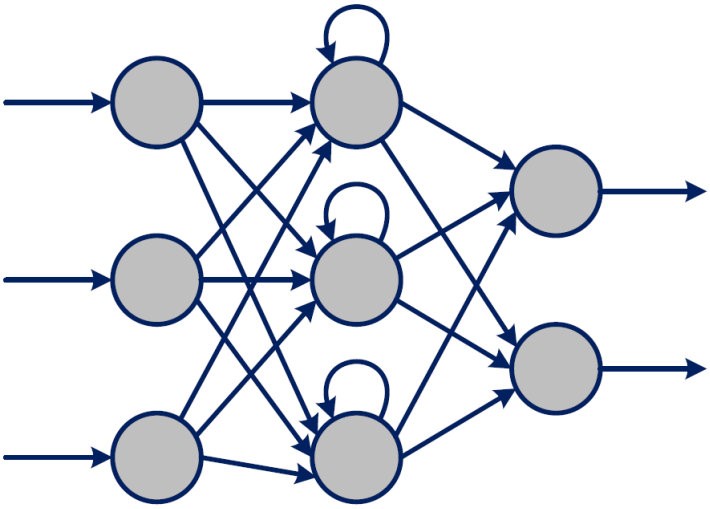
Redes Neuronales Recurrentes (RNN en inglés). De acuerdo con Ignacio (2018), la característica principal de las redes recurrentes es que la información puede persistir introduciendo bucles en el diagrama de la red lo que permite recordar estados previos, lo cual permite manejar series cronológicas.

Según Berzal (2018), si las conexiones entre neuronas de una misma capa se generalizan entonces se generan las redes neuronales con memoria (ver Figura 13).

Para Kamalov (2020), en series de tiempo, las redes neuronales recurrentes (RNN) se diseñaron para procesar datos secuenciales que constan de múltiples pasos de tiempo (ver Figura 14). En una RNN típica, la salida se calcula en función de la entrada actual y el estado oculto anterior, donde el estado oculto se calcula durante el paso del tiempo anterior. Por lo tanto, la red "recuerda" las entradas anteriores mientras calcula la salida

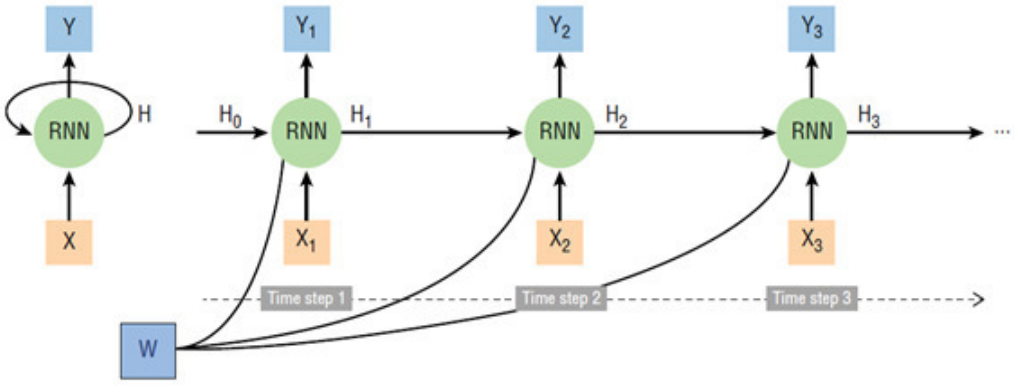
actual. En una RNN regular el valor del gradiente disminuye rápidamente a medida que se propaga hacia atrás en el tiempo, por lo que los pesos de las capas iniciales no se actualizarán de manera efectiva durante la sesión de entrenamiento.

Figura 13
Rede Neuronal Recurrente



Nota. Tomado de "Redes neuronales & deep learning ", Berzal (2018). Pag.207.

Figura 14
Estructura del Modelo Redes Neuronales Recurrentes



Nota. Tomado de "The Application of Stock Index Price Prediction with Neural Network", Gao et al. (2020). Pag.5.

Una diferencia entre las redes multicapa y las redes recurrentes es que las primeras se reinician inmediatamente después de cada activación y es como entrenar una red nueva en cada uso, lo cual no es así en las redes recurrentes porque existe una característica de retroalimentación, aunque surge otra desventaja del riesgo de saturarse y quedarse atascadas en un valor fijo si no se diseñan cuidadosamente (Berzal, 2018).

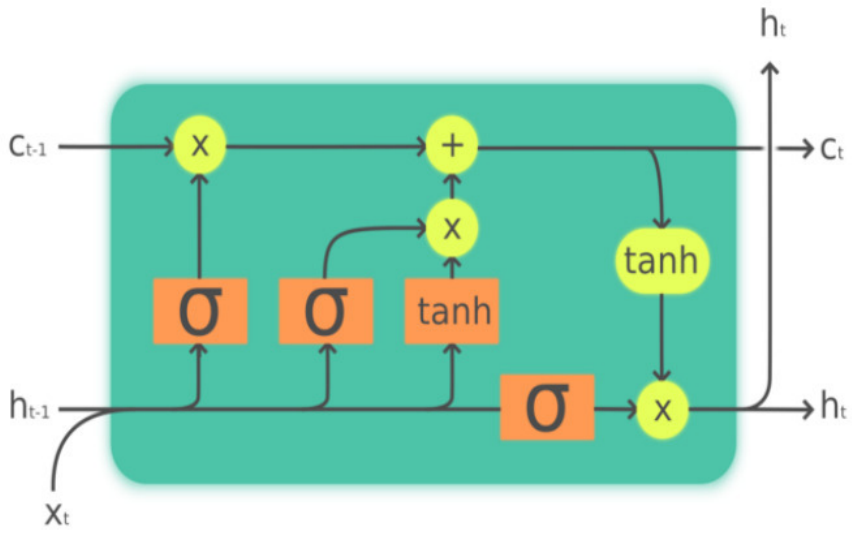
Análogamente, Mañas (2019) afirma que en las redes de múltiples capas que usan la propagación hacia atrás se presenta el problema del desvanecimiento del gradiente, por ejemplo, en las redes neuronales recurrentes los errores que se propagan de una capa a otra disminuyen exponencialmente con el número de capas, y eso impide el ajuste hacia atrás de los pesos de las neuronas basado en esos errores.

Memoria de Corto a Largo Plazo (LSTM en inglés). De acuerdo con Ignacio (2018), las LSTM son un tipo especial de redes recurrentes que aprenden de dependencias largas, por lo que se considera que tienen una memoria de largo plazo a comparación de las redes recurrentes que modelan dependencias a corto plazo. Según Kamalov (2020), LSTM resuelve el problema del gradiente que desaparece al introducir una unidad LSTM en un RNN regular (ver Figura 15).

Una unidad LSTM consta de tres puertas, según Data Science Team (2020): puerta de entrada, puerta de olvido y puerta de salida, que controlan el flujo de información dentro de la unidad .

- **Puerta de entrada:** Controla cuando la información nueva puede entrar en la memoria (Calvo, 2018).
- **Puerta de olvido:** Controla cuando se olvida una parte de la información, lo que permite a la celda discriminar entre datos importantes y superfluos, dejando así sitio para nuevos datos (Calvo, 2018).
- **Puerta de salida:** Controla cuando se utiliza en el resultado de los recuerdos almacenados en la celda (Calvo, 2018).

Figura 15
Estructura del Modelo LSTM



Nota. Tomado de "Aprendizaje profundo", LeCun (2020).

Sea f_t el vector de activación de la puerta de olvido, i_t el vector de activación de la puerta de entrada, o_t el vector de activación de la puerta de salida, h_t el vector de estados ocultos y C_t el vector de estados de la celda.

$$f_t = \sigma_g(W_f X_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma_g(W_i X_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma_o(W_o X_t + U_o h_{t-1} + b_o)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c X_t + U_c h_{t-1} + b_c)$$
$$h_t = o_t \odot tanh(c_t)$$

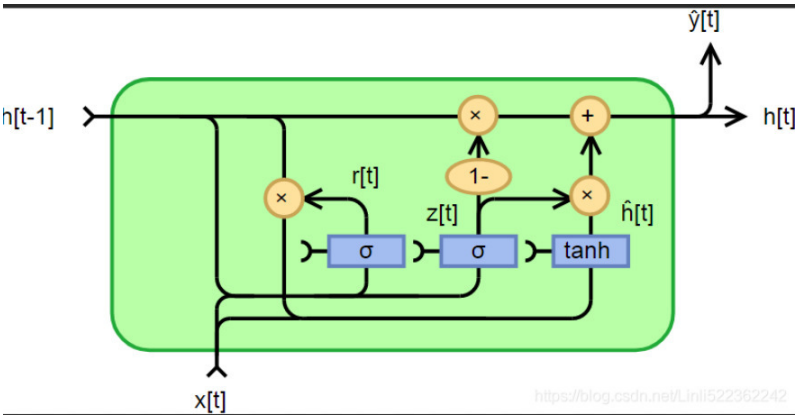
De acuerdo con la Figura 15, una unidad LSTM utiliza un estado de celda C_t para transportar (preservar o remover) información a través de la unidad. La puerta de olvido f_t decide cuanta información proveniente del estado de la celda previa C_{t-1} se quiere preservar, observando la entrada actual y el estado oculto previo, produce un numero entre 0 y 1 como el coeficiente de C_{t-1}. $tanh(W_c X_t + U_c h_{t-1} + b_c)$ calcula un nuevo candidato para actualizar el estado de la celda, y al igual que la puerta de olvido, la puerta de entrada i_t decide cuanto de la actualización será aplicada. Finalmente, la salida h_t estará basada en el estado de celda C_t, pero será puesta a través de una $tanh$ que luego será filtrada por la puerta de salida o_t

Gated Recurrent Unit. Según Cho et al. (2014), es una aplicación de módulos multiplicativos que intenta resolver el problema del desvanecimiento de la gradiente, análogo a LSTM (ver Figura 16). GRU fue introducida en el 2014 por Cho et al. (2014) y Chung et al. (2014).

GRU agrega las puertas de actualización y restablecimiento que mantienen un registro de la información más relevante para el pronóstico, análogo a LSTM, son dos vectores que deciden qué información se pasa a la salida (Sanz, 2023). En la puerta de actualización se determina la información previa que debe transmitirse al siguiente paso de tiempo y en la puerta de restablecimiento se determina la información que debe olvidarse.

Figura 16

Estructura del Modelo GRU



Nota. Tomado de "Aprendizaje profundo", LeCun (2020).

De acuerdo con la Figura 16 y LeCun (2020), sea x_t el vector de entrada, h_t el vector de salida, z_t el vector de puerta de actualización, r_t el vector de puerta de reinicio, ϕ_h es una tangente hiperbólica $tanh$.

$$z_t = \sigma_g(W_z X_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma_g(W_r X_t + U_r h_{t-1} + b_r)$$
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \phi_h(W_h X_t + U_h(r_t \odot h_{t-1}) + b_h)$$

Para ser específicos, z_t fluctúa entre 0 y 1 dado que es una función sigmoide a la suma de dos capas lineales y un sesgo sobre la entrada x_t y el estado previo h_{t-1}. h_t es una combinación convexa de h_{t-1} y $\phi_h(W_h X_t + U_h(r_t \odot h_{t-1}) + b_h)$ y cuando el coeficiente es 1, la unidad actual de salida es una copia del estado previo e ignora la entrada, si es menos que uno, se considerará una nueva información de la entrada. La puerta de reinicio r_t es utilizada para decidir cuanto olvidar de la información del

pasado. En el nuevo contenido de la memoria $\phi_h(W_h X_t + U_h(r_t \odot h_{t-1}) + b_h)$, si el coeficiente en r_t es 0, entonces no se almacena nada de la información del pasado. Si al mismo tiempo z_t es 0, entonces el sistema es completamente reiniciado ya que h_t sólo vería a la entrada.

Entre GRU y LSTM, LSTM tiene una puerta adicional por lo que aumenta su complejidad, pero, no hay evidencia suficiente para decidir sobre la superioridad de algunos de los tipos de red, por lo que se recomienda comenzar con GRU, por su menor complejidad y escalabilidad (Sanz, 2023).

2.2.4. Evaluación de la Capacidad Predictivas de las Técnicas

Medidas de Comparación: Considerando que luego de la aplicación de una técnica de pronóstico se realizan las estimaciones para n casos etiquetándose como $\hat{Y}_i$ y teniendo los datos reales etiquetados como Y_i (ver Tabla 5). Luego, es necesario evaluar el grado de precisión de dichos pronósticos, para ello se muestran a continuación cuatro tipo de medidas de evaluación para el pronóstico.

Tabla 5
Muestra de Datos Reales y Pronósticos

Id	Y_i	$\hat{Y}_i$
1	Y_1	$\hat{Y}_1$
2	Y_2	$\hat{Y}_2$
3	Y_3	$\hat{Y}_3$
4	Y_4	$\hat{Y}_4$
...	...	...
n-1	Y_{n-1}	$\hat{Y}_{n-1}$
n	Y_n	$\hat{Y}_n$

RMSE. Es la raíz del promedio de los cuadrados del error de cada artículo en el periodo i y también se utiliza para comparar la precisión de diferentes métodos de pronóstico. El resultado está en las unidades originales de la información histórica (Galvez, 2016).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

MAPE. Corresponde al promedio de los errores porcentuales en valor absoluto, por lo que únicamente toma en cuenta la magnitud del error y no su signo. El MAPE es una de las métricas más empleadas a nivel internacional; sin embargo, no se recomienda

como criterio principal para seleccionar un método de pronóstico, ya que tiende a generar sesgos que favorecen a aquellos modelos cuyos valores estimados se sitúan por debajo de los observados (Galvez, 2016).

$$MAPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|/Y_i}$$

MAE. Es la media de las diferencias absolutas entre los valores predichos y los valores actuales. El MAE toma los valores absolutos de los errores antes de promediarlos, y es una métrica más intuitiva a comparación del RMSE, sin embargo, este último es más utilizado dado que es más fácil derivar errores al cuadrado que los errores absolutos al momento de buscar minimizar los errores. Por otro lado, con respecto a la interpretación del MAE esta es similar a la interpretación del RMSE porque ambos producen puntajes que están en el mismo rango de valores que los valores reales (Korstanje, 2021).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

Accuracy., considerado como R^2, es una métrica de rendimiento en lugar de una métrica de error. El R^2 es un valor que tiende a estar entre 0 y 1, siendo 0 malo y 1 perfecto, por lo tanto, se puede usar fácilmente como un porcentaje multiplicándolo por 100.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2}$$

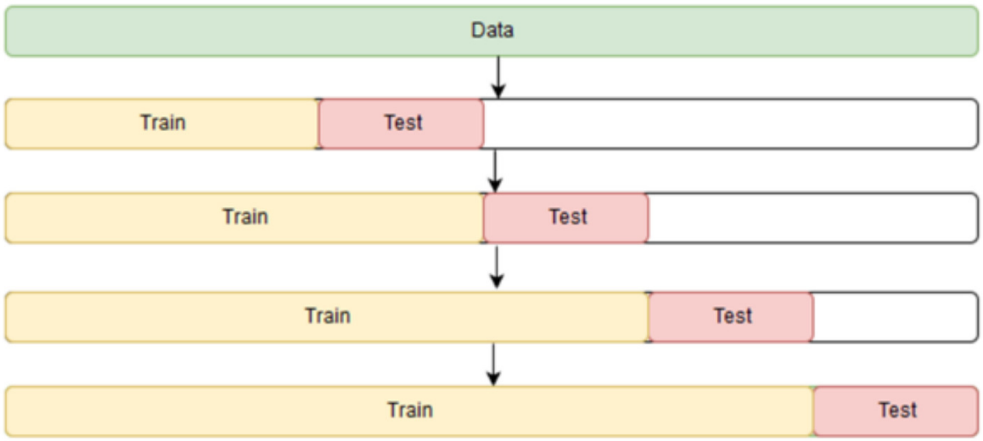
La fórmula calcula una relación entre la suma de los errores al cuadrado y la suma de las desviaciones entre el pronóstico y el promedio. Esto se reduce a un porcentaje de aumento de su modelo sobre el uso del promedio como modelo. Si su modelo es una predicción tan mala como usar el promedio, entonces el R2 será cero Korstanje (2021).

Validación Cruzada para Técnicas de Pronóstico. La validación cruzada tradicional que es aplicada a las técnicas de clasificación o regresión no es una estrategia válida en los modelos de pronóstico debido a la dependencia temporal que existe en las series de tiempo (Howell, 2023).

Se necesita pensar en la validación cruzada para series temporales de manera diferente debido a la secuencialidad en los datos que, por lo general, se correlacionan con el último valor observado (Verma, 2022).

Una vez que el modelo comienza a pronosticar, se verifica la precisión de los valores pronosticados con los valores reales y luego se requieren algunos de los valores reales pronosticados con datos más antiguos para hacer las próximas predicciones (Verma, 2022). En la figura 17 se muestra una representación del diseño de la validación cruzada para series temporales:

Figura 17
Validación Cruzada de una Serie Temporal



Nota. Tomado de "How to improve time series forecasting accuracy with cross-validation?", Verma (2022).

La librería de python **sklean** presenta la función **TimeSeriesSplit** que permite realizar el muestreo para la validación cruzada. De acuerdo con scikit-learn (s.f.), esta función proporciona índices de entrenamiento/prueba para dividir muestras de datos de series temporales que se observan a intervalos de tiempo fijos. En cada división, los índices de prueba deben ser más altos que antes. A diferencia de los métodos tradicionales de validación cruzada, los conjuntos de entrenamiento sucesivos son superconjuntos de los anteriores (ver Figura 17).

2.3. Marco Conceptual

2.3.1. Glosario de Términos Relacionados a las Técnicas

Algoritmo. Conjunto ordenado y finito de operaciones que permite hallar la solución de un problema. RAE. Análisis de Regresión. El análisis de regresión es una herramienta de frecuente uso en estadística. La cual permite investigar las relaciones entre diferentes variables cuantitativas. Esto, mediante la formulación de ecuaciones matemáticas. (Sanchez, 2016)

Boosting. Es un método de aprendizaje conjunto que combina un conjunto de estudiantes débiles en un estudiante fuerte para minimizar los errores de capacitación. En el impulso, se selecciona una muestra aleatoria de datos, se ajusta a un modelo y luego se entrena secuencialmente, es decir, cada modelo intenta compensar las debilidades de su predecesor. Con cada iteración, las reglas débiles de cada clasificador individual se combinan para formar una regla de predicción sólida.(IBM, 2021)

Ingeniería de Variables. Es el proceso de transformar datos en características para que actúen como entradas para los modelos de aprendizaje automático, de modo que las características de buena calidad ayuden a mejorar el rendimiento general del modelo. (Lazzeri, 2020)

Aprendizaje Automático. Es una forma de inteligencia artificial que permite a un sistema aprender de los datos en lugar de aplicar programación explícita; sin embargo, no es un proceso simple. Utiliza una variedad de algoritmos que iterativamente aprenden de los datos y predecir mejor los resultados. (Hurwitz y Kirsch, 2018)

Precisión. En estadística, la precisión del pronóstico es el grado de cercanía entre la expresión de cantidad y el valor real de esa cantidad. (Vermorel, 2013)

Series de Tiempo. Las series de tiempo son colecciones de observaciones sobre un determinado fenómeno efectuadas en sucesivos momentos del tiempo, usualmente equiespaciados.

Capítulo III

Metodología

3.1. Tipos de Investigación

3.1.1. *Nivel de Investigación*

Es una investigación aplicada ya que tiene por finalidad la solución de un problema relacionado al pronóstico de los precios y rendimientos de las acciones y también a la clasificación de los rendimientos en base a la información histórica de las cotizaciones de precios. Para ello, se ha realizado una búsqueda, consolidación y aplicación de las técnicas de pronóstico, en el ámbito de series de tiempo y aprendizaje automático para el pronóstico de los precios de la acción de una empresa peruana que cotiza en la NYSE.

3.1.2. *Diseño de Investigación*

La investigación es de diseño experimental porque en el presente trabajo de investigación se manipularán las variables independientes debido a las pruebas de diversas técnicas de modelamiento desde el enfoque de series de tiempo y de las técnicas de aprendizaje automático, dichas técnicas serán evaluadas y comparadas considerando la mejora en el pronóstico de los precios de una acción peruana que cotiza en la NYSE.

3.1.3. *Método de Investigación*

Se utiliza el "razonamiento deductivo, se parte de la teoría, y se derivan expresiones lógicas denominadas hipótesis que el investigador somete a prueba", el presente trabajo de investigación aborda el estudio de las técnicas de pronóstico a nivel de las series de tiempo y de aprendizaje automático, teorías que forman parte de la literatura científica, las cuales son aplicadas para el pronóstico de precios de una acción peruana que cotiza

en la NYSE. Se parte de una hipótesis donde se afirma que una de las técnicas de aprendizaje automático logran una mejor predicción y finalmente es contrastada con la realidad a través del uso de los datos de cotizaciones de precios.

3.2. Población y Muestra

3.2.1. *Población*

La unidad de análisis es el precio de las acciones de una empresa peruana listada en la NYSE. Entonces, la población de estudio se refiere a los precios históricos diarios de las acciones de una empresa peruana, que se obtienende desde Yahho Finance (`https://finance.yahoo.com/quote/BAP?p=BAP&.tsrc=fin-srch`).

3.2.2. *Muestra*

Los precios de la acción peruana que cotiza en la Bolsa de Valores de Nueva York, entre el 1 de enero de 2020 y el 1 de enero de 2023, que representa un tamaño muestral de 756 observaciones de precios diarios.

3.2.3. *Muestreo*

La selección de la muestra consistió en la definición de las fechas de la serie temporal, entre el 1 de enero de 2020 y el 1 de enero de 2023, la cual se realizó por conveniencia o intencional garantizando que se cumpla una cantidad de observaciones suficiente para la aplicación de las técnicas y para obtener resultados consistentes. Vale precisar de que en el caso de las series temporales de precios diarios es importante garantizar que dentro de un intervalo de tiempo se tengan los precios para cada día con el fin de que las técnicas propuestas puedan reconocer el patrón de los precios, mientras que si se tomasen aleatoriamente los precios las técnicas no podrían reconocer los patrones internos de la serie temporal.

3.3. Operacionalización de Variables

3.3.1. *Identificación de Variables*

Variable Independiente. Las variables independientes son las técnicas de series de tiempo y las técnicas de aprendizaje automático.

Respecto a la variable independiente de técnicas de series de tiempo se ha definido lo siguiente:

- **Definición Conceptual.** Se refiere a métodos estadísticos y modelos matemáticos diseñados para analizar datos en secuencia temporal, con el objetivo de capturar patrones y realizar predicciones. Incluyen métodos como el ARIMA, GARCH, y modelos de suavizado exponencial.

- **Definición Operacional.** Para la aplicación de las técnicas de series de tiempo se utilizará únicamente la variable del precio en la historia de tiempo seleccionada y se analizará las características de la serie como las tendencias, estacionalidad y comportamientos cíclicos sobre la serie. Una vez analizada el comportamiento de la serie se aplicarán las diversas técnicas se han propuesto.

  Se medirán a través de la implementación de modelos específicos de ARIMA y la precisión de sus predicciones en series históricas de precios de acciones o retornos financieros. Los resultados de cada modelo se evaluarán con métricas como el RMSE, MAPE y MAE.

Respecto a la variable independiente de técnicas de aprendizaje automático se ha definido lo siguiente:

- **Definición Conceptual.** El aprendizaje automático es la disciplina que permite a los ordenadores "aprender" directamente a partir de los datos. A diferencia de la programación tradicional, donde se codifican de manera explícita soluciones específicas para cada problema, el aprendizaje automático se centra en diseñar algoritmos generales capaces de identificar y extraer patrones en distintos tipos de datos (Bobadilla, 2020).

  Conjunto de algoritmos de aprendizaje automático que permiten la identificación de patrones complejos en los datos y la creación de modelos predictivos a partir de ellos. Esto puede incluir métodos supervisados como el Random Forest, el XGBoost, y redes neuronales, así como métodos no supervisados o semi-supervisados.

- **Definición Operacional.** Se evaluarán mediante la implementación de diferentes algoritmos de aprendizaje automático en el conjunto de datos, evaluando su capacidad predictiva con métricas como la precisión, F1-score, o el RMSE, según el tipo de variable dependiente.

Variable Dependiente. La variable dependiente es el pronóstico de precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.

- **Definición Conceptual.** El pronóstico del precio es el arte y la ciencia de seleccionar y utilizar los principales eventos históricos y las tendencias relacionadas con los activos subyacentes para analizar los eventos actuales y pronosticar el precio futuro en base a información adicional relevante que podamos recopilar. (Tjung, 2010)

- **Definición Operacional.** El pronóstico de los precios y rendimientos se realizará para la acción de una empresa peruana listada en la Bolsa de Valores de Nueva York. Este pronóstico se abordó desde tres enfoques: (i) la variable dependiente de naturaleza continua que es el precio en el día t, (ii) la variable dependiente de naturaleza continua que es el rendimiento hasta el día t y (iii) la variable dependiente de naturaleza categórica que indica si hasta el día t se garantiza un rendimiento mínimo viable o no.

  Dependiendo de la naturaleza de la variable dependiente se implementaron modelos de series de tiempo ARIMA y algoritmos de aprendizaje automático (como Random Forest o LSTM) para calcular el precio proyectado de la acción en un horizonte de tiempo determinado. Los resultados se evaluaron en términos de precisión predictiva utilizando métricas como RMSE, MAE, MAPE para predicciones continuas y GINI, F1-Score para predicciones categóricas.

3.3.2. *Tabla de Operacionalización de Variables*

En la Figura 18 se muestra la operacionalización de las variables del presente trabajo de investigación:

Figura 18

Operacionalización de las Variables.

Variables	Definición conceptual	Definición operacional	Dimensiones	Indicadores	Cálculo
Variable dependiente: Pronóstico del precio de la acción de una empresa peruana que cotiza en la Bolsa de Valores de *New York*	Es el arte y la ciencia de seleccionar y utilizar los principales eventos históricos y las tendencias relacionadas con los activos subyacentes para analizar los eventos actuales y pronosticar el precio futuro en base a información adicional relevante que podamos recopilar. (Tjung, 2010).	La predicción del precio se realizará para la acción de un empresa peruana que cotizan en la Bolsa de Valores de *New York*, dicha predicción se realizará considerando una ventana óptima de tiempo.	Precio	Precio futuro en el día t	P_t
			Rendimiento	Rendimiento futuro en el día t	$R_t = \frac{P_t}{P_0}$
			Categórico	Rangos de rentabilidad esperada	$\begin{cases} R_t > Umbral: Bueno \\ R_t > 0: Regular \\ R_t \leq 0: Malo \end{cases}$
Variable independiente Técnicas de series de tiempo	Se refiere a métodos estadísticos y modelos matemáticos diseñados para analizar datos en secuencia temporal, con el objetivo de capturar patrones y realizar predicciones. Incluyen métodos como el ARIMA, GARCH, y modelos de suavizado exponencial	Para la aplicación de las técnicas de series de tiempo se utilizará únicamente la variable del precio en la historia de tiempo seleccionada y se analizará las características de la serie como las tendencias, estacionalidad y comportamientos cíclicos sobre la serie. Una vez analizada el comportamiento de la serie se aplicarán las diversas técnicas se han propuesto.	Técnicas de series de tiempo: Media Móvil, Autoregresión, ARMA y ARIMA	Precisión de la predicción: RMSE y MAPE	Serie de tiempo: $Y_{T+k} = Y_T + e_{T+k}$ Indicadores de precisión: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$ $MAPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Y_i - \hat{Y}_i \rvert}{Y_i}}$
Variable independiente Técnicas de *machine learning*	*Machine Learning* es la ciencia que hace que los ordenadores "aprendan" a partir de los datos. En vez de programar, paso a paso, cada solución específica para cada necesidad planteada, tal y como se realiza en el enfoque de la programación convencional, el área de *machine learning* está enfocada al desarrollo de algoritmos genéricos que pueden extraer patrones de diferentes tipos de datos. (Bobadilla, 2020)	La aplicación de tres tipos de técnicas de *Machine Learning*: regresión, *boosting* y redes neuronales permitirá predecir el precio de las acciones, las cuales serán comparadas en función a su nivel de precisión y se elegirá la que presente un mejor indicador de precisión (menor RMSE y menor MAPE).	Regresión lineal múltiple, *Random Forest*, XGBoost, LightGBM, Perceptrón multicapa, Redes neuronales LSTM, GRU	Precisión de la predicción: RMSE y MAPE	Regresión lineal: $Y = \beta_0 + X_1\beta_1 + \cdots X_k\beta_k + e$ Boosting y Redes neuronales: $Y = f(X_1, \ldots X_k) + e$ Indicadores de precisión: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$ $MAPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Y_i - \hat{Y}_i \rvert}{Y_i}}$

3.3.3. *Descripción de Indicadores de las Variables*

Para la variable dependiente se han definido tres indicadores relacionados al precio futuro, los cuales representan el target en la aplicación de las técnicas de pronóstico.

- **Precio en el futuro día T:** Se considera el precio en el futuro día T, P_T (ver Tabla 6).

Tabla 6
Determinación del Precio en el Futuro Día T

Día	Precio$_t$	Precio Futuro$_{t+T}$
0	P_0	P_T
1	P_1	P_{T+1}
2	P_2	P_{T+2}
...	...	...
t	P_T	P_{2T}
t+1	P_{T+1}	P_{2T+1}
t+2	P_{T+2}	P_{2T+2}
...	...	...
N	P_N	P_{N+T}

- **Rendimiento acumulado al día T:** Se considera el rendimiento obtenido después de haber pasado T días con respecto al día actual (ver Tabla 7).

Tabla 7
Determinación del Rendimiento en el Futuro Día T

Día	Precio$_t$	Precio Futuro$_{t+T}$	Rendimiento$_t$
0	P_0	P_T	$P_T/P_0 - 1$
1	P_1	P_{T+1}	$P_{T+1}/P_1 - 1$
2	P_2	P_{T+2}	$P_{T+2}/P_2 - 1$
...	...	...	...
t	P_T	P_{2T}	P_{2T}/P_T
t+1	P_{T+1}	P_{2T+1}	$P_{2T+1}/P_{T+1} - 1$
t+2	P_{T+2}	P_{2T+2}	$P_{2T+2}/P_{T+2} - 1$
...	...	...	...
N	P_N	P_{N+T}	$P_{N+T}/P_N - 1$

Entonces, se puede definir el rendimiento acumulado hasta en los siguientes k días partiendo del tiempo actual t como R_{tk}:

$$R_{tk} = \frac{P_{t+k}}{P_t}$$

- **Nivel del rendimiento:** En base a los rendimiento calculados al día T se define

tres categorías donde 0 indica pérdida al día T, 1 indica ganancia menor al $5\,\%$ y 2 indica ganancia mayor o igual al $5\,\%$.

Para las variables independientes se han definido tres indicadores de precisión que permiten comparar las técnicas de pronósticos que son MAE, RMSE y MAPE, las cuales fueron definidos en el capítulo que corresponde al marco teórico, precísamente, medidas de evaluación de la calidad del modelo.

3.4. Técnicas e Instrumentos de Recolección de Datos

3.4.1. *Técnicas*

De acuerdo con Arias (2010), "se entenderá por técnica de investigación, el procedimiento o forma particular de obtener datos o información", en ese sentido, la técnica de recolección de datos se basó en extraer las cotizaciones históricas de la acción peruana que cotiza en la Bolsa de Valores de Nueva York, en un periodo determinado de tiempo. Las series temporales fueron obtenidas a través de la página web de Yahoo Finance de la nomenclatura BAP que corresponden a las acciones del grupo Credicorp que se cotizan en la bolsa de valores de Nueva York.

3.4.2. *Instrumentos*

De acuerdo con Arias (2010), "un instrumento de recolección de datos es cualquier recurso, dispositivo o formato (en papel o digital), que se utiliza para obtener, registrar o almacenar información"; en ese sentido, en la investigación se ha utilizado una laptop con acceso a internet que permita el acceso a la página web de Yahoo Finance. También, se ha utilizado Google Colab, la cual permitió acceder a los datos de los precios de Yahoo Finance de una manera automática a través de la API de *Yahoo Finance* mediante código python.

3.4.3. *Validación y Confiabilidad del Instrumento*

Los instrumentos para la obtención de información de la presente investigación corresponden a herramientas tecnológicas y hardware, las cuales son herramientas determinísticas y no dependen del juicio del investigador, por lo que no sería necesario una validación. En Kingaby (2022) se detalla la API de Yahoo Finance, además en

diveras publicaciones se utiliza como fuente de obtención de precios (Dubey et al., 2022), (Achyutha et al., 2022), entre otros.

3.5. Análisis e Interpretación de la Información

Lazzeri (2020) planteó un marco ágil e iterativo para desarrollar proyectos analíticos de predicción mediante series de tiempo de forma eficiente. La Figura 19 ilustra las principales fases que integran este tipo de proyectos, las cuales comprenden: (a) comprensión del negocio y definición de métricas de desempeño, (b) recolección de datos, (c) análisis exploratorio, (d) preprocesamiento e ingeniería de características, (e) construcción y selección del modelo, (f) despliegue del modelo y (g) validación y aceptación de la solución de pronóstico.

Además, Kumar et al. (2021) realizó un flujo de trabajo estandarizado enfocado al pronóstico de precios en el mercado de valores, la cual se alínea con el presente trabajo de investigación (ver Figura 20) . El presente flujo coincide en algunos puntos con la metodología de trabajo detallada en la Figura 19.

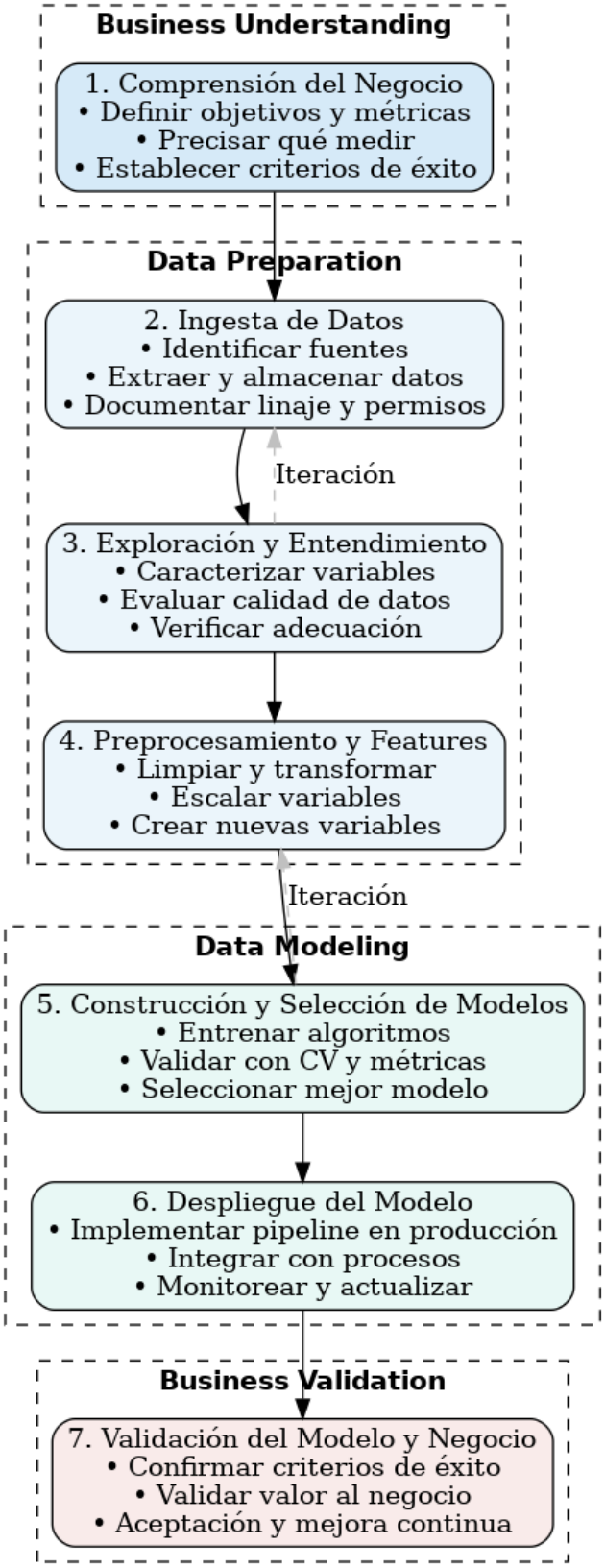
3.5.1. *Entendimiento del Negocio y Definición de las Métricas de Rendimiento*

Se describe el aspecto empresarial que se debe comprender y considerar antes de tomar una decisión de inversión, también se explica cómo calificar el problema de negocio para garantizar que el análisis predictivo y el aprendizaje automático sean realmente efectivos y aplicables; además, se debe especificar las métricas clave de rendimiento, los datos y las características que servirán como objetivos del modelo y cuyas métricas relacionadas se utilizarán para determinar el éxito del proyecto, y también identificar las fuentes de datos relevantes que el negocio tiene acceso o necesita obtener (Lazzeri, 2020).

El presente estudio tiene como objetivo estimar los precios futuros de las acciones, respondiendo a la necesidad de los inversionistas de contar con predicciones más certeras que les faciliten la toma de decisiones y, en consecuencia, incrementar la rentabilidad de sus inversiones. Para la evaluación del desempeño se consideraron tres métricas centradas en la capacidad predictiva de los modelos (MAE, RMSE y MAPE); no obstante, también podrían emplearse otras medidas, como aquellas orientadas a la rentabilidad obtenida, la valoración cualitativa de los inversionistas respecto al uso de estas técnicas de predicción, entre otras.

Figura 19

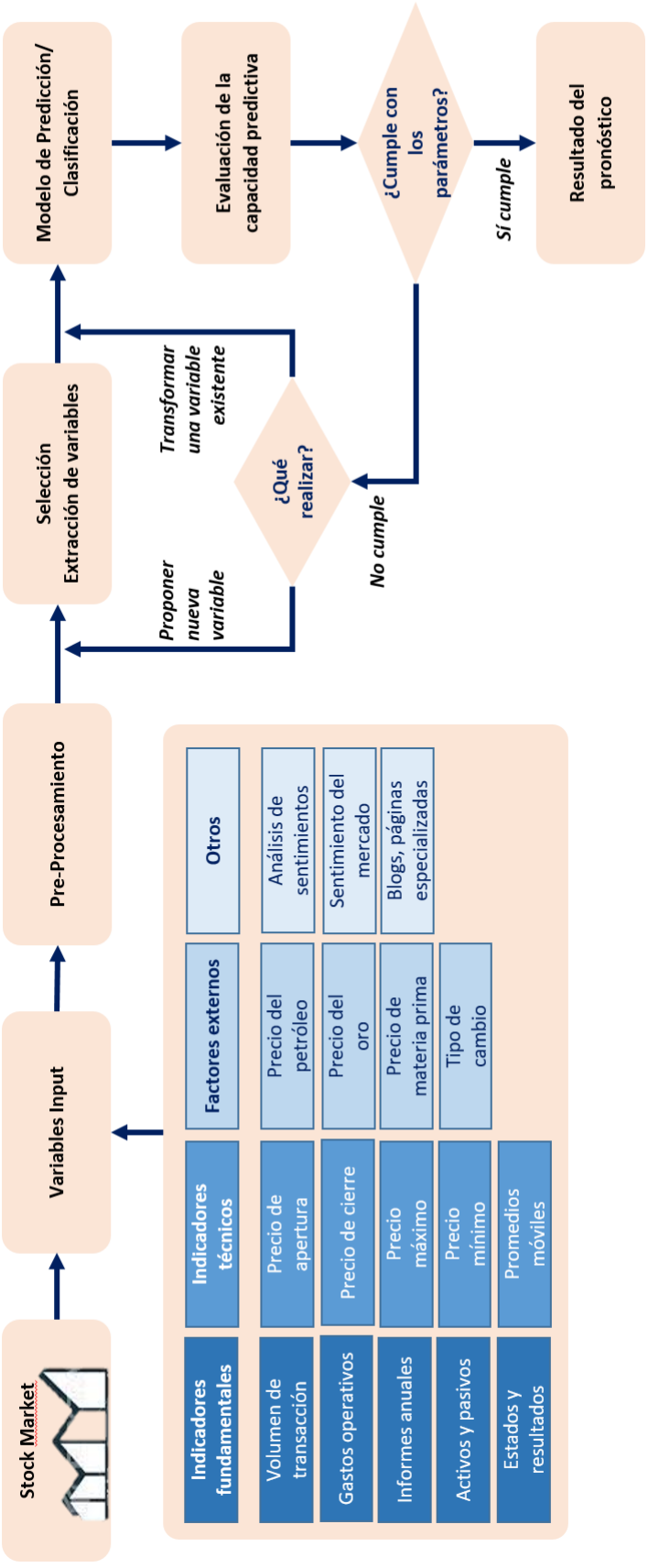
Etapas para la Predicción de Series Temporales



Nota. Adaptado de "Machine learning for time series forecasting with Python", Lazzeri (2020). Pag.7

Figura 20

Estructura General para el Mercado de Acciones



Nota. Adaptado de "Stock market forecasting using computational intelligence: A survey", Kumar et al. (2021). Pag.7

3.5.2. *Obtención de los Datos*

Es el proceso de obtener e importar los datos Lazzeri (2020). Para esta parte del proyecto se partió de las cotizaciones de los precios históricos.

Cotizaciones en los Precios Históricos. A través de Yahoo Finance se obtuvo información histórica de precios dentro de un intervalo de tiempo específico. Los campos disponibles en esta fuente incluyen: la fecha de la cotización, el precio de apertura del día, el precio máximo alcanzado, el precio mínimo registrado, el precio de cierre al finalizar la jornada y el volumen total negociado (ver Tabla 8).

Tabla 8
Estructura de Información de Cotizaciones

Campo	Descripción
Date	Fecha de la cotización
Open	Precio de apertura de la acción en el día
High	Precio más alto de la acción en el día
Low	Precio más bajo de la acción en el día
Close	Precio de cierre de cotización de la acción en el día
Volume	Número de accionesNumber of shares that changed hands

3.5.3. *Análisis Exploratorio de los Datos*

Implica resumir las características principales de un conjunto de datos, incluido su tamaño, precisión, patrones iniciales en los datos y otros atributos Lazzeri (2020).

Se analizó la serie de tiempo de los precios de la acción peruana listada en la Bolsa de Valores de Nueva York, donde se determinó las características de la serie de tiempo como la componente de tendencia, estacionalidad y componente cíclico. Adicionalmente, se obtendrán estadísticos descriptivos de la serie para comprender la serie. Posteriormente, se utilizará el conocimiento para la aplicación de las técnicas de series de tiempo y la elaboración de posibles variables adicionales que puedan ser útiles para la predicción del precio.

3.5.4. *Pre-procesamiento de los Datos e Ingeniería de Variables*

En esta fase, los científicos de datos se encargan de depurar el conjunto de datos, eliminando valores atípicos y gestionando los datos faltantes, además de generar nuevas características a partir de la información en bruto con el propósito de optimizar el rendimiento de los modelos de aprendizaje automático.

Inicialmente, se realizó la transformación para determinar la variable dependiente. En la Tabla 9 se presenta un ejemplo del tratamiento para determinar la variable dependiente para los tres tipos de variables: *Precio Futuro* (Precio$_{t+T}$), *Rendimiento Acumulado* (Rendimiento$_{t+T}$) y la categoría del rendimiento acumulado considerando si superan un rendimiento mínimo definido donde se puede tomar los valores de $Bueno$, $Regular$ y $Malo$ (Categoría):

Tabla 9
Método de Cálculo de las Variables Dependientes

Día	Precio$_t$	Futuro$_{t+T}$	Rendimiento$_{t+T}$	Categoría
0	P_0	P_T	$R_{0T} = P_T/P_0 - 1$	CAT R_{0T}
1	P_1	P_{T+1}	$R_{1T} = P_{T+1}/P_1 - 1$	CAT R_{1T}
2	P_2	P_{T+2}	$R_{2T} = P_{T+2}/P_2 - 1$	CAT R_{2T}
...	...	...	...	
t	P_T	P_{2T}	$R_{TT} = P_{2T}/P_T$	CAT R_{TT}
t+1	P_{T+1}	P_{2T+1}	$R_{(T+1)T} = P_{2T+1}/P_{T+1} - 1$	CAT $R_{(T+1)T}$
t+2	P_{T+2}	P_{2T+2}	$R_{(T+2)T} = P_{2T+2}/P_{T+2} - 1$	CAT $R_{(T+2)T}$
...	...	...	...	
N	P_N	P_{N+T}	$R_{NT} = P_{N+T}/P_N - 1$	CAT R_{NT}

Respecto a las características que se han utilizado como variables predictoras para ser utilizados en las técnicas propuestas han presentado determinadas transformaciones, pero todas son generadas de los precios históricos detallados en la Tabla 8.

- Precio de cierre en el día de la cotización t ($Precio_t$).
- Precios históricos antes del día de cotización: $Precio_{t-k}$, $Precio_{t-k-1}$....$Precio_t$.
- Promedio móvil de los precios históricos antes del día de cotización: $Precio_{t-k}$, $Precio_{t-k-1}$....$Precio_t$.

$$MA_\bar{P}recio_t = \frac{\sum_{i=t-k}^{t} Precio_i}{k}$$

- Ratio entre el Precio de Cierre y el Precio Máximo, en el día de la cotización t (Rat_Max_Cierre).

$$Rat_Max_Cierre_t = \frac{Precio_t}{Precio_Max_t}$$

- Ratio entre el Precio de Cierre y el Precio Mínimo, en el día de la cotización t (Rat_Min_Cierre).

$$Rat_Min_Cierre_t = \frac{Precio_Close_t}{Precio_Min_t}$$

- Rango entre el Precio Máximo y el Precio Mínimo, en el día de la cotización t ($Rango_Max_Min$).

$$Rango_Max_Min_t = Precio_Max_t - Precio_Min_t$$

- Rango entre el Precio Máximo y el Precio Mínimo, en el día de la cotización t ($Rango_Max_Min$).

$$Rango_Max_Min_t = Precio_Max_t - Precio_Min_t$$

- Rendimiento Diario en el día de la cotización t ($Rendimiento_t$).

$$Rendimiento_t = Ln\frac{Precio_t}{Precio_{t-1}}$$

- Volatilidad de los rendimientos históricos diarios en el día de la cotización entre $t-k$ y t (Vol_t).

$$Vol_t = \frac{\sum_{i=t-k}^{T}(R_i - \bar{R})^2}{k}$$

  donde k es el número de rezagos o información histórica a utilizar para el cálculo de la desviación y R es el rendimiento diario.

- El Índice de Fuerza Relativa (RSI) es un indicador técnico utilizado en el análisis bursátil para evaluar la magnitud de las recientes ganancias y pérdidas de un activo en un período específico con el fin de determinar si se encuentra en condiciones de sobrecompra o sobreventa. El RSI oscila entre 0 y 100, y valores por encima de 70 suelen indicar sobrecompra, mientras que valores por debajo de 30 sugieren sobreventa.

$$\text{RSI} = 100 - \frac{100}{1 + \frac{\text{Ganancia Promedio}}{\text{Pérdida Promedio}}}$$

donde:

$$\text{Ganancia Promedio} = \frac{\sum_{\text{días de ganancia}} \text{Cambio Positivo}}{n}$$

$$\text{Pérdida Promedio} = \frac{\sum_{\text{días de pérdida}} |\text{Cambio Negativo}|}{n}$$

- Las Bandas de Bollinger son un indicador técnico utilizado para medir la volatilidad del precio de un activo y determinar posibles puntos de compra o venta. Están compuestas por tres líneas:

$$\text{Banda Superior} = \text{SMA} + (k \times \sigma)$$

$$\text{Banda Inferior} = \text{SMA} - (k \times \sigma)$$

  Donde SMA es la Media Móvil Simple, k es el número de desviaciones estándar y σ es la desviación estándar del precio durante el mismo período.

- El Oscilador Estocástico es un indicador técnico que mide la posición del precio de cierre de un activo en relación con su rango alto-bajo durante un período específico, generalmente de 14 días. Este indicador ayuda a identificar condiciones de sobrecompra o sobreventa.

$$\%K = \frac{\text{Precio de Cierre Actual} - \text{Mínimo Bajo}}{\text{Máximo Alto} - \text{Mínimo Bajo}} \times 100$$

3.5.5. *Determinación de la Variable Dependiente*

En primer lugar, se determinó la ventana óptima de pronóstico a través de un método iterativo que se aplicó a la información histórica de los precios. Los pasos seguidos son los siguientes:

- Se seleccionó el precio del día $i-$ésimo de toda la ventana diaria de modelamiento ($Precio_i$), donde $i = 1, ..., 756$.
- Se le asignó el precio futuro del día $j-$ésimo ($Precio_{i+j}$), donde $j = 1, ..., 30$.
- Se calculó el rendimiento ($Rendimiento_{ij} = Precio_{i+j}/Precio_i - 1$)
- Si el rendimiento superó un cierto umbral k de rendimiento mínimo requerido, entonces se contabilizó como un evento exitoso.

Luego, se contabilizó el porcentaje de días en la que los rendimientos ($Rendimiento_{ij}$) superaron el umbral k para cada ventana futura j (ver Figura 23). A través de dicho análisis gráfico se buscó determinar en que día j de la ventana de predicción el porcentaje de días se estabiliza, de tal forma que se considere una ventana óptima de predicción.

En segundo lugar, al determinar la ventana óptima de predicción en un día J (donde J está en el rango de 1 a 30 días) se calcularon las variables objetivos para los tres tipos de problemas específicos que busca resolver el presente trabajo de investigación.

- Precio futuro en el día J ($Precio_{t+J}$).
- Rendimiento futuro en el día J ($Rendimiento_{t+J} = Precio_{t+J}/Precio_t - 1$).
- Clasificación del rendimiento futuro en el día J:
  - Si el $Rendimiento_{t+J} >= 6\,\%$ entonces se considera **Bueno**
  - Si el $Rendimiento_{t+J} >= 0\,\%$ entonces se considera **Regular**
  - Caso contrario se considera **Malo**.

3.5.6. *Construcción y Selección del Modelo*

Las técnicas de modelamiento que se han utilizado para el pronóstico de las variables dependientes continuas (precio y rendimiento) son las siguientes:

- Técnicas de series de tiempo: Media móvil, proceso autoregresivos, proceso ARMA, proceso ARIMA.
- Técnicas de aprendizaje automático: Regresión lineal múltiple, Árboles de regresión, Random Forest, XGboost, LightGBM, Perceptrón multicapa, Redes neuronales recurrentes LSTM y GRU.

Los indicadores que se han utilizado para evaluar la capacidad predictiva de los modelos son:

- RMSE: Mide la cantidad de error que hay entre dos conjuntos de datos.
- MAPE: es una desviación porcentual absoluta media. Es una medida de precisión de predicción de un método de predicción en estadística.
- MAE: es el promedio de las diferencias absolutas entre los valores predichos y los valores observados, proporcionando una medida clara de error en las mismas unidades que los datos. Cuanto menor sea el MAE, mejor es el rendimiento del modelo.

Por otro lado, las técnicas de modelamiento que se han utilizado para la clasificación de la variable dependiente categórica (categoría de un rendimiento aceptable) son las siguientes:

- Técnicas de aprendizaje automático de clasificación: Regresión Logística Multinomial, Árboles de clasificación, Random Forest, XGboost, LightGBM, Perceptrón multicapa, Redes neuronales recurrentes LSTM y GRU.

Los indicadores que se han utilizado para evaluar la capacidad predictiva de los modelos son:

- Accuracy: Es la proporción de predicciones correctas (tanto positivas como negativas) sobre el total de predicciones.
- Precisión: Mide la proporción de verdaderos positivos sobre el total de predicciones positivas.
- F1-Score: Es la media armónica entre precisión y sensibilidad (recall). Es especialmente útil cuando las clases están desbalanceadas, ya que considera tanto falsos positivos como falsos negativos.

3.5.7. *Siguientes fases*

Las etapas correspondientes a la implementación del modelo y a la validación de la solución de pronóstico no se desarrollaron en este informe, ya que suponen la puesta en producción de un modelo de negocio. Dado el enfoque de la investigación, el alcance de los resultados se limitó a la comparación del desempeño predictivo de los modelos analizados.

Capítulo IV

Resultados y Discusión

4.1. *Análisis Descriptivo de los Precios y Rendimientos*

Los precios diarios de la acción seleccionada presentan una tendencia volátil alrededor de una media. Por ciertos periodos presenta tendencia creciente y en otros tramos una tendencia decreciente. A inicios del 2020 se tuvo un nivel más alto de los precios, pero luego se evidenció una caída por la pandemia y posteriormente se ha mantenido en un dicho nivel, explicado por la coyuntura económica y política que se ha tenido entre el 2020 y 2023 (ver Figura 21).

Los rendimientos diarios de los precios de la acción seleccionada presentan unos valores atípicos en determinados periodos, lo que permite concluir presencia de heterocedasticidad en la serie de rendimientos. (ver Figura 22).

4.2. *Determinación de la Ventana de Pronóstico*

Se evaluó de manera iterativa la ventana de inversión óptima en cada día de la ventana temporal de modelamiento. Esta duración máxima diaria de inversión se evaluó en un rango de 1 a 30, cuantificando el porcentaje de casos donde cada rendimiento supera un umbral mínimo ($1\,\%$, $2\,\%$, $3\,\%$, $4\,\%$ o $5\,\%$). Luego, se observó que la curva de porcentajes deje de crecer, se estabilice o incluso se reduzca para determinar el día óptimo.

Precisar de que este análisis se realizó con el fin de sustentar analíticamente la determinación del día óptimo y no decidir arbitrariamente otra ventana de predicción.

En el análisis de los rendimientos, se determinó que en el día 20 el $35\,\%$ de los días alcanzaron un rendimiento superior al $3\,\%$. Después de ese día, este porcentaje se mantiene estable, lo que significa que no se observan grandes variaciones en la proporción de días con rendimiento superior al umbral a medida que se termina de recorrer los 30 días futuros a cada punto observado. En el caso de los otros umbrales,

el análisis es similar (ver Figura 23).

Figura 21

Evolutivo de los Precios de la Acción

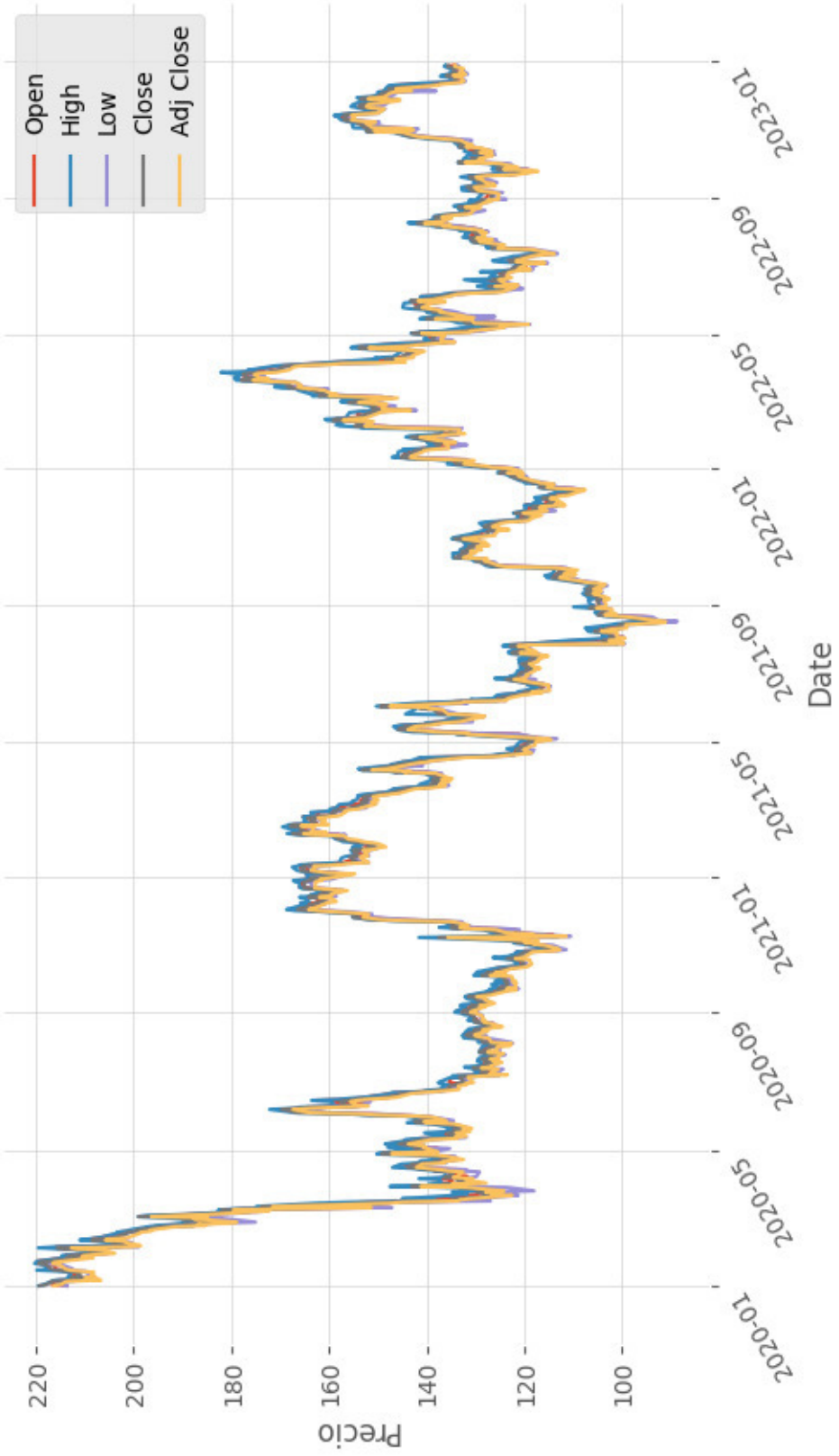


Figura 22

Evolutivo de los Rendimiento de los Precios de la Acción

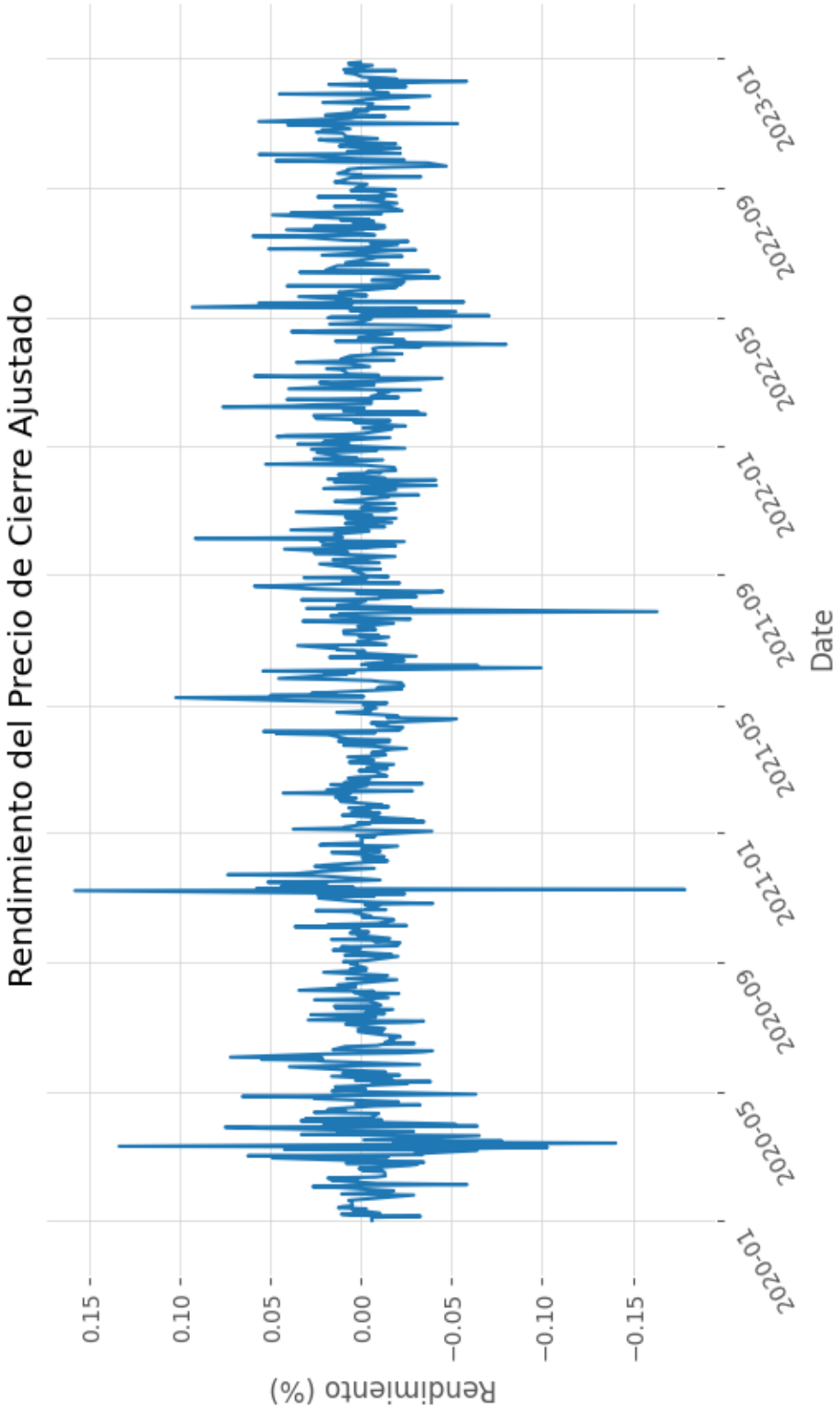
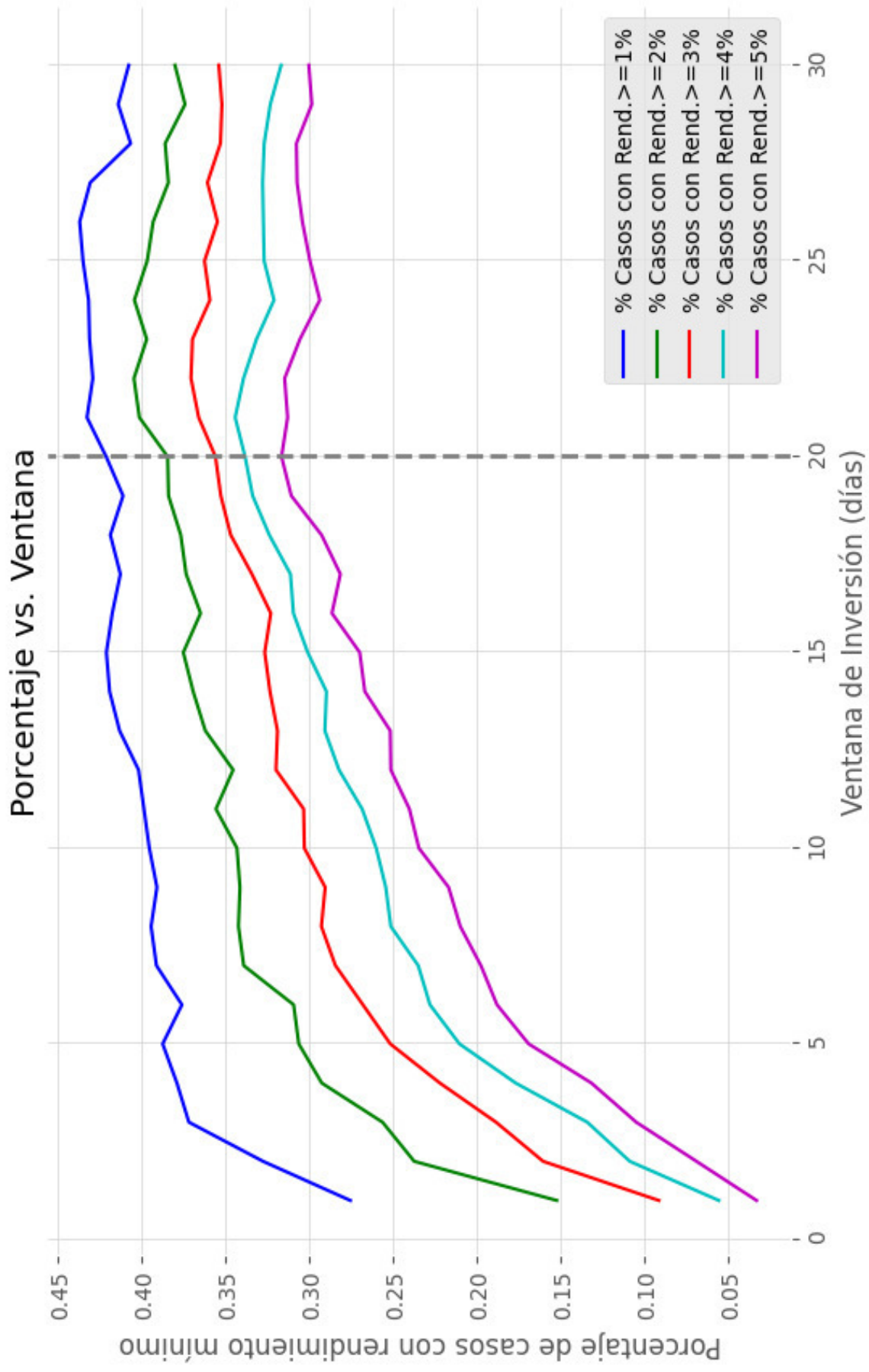


Figura 23

Determinación de la Ventana de Predicción



Una vez que se determinó la ventana de predicción de 20 días se procedió con la experimentación de aplicar diversas técnicas de series de tiempos y aprendizaje automático en los tres problemas planteados: (i) problema de pronósticos de precios (experimento 1), (ii) problema de pronósticos de rendimientos (experimento 2) y (iii) problema de clasificación de días con alto rendimiento (experimento 3).

4.3. Análisis de Resultados del Experimento 1

En este experimento se determinó la técnica de modelamiento que mejor pronostica los precios de las acciones en el día 20 de la ventana de predicción de la empresa peruana listada en la NYSE.

4.3.1. *Análisis Descriptivo de la Variable Dependiente Precio*

En la Figura 21, se observa que el precio se redujo desde inicios de la ventana de análisis seleccionada y el precio no ha podido recuperarse en los siguientes años.

4.3.2. *Comparación de las Técnicas de Pronóstico de los Precios*

En la Tabla 10 se determinó que la técnica de aprendizaje profundo de redes neuronales con neuronas Long Short Term Memory (LSTM) es la que alcanzó los mejores niveles de precisión en el pronóstico de precios al día 20 en MAE con un nivel del 7.968, en RMSE con 10.114 y MAPE con 0.058. Los otros dos modelos son la técnica de aprendizaje automático LightGBM y la red neuronal con neuronas GRU. Por el contrario, Random Forest presentó el mayor error de pronóstico, seguido por la técnica de serie de tiempo ARIMA.

Vale precisar, que por el lado de la técnica de serie de tiempo, se evaluó con varias combinaciones de los parámetro de ARIMA a través del proceso automatizado auto-arima que se puede ejecutar en *Python*, es decir, se evaluaron procesos autoregresivos, procesos de media móvil y procesos integrados (ver Anexo D).

Tabla 10
Precisión de las Técnicas de Pronóstico del Precio

Modelo	Técnica	MAE	RMSE	MAPE
Series de Tiempo	ARIMA	10.600	12.910	0.079
Aprendizaje Automático	Regresión Lineal	9.350	11.242	0.070
	Árbol de Regresión	9.699	10.992	0.073
	Rando Forest	10.658	13.620	0.079
	XGBoost	9.164	11.489	0.068
	LightGBM	*8.341*	*10.045*	*0.063*
Aprendizaje Profundo	Perceptrón Multicapa	9.399	11.390	0.071
	LSTM	***7.968***	***10.114***	***0.058***
	GRU	8.353	10.130	0.062

En la evaluación de estabilidad mediante *cross validation* se determinó que la técnica de aprendizaje automático LightGBM presentó un RMSE promedio menor a comparación de las otras técnicas. Seguido por las técnicas de aprendizaje profundo GRU y LSTM (ver Tabla 11).

Tabla 11
Estabilidad del RMSE en las Técnicas de Pronóstico de los Precios

Modelo	Técnica	Promedio	Mediana	Mínimo	Máximo	Desviación
Aprendizaje Automático	Árbol de Regresión	13.995	13.637	10.992	19.392	1.439
	Rando Forest	14.229	14.213	13.619	14.663	0.210
	XGBoost	11.489	11.489	11.489	11.489	0.000
	LightGBM	***10.045***	***10.045***	***10.045***	***10.045***	***0.000***
Aprendizaje Profundo	Perceptrón Multicapa	14.671	14.718	11.390	17.160	1.203
	LSTM	11.406	10.643	10.130	15.272	1.431
	GRU	*10.566*	*10.213*	*10.114*	*12.284*	*0.635*

Por tanto, de acuerdo con los resultados obtenidos en este experimento 1, se ha evidenciado la mejora en los pronósticos con el uso de las técnicas de aprendizaje automático, principalmentes las redes neuronales recurrentes con neuronas LSTM y GRU y también la técnica LightGBM que, mediante *cross validation*, alcanzó mejores resultados en promedio.

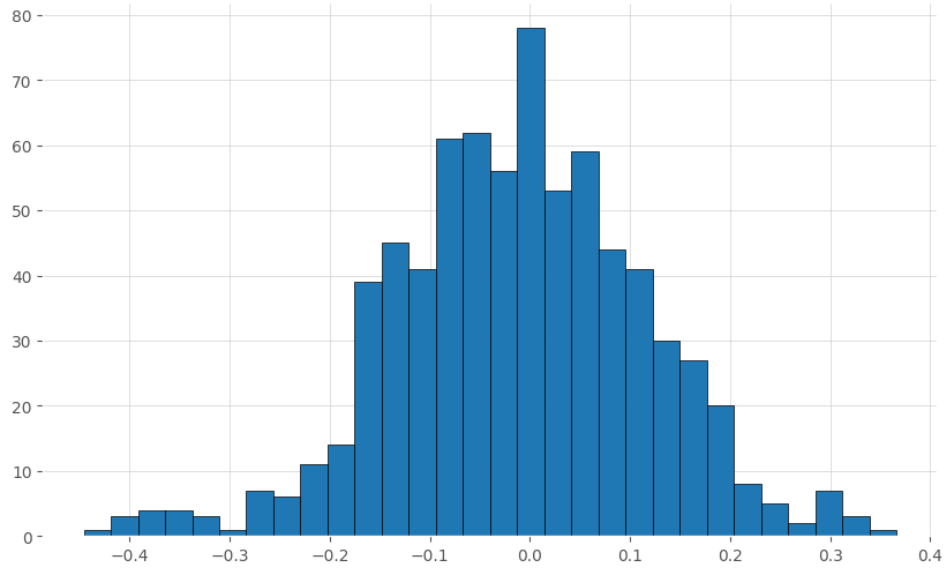
4.4. Análisis de Resultados del Experimento 2

En este experimento se determinó la técnica de modelamiento que mejor pronostica los rendimientos en el día 20 de la ventana de predicción de las acciones de la empresa peruana listada en la NYSE.

4.4.1. *Análisis Descriptivo de la Variable Dependiente Rendimiento*

En la Figura 24 se tiene el histograma de los rendimientos al día 20 donde se observa una concentración en los valores centrales, pero también se tienen valores atípicos por ambos extremos.

Figura 24
Histograma de los Rendimientos de la Acción



4.4.2. *Comparación de las Técnicas de Pronóstico de los Rendimientos*

En la Tabla 12 se determinó que la técnica de aprendizaje profundo de redes neuronales con neuronas *Gated Recurrent Uni* (GRU) es la que alcanzó los mejores niveles de precisión en el pronóstico de rendimientos al día 20 en MAE con un nivel del 0.056, en RMSE con 0.069 y MAPE con 2.165. Los otros dos modelos son la técnica de aprendizaje profundo LSTM y Perceptrón Multicapa. Por el contrario, la técnica de series de tiempo ARIMA presentó el mayor error de pronóstico.

Tabla 12
Precisión de las Técnicas de Pronóstico del Rendimiento

Modelo	Técnica	MAE	RMSE	MAPE
Series de Tiempo	ARIMA	0.129	0.152	3.904
Aprendizaje Automático	Regresión Lineal	0.072	0.086	1.314
	Árbol de Regresión	*0.068*	*0.084*	*1.255*
	Rando Forest	0.070	0.091	1.495
	XGBoost	*0.073*	*0.087*	*1.388*
	LightGBM	0.085	0.106	2.441
Aprendizaje Profundo	Perceptrón Multicapa	0.066	0.081	1.141
	LSTM	0.065	0.078	1.336
	GRU	***0.056***	***0.069***	***2.165***

En la evaluación de estabilidad mediante cross validation se determinó que la técnica de aprendizaje profundo GRU y Perceptrón Multicapa presentó un RMSE promedio menor a comparación de las otras técnicas. Seguido por las técnicas de aprendizaje automático XGBoost (ver Tabla 13).

Tabla 13
Estabilidad del RMSE en las Técnicas de Pronóstico de los Rendimientos

Modelo	Técnica	Promedio	Mediana	Mínimo	Máximo	Desviación
Aprendizaje Automático	Árbol de Regresión	0.098	0.097	0.084	0.119	0.006
	Rando Forest	0.093	0.093	0.091	0.095	0.001
	XGBoost	*0.087*	*0.087*	*0.087*	***0.087***	***0.000***
	LightGBM	0.106	0.106	0.106	0.106	***0.000***
Aprendizaje Profundo	Perceptrón Multicapa	0.082	0.082	0.081	*0.084*	0.001
	LSTM	0.095	0.091	0.078	0.171	0.016
	GRU	***0.082***	***0.081***	**0.069**	*0.121*	*0.010*

Por tanto, de acuerdo con los resultados obtenidos en este experimento 2, se ha evidenciado la mejora en los pronósticos de los rendimientos con el uso de las técnicas de aprendizaje automático, principalmentes las redes neuronales recurrentes con neuronas GRU y LSTM y también la técnica XGBoost.

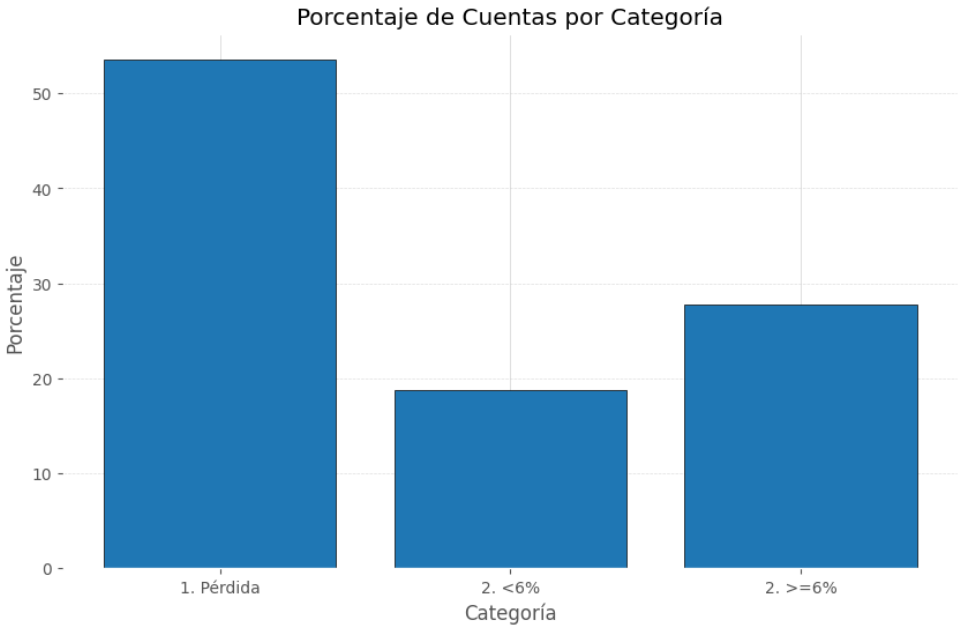
4.5. Análisis de Resultados del Experimento 3

En este experimento se determinó la técnica de modelamiento que mejor clasifica los rendimientos superiores al $6\,\%$ (rentabilidad esperada), entre $0\,\%$ (rentabilidad regular) y menores que $0\,\%$ (pérdida).

4.5.1. *Análisis Descriptivo de la Variable Dependiente*

En la Figura 25 se tiene el diagrama de barras de las categorías de rendimientos al día 20 (Pérdida, rendimiento menor que $6\,\%$ y rendimiento superior al $6\,\%$) donde se observa una concentración en la categoría de pérdida con un porcentaje mayor al $50\,\%$, mientras que la categoría de rentabilidad esperada (superior al $6\,\%$) contiene un porcentaje cercano al $30\,\%$. De todas maneras se observa materialidad de al menos el $15\,\%$ en las tres categorías, sin estar completamente balanceadas.

Figura 25
Clasificación de los Días con Mayor Rendimiento



4.5.2. *Comparación de las Técnicas de Clasificación*

En la Tabla 14 se determinó que la técnica de aprendizaje automático XGBoost alcanzó el mayor nivel de ajuste con un *accuracy* de 87 %, seguido por las redes neuronales recurrentes con neuronas *Long Short Term Memory* (GRU).

Tabla 14
Ajuste de las Técnicas de Clasificación de los Rendimientos

Modelo	Técnica	Accuracy	Precisión '2'	Recall '2'	f1-Score '2'
	Regresión Multinomial	0.507	0.195	0.571	0.291
	Árbol de Clasificación	*0.590*	*0.415*	*0.531*	*0.466*
Aprendizaje Automático	Rando Forest	0.576	0.220	0.474	0.300
	XGBoost	***0.869***	***0.926***	***0.899***	***0.912***
	LightGBM	0.569	0.268	0.786	0.400
	Perceptrón Multicapa	0.479	0.479	0.479	0.479
Aprendizaje Profundo	LSTM	*0.734*	*0.897*	*0.700*	*0.787*
	GRU	0.677	0.769	0.588	0.667

En la evaluación de estabilidad mediante *cross validation* se determinó que la técnica de aprendizaje profundo redes neuronales recurrentes con neuronas LSTM y GRU, y la técnica LightGBM presentaron un mejor ajuste (en promedio) a comparación de las otras técnicas (ver Tabla 15).

Tabla 15
Estabilidad del ACCURACY en las Técnicas de Clasificación

Modelo	Técnica	Promedio	Mediana	Mínimo	Máximo	Desviación
Aprendizaje Automático	Árbol de Clasificación	0.439	0.451	0.188	*0.590*	0.074
	Rando Forest	0.540	0.542	0.493	0.576	0.015
	XGBoost	0.542	0.542	0.542	0.542	***0.000***
	LightGBM	*0.569*	*0.569*	***0.569***	0.569	***0.000***
Aprendizaje Profundo	Perceptrón Multicapa	0.419	0.417	0.361	0.479	*0.026*
	LSTM	***0.624***	***0.641***	*0.460*	***0.734***	0.070
	GRU	0.597	0.605	*0.460*	0.677	0.045

Por tanto, de acuerdo con los resultados obtenidos en este experimento 3, se ha evidenciado la mejora en la clasificación de los rendimientos con el uso de las técnicas de aprendizaje automático, principalmente la técnica basada en árboles XGBoost, seguido por por la red neuronal recurrente con neuronas LSTM que, resultó mejor en el *cross validation*.

4.6. Pruebas de Hipótesis

Finalmente, de acuerdo con los experimentos realizados, se realiza un resumen de las pruebas de hipótesis de investigación en la Tabla 16.

Tabla 16
Verificación de la Hipótesis

Hipótesis	Verificación
HG. Las técnicas de Aprendizaje Automático, como las redes neuronales recurrentes LSTM, GRU y XGBoost, proporcionan una mayor capacidad de pronóstico de los precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.	Cumple
HE.1. La técnica de Aprendizaje Automático, red neuronal LSTM, proporciona una mayor precisión en los pronósticos de los precios de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.	Cumple
HE.2. La técnica de Aprendizaje Automático, red neuronal GRU, proporciona una mayor precisión en los pronósticos de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.	Cumple

HE.3. La técnica de Aprendizaje Automático, XGBoost, proporciona una mayor exactitud en la clasificación de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.	Cumple

4.7. Discusión de Resultados

Los experimentos realizados respaldan el uso de las técnicas de aprendizaje automático en el pronóstico de precios y rendimientos de las acciones de una empresa peruana listada en la Bolsa de Valores de Nueva York, en línea con los tres problemas de predicción planteados. Los resultados de este estudio evidencian que los modelos de aprendizaje automático, en particular los de aprendizaje profundo, como las redes neuronales recurrentes con neuronas LSTM o GRU, así como técnicas como XGBoost y LightGBM, superaron a los modelos tradicionales de series de tiempo, como ARIMA, y a otras técnicas de pronóstico tradicionales como las de regresión.

La superioridad de las técnicas de redes neuronales recurrentes con neuronas LSTM y GRU pueden interpretarse a partir de su capacidad para identificar patrones no lineales y dependencias de largo plazo en la serie de tiempo (Dudek et al., 2022). Estos patrones no lineales son comunes en los datos financieros debido a factores como la volatilidad y las reacciones del mercado a eventos exógenos (Bartiromo, 2004). Las estructuras de la red LSTM y GRU que cuentan con células de memoria y permite retener información durante múltiples pasos temporales, parece adaptarse mejor a estas características complejas de los datos bursátiles. En contraste, los modelos ARIMA están limitados a relaciones lineales y presentan mayores dificultades para adaptarse a cambios repentinos en los precios.

Los resultados obtenidos en este estudio son coherentes con investigaciones previas que destacan el desempeño superior de los modelos de aprendizaje profundo en datos financieros volátiles. Zhang et al. (2020) y Patel et al. (2021) también encontraron que los modelos de aprendizaje automático superan a los modelos tradicionales de series de tiempo en la predicción de precios de acciones. Además, estudios recientes en mercados emergentes han señalado que técnicas como LSTM son particularmente útiles cuando los datos exhiben alta volatilidad y variabilidad no lineal, lo cual refuerza la aplicabilidad de este enfoque para el mercado bursátil peruano en un contexto internacional.

Los hallazgos de este estudio tienen importantes implicaciones tanto para la teoría como para la práctica de la predicción financiera. Teóricamente, este estudio contribuye al campo de las finanzas cuantitativas al demostrar que los modelos LSTM pueden

mejorar la precisión en la predicción de precios en mercados de alta volatilidad. En términos prácticos, estos hallazgos sugieren que los analistas financieros e inversionistas podrían beneficiarse de la adopción de modelos de aprendizaje automático avanzados para mejorar sus estrategias de inversión y gestión de riesgo, especialmente cuando operan en mercados emergentes, donde las fluctuaciones pueden ser más pronunciadas.

Los resultados obtenidos fueron los esperados y están alineados con las hipótesis de investigación, sin embargo, existe la oportunidad de realizar posteriores investigaciones que exploren el impacto de incorporar variables exógenas tales como indicadores macroeconómicos del Perú, indicadores microeconómicos del sector económico de la empresa, indicadores financieros, entre otros. Además, el análisis de sentimiento de las noticias financieras o de las redes sociales u otras páginas *web* también pueden representar enfoques alternativos para mejorar la precisión de los pronósticos sobre los precios y los rendimientos. Por otro lado, sería interesante investigar el rendimiento de otros modelos de aprendizaje profundo, como las redes neuronales convolucionales (CNN) y *transformers* los cuales han mostrado resultados prometedores en la predicción de series de tiempo no lineales en otros campos (Zeng et al., 2023). También se sugiere la implementación de una estrategia de validación cruzada ampliada para probar los modelos en diferentes periodos de tiempo, aumentando la robustez de los resultados obtenidos.

Capítulo V

Conclusiones y Recomendaciones

5.1. Conclusiones

Los precios de las acciones de una empresa peruana listada en la Bolsa de Valores de Nueva York muestran variabilidad y ciertos patrones de comportamiento dentro de intervalos específicos de la ventana temporal seleccionada. Esta variabilidad en los precios influye directamente en la incertidumbre que enfrentan los inversionistas al tomar decisiones, lo que resalta la importancia de este trabajo de investigación. En particular, se evaluó la capacidad predictiva de las técnicas de aprendizaje automático para el pronóstico de los precios y rendimientos futuros, centrándose en tres problemas específicos: el pronóstico de los precios y de los rendimientos y la clasificación de los rendimientos.

En relación al objetivo general de la investigación, comparación de las técnicas de Series de Tiempo y Aprendizaje Automático en cuanto a su capacidad de pronóstico de los precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York, se concluyó que las técnicas de aprendizaje automático redes neuronales recurrentes con neuronas LSTM y GRU, y también las técnicas *boosting* XGBoost y LightGBM mejoraron la capacidad de pronóstico de los precios y de los rendimientos considerando los tres problemas específicos abordados. Estos resultados son evidencia de la importancia de incorporar técnicas de aprendizaje automático que permiten la correlación de múltiples variables independientes con la variable objetivo y la posibilidad de considerar relaciones no lineales, lo cual contribuye a una mejora significativa en la calidad de los pronósticos obtenidos.

En relación al primer objetivo específico, comparación de las técnicas de series de tiempo y aprendizaje automático en cuanto a su precisión en el pronóstico de los precios de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York, los resultados obtenidos evidenciaron la superioridad de la técnica de redes neuronales recurrentes con neuronas LSTM, cuyo mejor RMSE fue de 10.114. Precisar que

este resultado implica la mejor precisión después de 30 iteraciones, sin embargo, en niveles promedio al aplicar *cross validation*, se determinó que la técnica LightGBM obtuvo ligeramente un menor RMSE, seguido por la red neuronal recurrente con neuronas GRU, evidenciando la superioridad de este tipo de redes neuronales. Esto sugiere que las técnicas de aprendizaje automático tienen una mayor capacidad para capturar patrones no lineales y complejos en los datos históricos de los precios.

En relación al segundo objetivo específico, comparación de las técnicas de Series de Tiempo y Aprendizaje Automático en cuanto a su precisión en el pronóstico de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York, los resultados obtenidos evidenciaron la superioridad de la técnica de redes neuronales recurrentes con neuronas GRU, cuyo mejor RMSE fue de 0.069. Este resultado implica la mejor precisión después de 30 iteraciones y en niveles promedio al aplicar *cross validation* esta técnica permaneció como la mejor con el menor RMSE seguido por las redes neuronales de perceptrón multicapa, evidenciando la superioridad de este tipo de redes neuronales. Esto sugiere que las técnicas de aprendizaje automático presentaron una ventaja significativa en la predicción de cambios abruptos y tendencias en los rendimientos; lograron adaptarse mejor a la naturaleza volátil de los datos, lo que indica su superioridad en situaciones de alta variabilidad.

En relación al tercer objetivo específico, comparación de las técnicas de Aprendizaje Automático en cuanto a la exactitud en la clasificación de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York, los resultados obtenidos evidenciaron la superioridad de la técnica de XGBoost cuyo mejor grado de ajuste fue del $87\,\%$. Este resultado implica el mejor ajuste después de 30 iteraciones, sin embargo, en niveles promedio al aplicar *cross validation* las técnicas de redes neuronales recurrentes con neuronas LSTM y GRU obtuvieron los menores RMSE, evidenciando la superioridad de este tipo de redes neuronales. Esto se atribuye a la capacidad de los modelos de aprendizaje automático para integrar múltiples fuentes de datos y métricas derivadas que ayudan a identificar periodos de alto rendimiento con mayor exactitud.

En resumen, la investigación concluye que la aplicación de técnicas de aprendizaje automático ofrece ventajas significativas en la predicción de los precios y de los rendimientos de las acciones, así como en la clasificación de los altos rendimientos. Esto resalta la importancia de adoptar enfoques modernos y flexibles en el análisis financiero, particularmente en un entorno competitivo como el mercado de valores.

5.2. Recomendaciones

Las recomendaciones que se han determinado respecto al trabajo de investigación se enfocan en diversas perspectivas.

Modelos Alternativos. Dado que se ha evidenciado la utilidad de las técnicas de aprendizaje automático en el pronóstico de los precios de acciones, se recomienda explorar modelos alternativos que permitan mejorar la precisión de las predicciones teniendo como base las mejores técnicas que se encontraron en los experimentos (red neuronal recurrente con LSTM o GRU y XGboost o LightGBM). En particular, el uso de modelos híbridos y ensamblados podría potenciar los resultados al combinar la capacidad de los modelos de series de tiempo tradicionales, como ARIMA y GARCH, con la flexibilidad de los algoritmos de aprendizaje profundo como LSTM, GRU y *Transformers*. Asimismo, para la predicción de rendimientos, es fundamental profundizar en modelos de volatilidad, incluyendo variaciones multivariadas de GARCH, como DCC-GARCH y GJR-GARCH, que capturan mejor la dinámica del riesgo en los mercados financieros. En este sentido, se recomienda realizar una comparación exhaustiva entre estos enfoques para determinar cuál ofrece el mejor desempeño en términos de precisión predictiva y estabilidad, considerando métricas como RMSE y MAPE en diferentes horizontes temporales.

Análisis de Variables Exógenas. Para mejorar la precisión en los pronósticos de los precios de las acciones, se recomienda incorporar y analizar el impacto de variables exógenas, tanto macroeconómicas como específicas de la empresa. Factores como tasas de interés, inflación, PIB y tipo de cambio pueden influir significativamente en el comportamiento de los mercados financieros, afectando la dinámica de los precios. Asimismo, variables específicas de cada empresa, como informes de ganancias, anuncios de nuevos productos o cambios en la administración, pueden generar variaciones abruptas en el precio de sus acciones. Se sugiere la integración de estos factores en los modelos predictivos mediante enfoques como modelos de series de tiempo con regresores exógenos (ARIMAX, VARX, GARCH-X) o mediante técnicas de aprendizaje automático que incorporen estas variables como características adicionales en algoritmos como XGBoost, Random Forest o redes neuronales profundas. La inclusión y evaluación de estas variables permitirá identificar cuáles tienen un impacto significativo y optimizar la capacidad predictiva del modelo.

Análisis de Sentimiento. Para mejorar la precisión en la predicción de los precios de las acciones, se recomienda incorporar el análisis de sentimiento como una variable adicional en los modelos predictivos. La percepción del mercado, reflejada en redes sociales, noticias económicas y reportes financieros, puede influir significativamente en la evolución de los precios de los activos. Técnicas de procesamiento de lenguaje natural (NLP), como el uso de transformers (BERT, FinBERT) o modelos de análisis de sentimiento basados en diccionarios financieros (Loughran-McDonald), pueden permitir cuantificar el impacto del sentimiento del mercado en el precio de las acciones. Se sugiere combinar estas métricas con modelos tradicionales de series de tiempo (ARIMAX, GARCH-X) o técnicas de aprendizaje automático (XGBoost, LSTM,

Transformers) para evaluar su efecto sobre la volatilidad y los retornos. La integración del análisis de sentimiento puede proporcionar una ventaja adicional al capturar reacciones del mercado que no son detectables a través de variables exclusivamente cuantitativas.

Fuentes de Información. El problema de predicción de precios no se limita únicamente a las acciones de la Bolsa de Valores de Nueva York (NYSE), sino que también es relevante en otros mercados financieros, como la Bolsa de Valores de Lima (BVL), donde se negocian acciones de empresas peruanas. Por ello, las técnicas propuestas pueden adaptarse y aplicarse en este contexto, considerando las particularidades del mercado local, como su menor liquidez y la posible influencia de factores macroeconómicos específicos de la economía peruana. Además, el antecedente de aplicación al bitcoin sugiere que estas metodologías pueden extenderse a otros activos financieros, como el mercado de divisas (Forex), donde la variabilidad en los tipos de cambio representa un desafío similar. En este sentido, se recomienda explorar la aplicabilidad de modelos de series de tiempo y aprendizaje automático en distintos mercados, comparando su desempeño en la predicción de precios de acciones, criptomonedas y divisas, lo que permitiría validar la robustez de las técnicas empleadas en diferentes entornos financieros.

Modelos de Negocio. Dado el impacto de las técnicas de pronóstico en la predicción de precios y rendimientos de acciones en una Bolsa de Valores, es viable diseñar e implementar un *software* o aplicación que consolide información financiera y brinde análisis predictivo en tiempo real. Este sistema podría integrar modelos de series de tiempo (ARIMA, GARCH, VAR) y de aprendizaje automático (XGBoost, LSTM, Transformers) para generar pronósticos de precios futuros con alta precisión. Además, incluiría indicadores de riesgo, volatilidad y sentimiento del mercado, permitiendo a los inversores tomar decisiones informadas. El software podría ofrecer funcionalidades como: visualización de datos financieros con gráficos interactivos, alertas y recomendaciones basadas en predicciones, simulación de estrategias de inversión considerando distintos escenarios. Esta herramienta podría ser de gran utilidad para inversionistas individuales e institucionales, optimizando su toma de decisiones en los mercados financieros (ver Anexo E).

Referencias Bibliográficas

Achyutha, P. N., Chaudhury, S., Bose, S. C., Kler, R., Surve, J., y Kaliyaperumal, K. (2022). User classification and stock market-based recommendation engine based on machine learning and twitter analysis. *Mathematical Problems in Engineering*, *2022*(1), 4644855.

Al Mamun, A., Sohel, M., Mohammad, N., Sunny, M. S. H., Dipta, D. R., y Hossain, E. (2020). A comprehensive review of the load forecasting techniques using single and hybrid predictive models. *IEEE Access*, *8*, 134911–134939.

Amat, O. (2010). *La bolsa: funcionamiento y técnicas para invertir*. Grupo Planeta (GBS).

Arias, F. (2010). *El proyecto de investigación científica*. Episteme.

Arjomandi-Nezhad, A., Ahmadi, A., Taheri, S., Fotuhi-Firuzabad, M., Moeini-Aghtaie, M., y Lehtonen, M. (2022). Pandemic-aware day-ahead demand forecasting using ensemble learning. *IEEE Access*, *10*, 7098–7106.

Bartiromo, R. (2004). Non linear behaviour of stock market volatility. *arXiv preprint cond-mat/0410768*.

Bentéjac, C., Csörgo, A., y Martínez-Muñoz, G. (2019). A comparative analysis of xgboost. *ArXiv abs*.

Berzal, F. (2018). *Redes neuronales & deep learning: Volumen i*. Independently published.

Bobadilla, J. (2020). *Machine learning y deep learning: usando python, scikit y keras*. Ra-Ma.

Box, G. E., Jenkins, G. M., Reinsel, G. C., y Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Britannica, T. (2016, Jan. 19,). *New york stock exchange.* https://www.britannica.com/topic/New-York-Stock-Exchange.

Calvo, D. (2018). *Aprendizaje profundo.* https://www.diegocalvo.es/red-neuronal-recurrente/.

Chen, T., y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. En *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, Z., Li, C., y Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., y Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chung, J., Gulcehre, C., Cho, K., y Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Da Silva, D. G., Geller, M. T. B., dos Santos Moura, M. S., y de Moura Meneses, A. A. (2022). Performance evaluation of lstm neural networks for consumption prediction. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, *2*, 100030.

Data Science Team. (2020). *Introducción al concepto de lstm.* https://datascience.eu/es/aprendizaje-automatico/comprension-de-las-redes-de-lstm/.

Dubey, S., Patil, N., Thakur, P., y Kanakia, H. (2022). Stock market prediction using yahoo finance and twitter polarity. En *2022 ieee 3rd global conference for advancement in technology (gcat)* (pp. 1–5).

Dudek, G., Smyl, S., y Pełka, P. (2022). Recurrent neural networks for forecasting time series with multiple seasonality: a comparative study. En *International conference on time series and forecasting* (pp. 171–185).

educatyba. (2024, mayo). *Las empresas peruanas que cotizan en la bolsa de nueva york.* `https://tyba.pe/blog/empresas-peruanas-que-cotizan-en-la-bolsa/`. Descargado de `https://tyba.pe/blog/empresas-peruanas-que-cotizan-en-la-bolsa/` (Accessed: 2024-10-08)

Ensafi, Y., Amin, S. H., Zhang, G., y Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning–a comparative analysis. *International Journal of Information Management Data Insights*, *2*(1), 100058.

Fama, E. F. (1970). Session topic: stock market price behavior. *The Journal of Finance*, *25*(2), 383–417.

Fernandes, R., y Antonelli, M. (2018). Machine learning: a practical approach on the statistical learning theory. *Cham: Springer International Publishing*.

Galvez, T. (2016, Aug. 9,). *¿cómo medir la precisión de los pronósticos?* https://es.linkedin.com/pulse/c

Gao, P., Zhang, R., y Yang, X. (2020). The application of stock index price prediction with neural network. *Mathematical and Computational Applications*, *25*(3), 53.

García, M. C., Jalal, A. M., Garzón, L. A., y López, J. M. (2013). Métodos para predecir índices bursátiles. *Ecos de Economía*, *17*(37), 51–82.

Hastie, T., Tibshirani, R., Friedman, J. H., y Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hochreiter, S. (1997). Long short-term memory. *Neural Computation MIT-Press*.

Howell, E. (2023, Ene. 23,). *How to correctly perform cross-validation for time series.* https://towardsdatascience.com/how-to-correctly-perform-cross-validation-for-time-series-b083b869e42c.

Hu, Z., Zhao, Y., y Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, *4*(1), 9.

Hurwitz, J., y Kirsch, D. (2018). Machine learning for dummies. *IBM Limited Edition*, *75*.

IBM. (2021, May. 26,). *Inteligencia artificial.* https://www.ibm.com/cloud/learn/boosting.

Ignacio, J. (2018, Nov. 18,). *Cómo usar redes neuronales (lstm) en la predicción de averías en las máquinas.* https://blog.gft.com/es/2018/11/06/como-usar-redes-neuronales-lstm-en-la-prediccion-de-averias-en-las-maquinas/.

Ioffe, S., y Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. En *International conference on machine learning* (pp. 448–456).

Jimenez, D. (2015). *Economipedia: Acción.* https://economipedia.com/definiciones/accion.html.

Kamalov, F. (2020). Forecasting significant stock price changes using neural networks. *Neural Computing and Applications*, *32*(23), 17655–17667.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Khan, M. Y., Qayoom, A., Nizami, M., Siddiqui, M. S., Wasi, S., y Syed, K.-U.-R. R. (2021, 09). Automated prediction of good dictionary examples (gdex): A comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques. *Complexity*. doi: 10.1155/2021/2553199

Khare, K., Darekar, O., Gupta, P., y Attar, V. (2017). Short term stock price prediction using deep learning. En *2017 2nd ieee international conference on recent trends in electronics, information & communication technology (rteict)* (pp. 482–486).

Kingaby, S. A. (2022). The stock market api. En *Data-driven alexa skills*. Berkeley, CA: Apress. Descargado de `https://doi.org/10.1007/978-1-4842-7449-1_16` doi: 10.1007/978-1-4842-7449-1_16

Korstanje, J. (2021). *Advanced forecasting with python*. Springer.

Kumar, G., Jain, S., y Singh, U. P. (2021). Stock market forecasting using computational intelligence: A survey. *Archives of Computational Methods in Engineering*, *28*(3), 1069–1101.

Lazzeri, F. (2020). *Machine learning for time series forecasting with python*. John Wiley & Sons.

LeCun, A., Y. Canziani. (2020). *Aprendizaje profundo.* https://atcold.github.io/pytorch-Deep-Learning/es/.

Liu, J., Chao, F., Lin, Y.-C., y Lin, C.-M. (2019). Stock prices prediction using deep learning models. *arXiv preprint arXiv:1909.12227*.

Mañas, A. (2019, Jun. 19,). *Notas sobre pronóstico del flujo de tráfico en la ciudad de madrid.* https://bookdown.org/amanas/traficomadrid/resumen.html.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., y Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. En *2019 ieee fifth international conference on big data computing service and applications (bigdataservice)* (pp. 205–208).

Montgomery, D., Peck, E. A., y Vining, G. G. (2006). Introducción al análisis de regresión lineal. *México: Limusa Wiley*.

Moukalled, M. I. (2019). *Automated stock price prediction using machine learning* (Tesis Doctoral no publicada).

Naik, N., y Mohan, B. R. (2019). Study of stock return predictions using recurrent neural networks with lstm. En *International conference on engineering applications of neural networks* (pp. 453–459).

Olabe, X. B. (1998). Redes neuronales artificiales y sus aplicaciones. *Publicaciones de la Escuela de Ingenieros*.

Peixeiro, M. (2022). *Time series forecasting in python*. Manning. Descargado de `https://books.google.com.pe/books?id=hqXczgEACAAJ`

Peng, C., Yin, Z., Wei, X., y Zhu, A. (2019). Stock price prediction based on recurrent neural network with long short-term memory units. En *2019 international conference on engineering, science, and industrial applications (icesi)* (pp. 1–5).

Peng, Z., Huang, Q., y Han, Y. (2019). Model research on forecast of second-hand house price in chengdu based on xgboost algorithm. En *2019 ieee 11th international conference on advanced infocomm technology (icait)* (pp. 168–172).

Qian, X.-Y., y Gao, S. (2017). Financial series prediction: Comparison between precision of time series models and machine learning methods. *arXiv preprint arXiv:1706.00948*, 1–9.

Quinto, B. (2020). *Next-generation machine learning with spark: Covers xgboost, lightgbm, spark nlp, distributed deep learning with keras, and more*. Springer.

Sachdeva, A., Jethwani, G., Manjunath, C., Balamurugan, M., y Krishna, A. V. (2019). An effective time series analysis for equity market prediction using deep learning model. En *2019 international conference on data science and communication (icondsc)* (pp. 1–5).

Sanchez, J. (2016, Oct. 4,). *Economipedia:análisis de regresión.* https://economipedia.com/definiciones/analisis-de-regresion.html.

Sanz, F. (2023). *Modelos de secuencia. todo lo que necesitas es atención.* https://www.themachinelearners.com/modelos-secuencia/.

scikit-learn. (s.f.). *sklearn.model_selection.timeseriessplit.* Descargado 2024-10-08, de `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html`

Series, O. T. (2019). *Load, wind and solar, prices in hourly resolution.*

Shah, D., Isah, H., y Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(2), 26.

Staszkiewicz, P., y Staszkiewicz, L. (2015). Chapter 1 - introduction to finance and financial markets. En P. Staszkiewicz y L. Staszkiewicz (Eds.), *Finance* (p. 1-17). San Diego: Academic Press. Descargado de `https://www.sciencedirect.com/science/article/pii/B9780128015841000019` doi: https://doi.org/10.1016/B978-0-12-801584-1.00001-9

STATISTA. (2021, Aug. 31,). *Largest stock exchange operators worldwide as of august 2021, by market capitalization of listed companies.* https://www.statista.com/statistics/270126/largest-stock-exchange-operators-by-market-capitalization-of-listed-companies/.

Sugiyama, M. (2015). *Introduction to statistical machine learning*. Morgan Kaufmann.

Tjung, L. (2010). *Forecasting stock prices*. Lulu. com.

Tsay, R. (2010). *Analysis of financial time series*. John Wiley & Sons.

Verma, Y. (2022, May. 22,). *How to improve time series forecasting accuracy with cross-validation?* https://analyticsindiamag.com/how-to-improve-time-series-forecasting-accuracy-with-cross-validation/.

Vermorel, J. (2013, Jun. 1,). *Precisión de pronóstico de la demanda.* https://www.lokad.com/es/definicion-precision-de-pronostico.

Zeng, Z., Kaur, R., Siddagangappa, S., Rahimi, S., Balch, T., y Veloso, M. (2023). Financial time series forecasting using cnn and transformer. *arXiv preprint arXiv:2304.04912*.

Zhang, L., y Zhan, C. (2017). Machine learning in rock facies classification: An application of xgboost. En *International geophysical conference, qingdao, china, 17-20 april 2017* (pp. 1371–1374).

Anexos

Anexo A. Hipótesis de Investigación

A.1. Hipótesis General

Las técnicas de Aprendizaje Automático, como las redes neuronales recurrentes LSTM, GRU y XGBoost, proporcionan una mayor capacidad de pronóstico de los precios y rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York, a comparación de las técnicas de series de tiempo.

A.2. Hipótesis Específicas

HE.1. La técnica de Aprendizaje Automático, red neuronal LSTM, proporciona una mayor precisión en los pronósticos de los precios de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.

HE.2. La técnica de Aprendizaje Automático, red neuronal GRU, proporciona una mayor precisión en los pronósticos de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.

HE.3. La técnica de Aprendizaje Automático, XGBoost, proporciona una mayor exactitud en la clasificación de los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de Nueva York.

Anexo B. Matriz de Consistencia

Figura 26

Matriz de Consistencia

Aplicación de las Técnicas de Series de Tiempo y *Machine Learning* para el Pronóstico del Precio de la Acción de Una Empresa Peruana que Cotiza en la Bolsa de Valores de *New York*

Problema	Objetivos	Hipótesis	Metodología
1. Problema General	1. Objetivo General	1. Hipótesis General	
¿Qué técnicas, entre Series de Tiempo y *Machine Learning,* mejoran el pronóstico de los precios de las acciones de una empresa peruana que cotiza en la Bolsa de Valores de *New York*?	Comparar las técnicas de series de tiempo y de *Machine Learning* en la mejora del pronóstico de los precios de las acciones de una empresa peruana que cotiza en la Bolsa de Valores de *New York*.	Las técnicas de *Machine Learning* de redes neuronales recurrentes LSTM y GRU presentan un mejor pronóstico de los precios de las acciones de una empresa peruana que cotiza en la Bolsa de Valores de *New York*, a comparación de las técnicas de series de tiempo.	
2. Problemas Específicos	2. Objetivos Específicos	2. Hipótesis Específicas	**1) Nivel de investigación:** Aplicada **2) Diseño de investigación:** Experimental **3) Método de investigación:** Razonamiento deductivo. **4) Unidad de análisis:** Precio de una acción peruana **5) Población:** Cotizaciones históricas de los precios de una acción peruana **6) Muestra:** Precios de acción entre 01/01/2020 y 01/01/2023 **7) Técnicas:** Análisis de documentos como base de datos. **8) Instrumentos:** Página web Yahoo Finance.
¿Qué técnica, entre Series de Tiempo y *Machine Learning*, proporciona una mayor precisión en los pronósticos de precios de las acciones de una empresa peruana que cotiza en la Bolsa de Valores de *New York*?	Evaluar el desempeño de las técnicas de series de tiempo y de *Machine Learning* para pronosticar los precios de las acciones de una empresa peruana en la Bolsa de Valores de *New York*.	La técnica de *Machine Learning* red neuronal GRU proporciona una mayor precisión en los pronósticos de precios de la acción de una empresa peruana que cotiza en la Bolsa de Valores de *New York*.	
¿Qué técnica, entre Series de Tiempo y *Machine Learning*, mejora la precisión en los pronósticos de los rendimientos de las acciones de una empresa peruana que cotiza en la Bolsa de Valores de *New York*?	Comparar las técnicas de series de tiempo y *Machine Learning* para pronosticar los rendimientos de las acciones de una empresa peruana en la Bolsa de Valores de New York.	La técnica de *Machine Learning* red neuronal GRU proporciona una mayor precisión en los pronósticos de los rendimientos de la acción de una empresa peruana que cotiza en la Bolsa de Valores de *New York*.	
¿Cuál de las técnicas de *Machine Learning* mejora la clasificación de los días futuros con altos rendimientos de la acción de una empresa peruana que cotiza en la Bolsa de Valores de *New York*?	Determinar qué técnica de *Machine Learning* permite una mejor clasificación de los días con altos rendimientos de las acciones de una empresa peruana que cotiza en la Bolsa de Valores de *New York*.	La técnica de *Machine Learning* red neuronal LSTM mejora la clasificación de los días futuros con altos rendimientos de la acción de una empresa peruana que cotiza en la Bolsa de Valores de *New York*.	

Anexo C. Introducción a las Redes Neuronales Artificiales

Las redes neuronales artificiales, inspiradas en las redes neuronales del cerebro humano, se enfocan en *aprender* los patrones de comportamiento de un objeto de estudio en función al input que ingresa el investigador, es decir, generaliza sus respuestas desde los casos previos revisados en el entrenamiento a los nuevos casos que tienden a parecerse (Olabe, 1998).

La neurona es la unidad fundamental del cerebro porque recibe y combina señales desde y hacia otras neuronas, análogamente, en las redes neuronales artificiales el elemento procesador es la unidad fundamental debido a que tiene un conjunto de entradas que son combinadas con una suma ponderada, la cual es modificada por una función de transferencia que, genera un valor de salida del elemento procesador (Olabe, 1998) (ver la Figura 27). Por tanto, una red neuronal representa un conjunto de elementos procesadores conectados, que pueden estar organizados en niveles o capas, por ejemplo, una red tradicional se estructura en:

- **Inputs:** se refiere a la capa de entrada que representa los datos que son input para la red neuronal (X_1, X_2, ... X_n en la Figura 27).

- **Weights:** se refiere a los pesos de cada input (W_1, W_2, ... W_n en la Figura 27) que, junto con el *bias*, permiten calcular una suma ponderada $Z = \sum_{i=1}^{n} X_i W_i + bias$

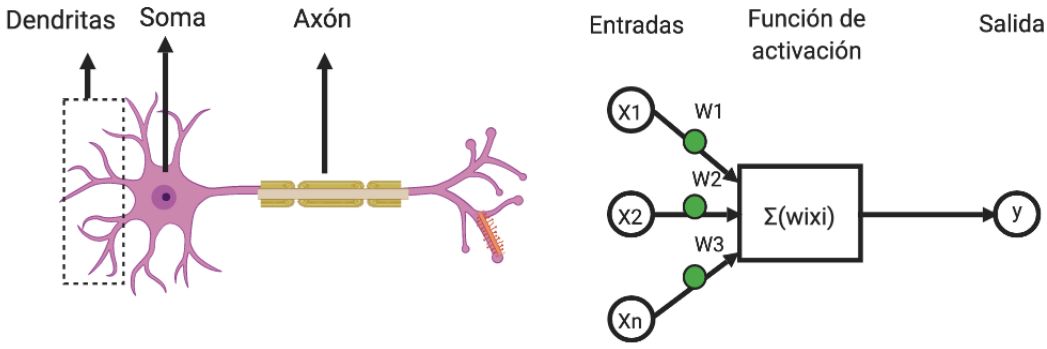
- **Función de activacción (σ):** esta se aplica a la suma ponderada calculada previamente. Un ejemplo de la función de activación es la función **sigmoid**.

$$\sigma(Z) = \frac{1}{1 + e^{-Z}}$$

- **Output:** es la capa de salida que representa el resultado de la red neuronal (y en la Figura 28).

Figura 27

Comparación de Neuronales del Cerebro y Neuronas Artificiales

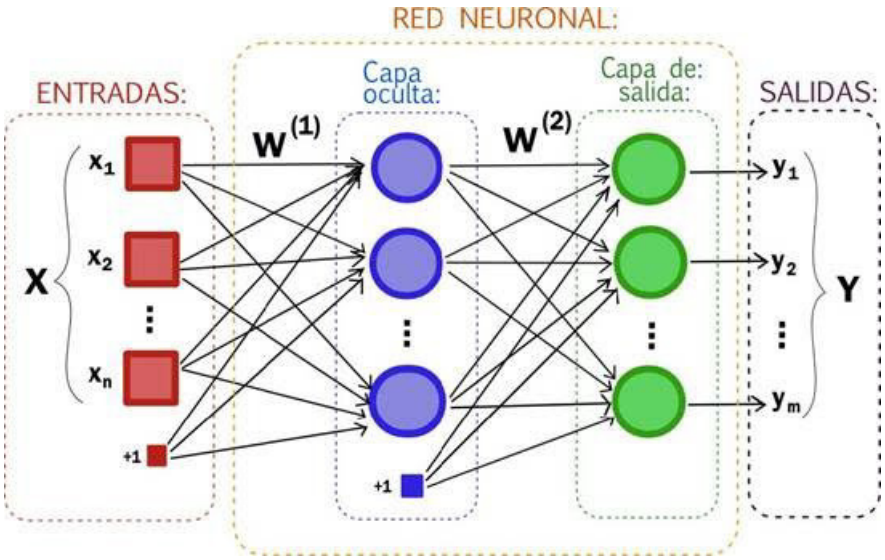


Nota. Tomado de "Introducción a las Redes Neuronales Pt. I", https://futurelab.mx/redes %20neuronales/inteligencia %20artificial/2019/06/25/intro-a-redes-neuronales-pt-1/

En la Figura 28 se observa una red neuronal con una capa oculta y múltiples salidas que representan las respuestas a cada unida de observación.

Figura 28

Estructura de Redes Neuronales Artificiales



Nota. Tomado de "Redes Neuronales", https://sites.google.com/site/mayinteligenciartificial/unidad-4-redes-neuronales

Entrenamiento de las Redes Neuronales. De acuerdo con Mañas (2019), un resumen en el entrenamiento de una red neuronal se da como sigue:

- Iniciar con unos pesos aleatorios.
- Introducir aleatoriamente los datos de entrada para iniciar con el entrenamiento.

- Dejar que la red neuronal genere un output, conocido como la propagación hacia delante.
- Comparar el output estimado con el output observado.
- Calcular el error o la diferencia entre la salida generada y la deseada (denominada error) se usa para ajustar los pesos sinápticos de las neuronas de la capa de salidas.
- El error se propaga hacia atrás y se ajustan los pesos en la capa anterior y se continúa propagando hacia atrás hasta que se alcance la capa de entrada
- Repetir el proceso con los diferentes datos de entrenamiento hasta reducir el error hasta un umbral mínimo aceptable.

A continuación, se da un mayor detalle del entrenamiento de la red neuronal.

- **Funciones de coste:** Permite cuantificar el grado de error del modelo de pronóstico entre el dato observado ($Y(x)$) y una función a que tiene como parámetros los pesos de cada variable (W) y el bias (b). Entonces, la función de coste es el siguiente:

$$C(W,b) = \frac{1}{2n}\sum_{x}[Y(x) - a(x,W,b)]^2 \tag{1}$$

  Siendo n el número de datos de la muestra.

  Si $C(W,b)$ es muy grande entonces la red no logra pronosticar correctamente, por lo que se buscará re-entrenar la red para minimizar el error hasta 0 si es posible ($C(W,b) = 0$). Por tanto, para minimizar la función de costo se deben estimar los parámetros W y b.

- **Entrenamiento con gradiente descendente:** Del punto anterior, se puede determinar que entrenar una red neuronal en realidad es un problema de optimización donde se busca minimizar la función de coste $C(W,b)$, la cual es compleja y dependerá de la arquitectura de la red y de sus parámetros W y b. Entonces, para buscar los valores óptimos de cada parámetro se emplea un algoritmo, conocido como gradiente descente.

  Se define la regla de actualización de la gradiente descendente, partiendo de un valor inicial W_0 y b_0:

$$W_k^{'} = W_k - \eta\frac{\partial C}{\partial W_k}$$

$$b_l^{'} = b_l - \eta \frac{\partial C}{\partial b_l}$$

Para cada valor de $W_k^{'}$ y $b_l^{'}$ se determina la función de coste es $C = \frac{1}{n} \sum_x C_x$ tal que $C_x = \frac{[Y(x)-a]^2}{2}$. Enseguida, se calcula el gradiente de los costos con respecto a cada punto x:

$$\triangle C = \frac{1}{n} \sum_x \triangle C_x$$

De la fórmula anterior se puede concluir que no es eficiente cuando se tiene que entrenar un número grande de datos, por eso se formuló la gradiente descendente estocástica que utiliza muestras aleatorias de los datos completos. Formalmente se tiene lo siguiente:

Se tienen m muestras de entrenamiento X_1, X_2, ..., X_m que se denominan batch o mini batch. La estimación es:

$$\triangle C = \frac{1}{m} \sum_{j=1}^{m} \triangle C_{X_j}$$

Por tanto, la regla de aprendizaje con gradiente descendente estocástica queda definido como:

$$W_k^{'} = W_k - \frac{\eta}{m} \sum_j \frac{\partial C}{\partial W_k}$$

$$b_l^{'} = b_l - \frac{\eta}{m} \sum_j \frac{\partial C}{\partial b_l}$$

Donde j representa la muestra de entrenamiento.

A pesar de que hay una mejora en el entrenamiento, aún no es suficiente porque la gradiente descendente estocástica se caracteriza por estimar un óptimo local. Sin embargo, es importante destacar que si se tiene un óptimo global en el entrenamiento, la red neuronal puede aprender tanto que sólo es válido para esa muestra pero no para muestras de validación lo cual también representa un problema de sobreajuste.

- **Algoritmo Backpropagation:** Toma los conceptos matemáticos de la regla de la cadena para el entrenamiento de las redes neuronales. El entrenamiento del modelo consta de varios pasos hacia adelante y hacia atrás. En el paso directo, los datos pasan a través de la red desde la capa de entrada a la capa de salida. En

el paso hacia atrás, los algoritmos calculan derivadas parciales de la función de costo con respecto a las ponderaciones y las utilizan para ajustar los valores de las ponderaciones (Olabe, 1998).

- **Unidades de activación:** Se considera como un elemento crítico en las redes neuronales.

Tabla 17
Unidades de activación

Unidad	Función	Definición
Sigmoid	$f(Z) = \frac{1}{1+exp(-Z)}$	Se caracteriza por convertir cualquier número real en un número entre 0 y 1, haciendo que valores positivos grandes sean prácticamente 1, mientras que valores negativos grandes tienden a 0.
Tanh	$f(Z) = \frac{e^Z - e^{-Z}}{e^Z + e^{-Z}}$	Presenta una forma bastante similar a sigmoid, saturando entre los valores -1 y 1.
ReLU	$f(Z) = max(0, Z)$	La unidad ReLU tiene salida 0 para valores menores que 0 y la función identidad para valores mayores que 0. Es la unidad de activación más usada en la práctica.
Leaky ReLU	Si $Z > 0$ entonces $f(Z) = Z$, caso contrario $f(Z) = 0.01Z$	Variante de la ReLU que intenta solucionar el problema de las ReLU muertas.
Maxout	$f(Z) = max(W_1^T X + b_1, W_2^T X + b_2)$	Generaliza las unidades ReLU y Leaky ReLU.

- **Batch normalization:** Consiste en normalizar los inputs de entrada de las unidades de activación, donde se calculan la media y la varianza empíricas para cada training batch.

$$\hat{X}^{(k)} = \frac{X^{(k)} - E[X^{(k)}]}{\sqrt{Var[X^{(k)}]}}$$

Según Ioffe y Szegedy (2015), Batch normalization se ejecuta antes de la función de activación, normalizando la operación lineal $WX + b$.

Algorithm 1: Batch Normalization

Input: Mini-batch $\mathcal{B} = \{x_1, \ldots, x_m\}$, parameters γ, β

Output: $\{y_i = \text{BN}_{\gamma,\beta}(x_i)\}$

Compute mini-batch mean:

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

Compute mini-batch variance:

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$$

Normalize each input:

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

Scale and shift:

$$y_i \leftarrow \gamma\hat{x}_i + \beta$$

- **Optimización avanzada para mejorar el overfitting:** Se considera como un elemento crítico en las redes neuronales.

Tabla 18
Variantes para mejorar Overfitting

Variantes	Definición
SGD + Momentum	Se tiene un vector de velocidad que es una especie de media ponderada de gradientes anteriores: $$V_{t+1} = pV_t + \triangle f(X_t)$$ $$X_{t+1} = X_t - \alpha V_{t+1}$$ La idea es que el descenso va adquiriendo «velocidad» en las direcciones correctas a través de todos las iteraciones, acelerando el proceso de convergencia.
Adagrad	Modifica la learning rate de manera independiente para cada parámetro.
RMSProp y Adam	Son métodos de optimización que mejoran la variante Adagrad.
Learning rate decay	Permite que el learning rate se vaya reduciendo a medida que avanza el entrenamiento. De este modo, el entrenamiento comienza con una learning rate mayor, haciendo que la loss se reduzca rápido, y poco a poco esta se va reduciendo permitiendo buscar mejor el espacio de soluciones.

- **Regularización:** Se enfocan en los problemas de overfitting.

Tabla 19
Alternativas de Regularización

Regularización	Definición
L2	Consiste en penalizar la norma al cuadrado de los weights en una red neuronal en la función de coste. Por ejemplo, usando cross entropy loss, la función de coste: $$C_0 = -\frac{1}{n}\sum_{X_j}[Y_j Ln(a_j^L) + (1 - Y_j)Ln(1 - a_j^L)]$$ $$C = C_0 + \frac{\lambda}{2n}\sum_W W^2$$
L1	Análogo a la L2, pero se suma los valores absolutos. $$C = C_0 + \frac{\lambda}{n}\sum_W \lvert W\rvert$$
Dropout	Consiste en asignar aleatoriamente una salida 0 a un porcentaje de neuronas de la red durante el entrenamiento con lo cual, en cada batch, se tendrá un número aleatorio de neuronas desactivadas. La probabilidad de desactivación (p) es un hiperparámetro de la red y normalmente se utilizan valores de 0.25 ó 0.5 donde se puede tener un 25 % o 50 % de neuronas desactivadas en cada batch.
Early stopping	Consiste en utilizar los datos de validación para evaluar el rendimiento de la red neuronal durante el entrenamiento sin aprender directamente sobre los datos. Por tanto, se puede determinar si la red está empezando a tener overfitting parar el entrenamiento con lo que se decide parar el entrenamiento en una época determinada con el fin de evitar mayor grado de sobreajuste.

Por otro lado, a diferencia de los algoritmos tradicionales que son instrucciones previamente programadas, las redes neuronales artificiales deben ser previamente entrenadas en base a unos casos de ejemplo y una regla de aprendizaje. A continuación, se abordan las técnicas de perceptrón multicapa, redes neuronales recurrentes y memoria de corto plazo a largo plazo.

Anexo D. Resultados de las Técnicas de Series de Tiempo

D.1. Pronóstico de los Precios

Figura 29

Resultados Obtenido de AUTOARIMA en el Pronóstico de Precios

```
Performing stepwise search to minimize aic
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=3177.374, Time=1.09 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=3184.908, Time=0.04 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=3180.659, Time=0.12 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=3182.056, Time=0.33 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=3183.511, Time=0.04 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=3173.014, Time=1.10 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=3171.536, Time=0.33 sec
 ARIMA(3,1,0)(0,0,0)[0] intercept   : AIC=3173.082, Time=0.49 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=3174.925, Time=2.38 sec
 ARIMA(2,1,0)(0,0,0)[0]             : AIC=3170.104, Time=0.12 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=3179.406, Time=0.05 sec
 ARIMA(3,1,0)(0,0,0)[0]             : AIC=3171.617, Time=0.24 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=3171.537, Time=0.65 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=3176.101, Time=0.12 sec
 ARIMA(3,1,1)(0,0,0)[0]             : AIC=3173.420, Time=0.30 sec

Best model:  ARIMA(2,1,0)(0,0,0)[0]
Total fit time: 7.475 seconds
```

Figura 30

Especificación de los Parámetros de ARIMA (2,1,0) para los Precios

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  572
Model:               SARIMAX(2, 1, 0)   Log Likelihood               -1582.052
Date:                Fri, 06 Sep 2024   AIC                           3170.104
Time:                        04:11:23   BIC                           3183.146
Sample:                             0   HQIC                          3175.192
                                  - 572
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0882      0.022     -4.068      0.000      -0.131      -0.046
ar.L2          0.1408      0.025      5.686      0.000       0.092       0.189
sigma2        14.9293      0.455     32.844      0.000      14.038      15.820
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):              1335.92
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               0.51   Skew:                            -0.57
Prob(H) (two-sided):                  0.00   Kurtosis:                        10.40
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

D.2. Pronóstico de los Rendimientos

Figura 31

Resultados Obtenido de AUTOARIMA en el Pronóstico de Rendimientos

```
Performing stepwise search to minimize aic
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=-2065.517, Time=0.22 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=-674.307, Time=0.12 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=-2063.176, Time=0.08 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=-1215.479, Time=0.44 sec
 ARIMA(2,0,1)(0,0,0)[0]             : AIC=-2072.249, Time=1.08 sec
 ARIMA(2,0,0)(0,0,0)[0]             : AIC=-2067.340, Time=0.43 sec
 ARIMA(3,0,1)(0,0,0)[0]             : AIC=-2092.254, Time=1.56 sec
 ARIMA(3,0,0)(0,0,0)[0]             : AIC=-2085.952, Time=0.77 sec
 ARIMA(4,0,1)(0,0,0)[0]             : AIC=-2085.820, Time=1.07 sec
 ARIMA(3,0,2)(0,0,0)[0]             : AIC=-2092.597, Time=2.26 sec
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=-2086.333, Time=5.46 sec
 ARIMA(4,0,2)(0,0,0)[0]             : AIC=inf, Time=3.91 sec
 ARIMA(3,0,3)(0,0,0)[0]             : AIC=-2082.086, Time=1.42 sec
 ARIMA(2,0,3)(0,0,0)[0]             : AIC=-2087.447, Time=2.15 sec
 ARIMA(4,0,3)(0,0,0)[0]             : AIC=-2076.378, Time=2.90 sec
 ARIMA(3,0,2)(0,0,0)[0] intercept   : AIC=-2086.867, Time=5.87 sec

Best model:  ARIMA(3,0,2)(0,0,0)[0]
Total fit time: 29.827 seconds
```

Figura 32

Especificación de los Parámetros de ARIMA (2,1,0) para los Rendimientos

SARIMAX Results

Dep. Variable:	y	No. Observations:	572
Model:	SARIMAX(3, 0, 2)	Log Likelihood	1052.299
Date:	Fri, 06 Sep 2024	AIC	-2092.597
Time:	04:25:21	BIC	-2066.502
Sample:	0	HQIC	-2082.417
	- 572		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.4474	0.159	9.113	0.000	1.136	1.759
ar.L2	-0.0556	0.298	-0.187	0.852	-0.639	0.528
ar.L3	-0.4074	0.145	-2.806	0.005	-0.692	-0.123
ma.L1	-0.6127	0.164	-3.748	0.000	-0.933	-0.292
ma.L2	-0.2007	0.161	-1.245	0.213	-0.517	0.115
sigma2	0.0015	5.41e-05	27.214	0.000	0.001	0.002

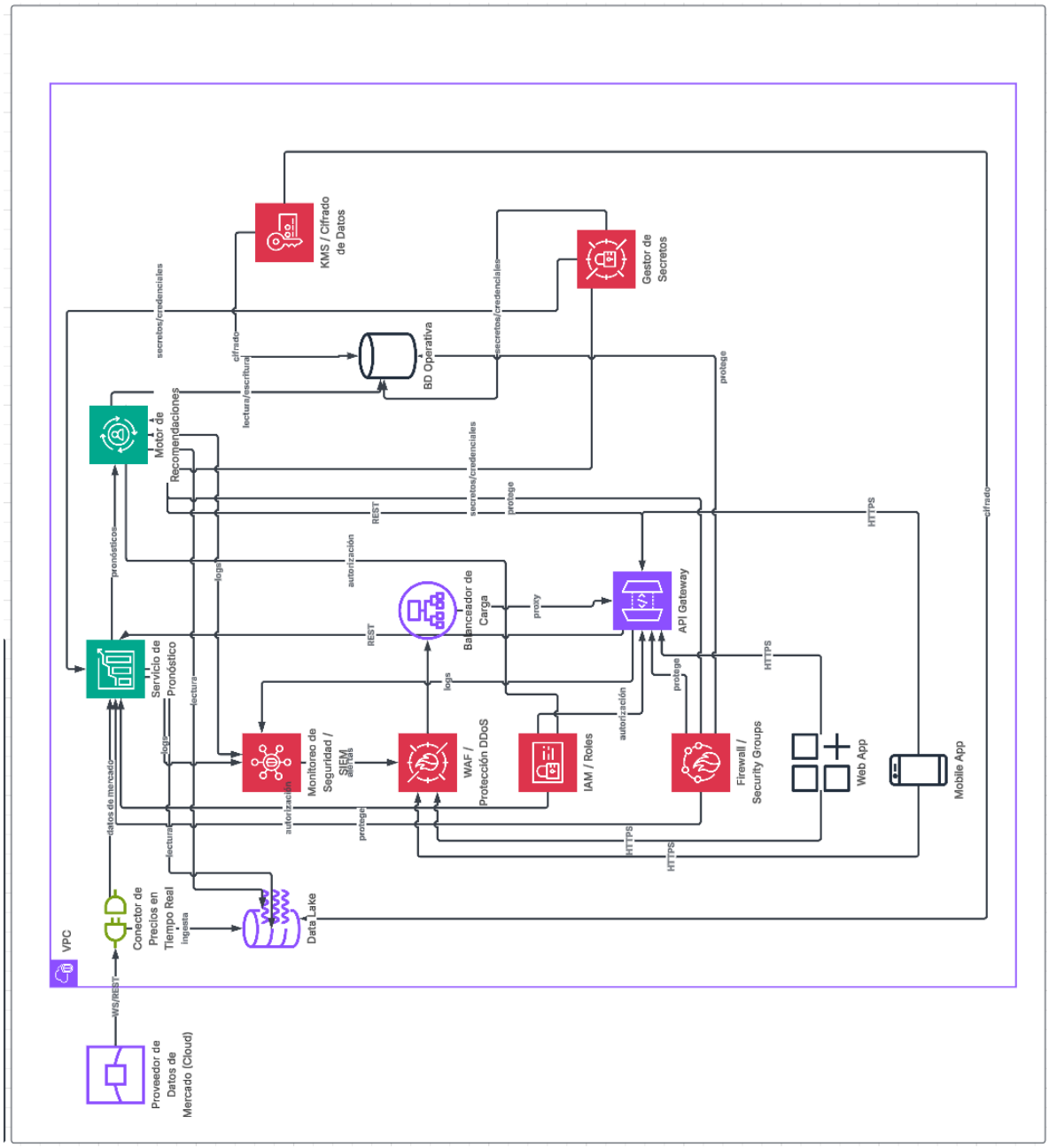
Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	448.35
Prob(Q):	0.85	Prob(JB):	0.00
Heteroskedasticity (H):	0.67	Skew:	-0.09
Prob(H) (two-sided):	0.01	Kurtosis:	7.33

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Anexo E. Arquitectura Tecnológica de Implementación

Figura 33

Arquitectura de Software de la Aplicación



Se ha propuesto una arquitectura por capas con ingesta de datos financieros, almacenamiento histórico, analítica predictiva, motor de decisión y exposición segura vía APIs, con controles transversales de seguridad, cifrado, monitoreo y gestión de credenciales

La Figura 33 representa una plataforma *cloud* dentro de una VPC para:

- Extraer datos online (precios en tiempo real / WS-REST desde un proveedor externo).

- Persistirlos (Data Lake para histórico y trazabilidad).
- Generar pronósticos (Servicio de Pronóstico).
- Convertir pronósticos en decisiones/señales (Motor de Recomendaciones).
- Servir resultados al usuario (Web/Mobile) a través de API Gateway + Balanceador, con controles de seguridad.
- Operar con seguridad y cumplimiento: WAF/DDoS, IAM, Firewall/Security Groups, SIEM/logs, KMS y Secrets Manager.

La arquitectura del sistema se organizó siguiendo un enfoque por capas, lo que permitió una clara separación de funciones, escalabilidad, mantenibilidad y refuerzo de los controles de seguridad. Cada capa cumple una función específica dentro del flujo de información, desde la interacción con el usuario hasta la persistencia y protección de los datos.

- **Capa de Presentación y Canales de Usuario:** Es el punto de interacción entre el sistema y los usuarios finales. Incluye las aplicaciones web y móviles donde se mostrará la información procesada, tales como precios históricos, pronósticos, recomendaciones de inversión y reportes analíticos.

  Su función principal es la visualización y comunicación de resultados, sin contener lógica de negocio ni procesamiento analítico. La comunicación con las capas inferiores se realiza a través de interfaces seguras, garantizando la integridad y confidencialidad de la información transmitida.

- **Capa de Exposición y Gestión de Servicios (API):** Actúa como un intermediario controlado entre los clientes y los servicios internos del sistema. A través de mecanismos como el balanceo de carga y la gestión de APIs, esta capa centraliza el acceso a las funcionalidades del sistema.

  Sus funciones consisten en la autenticación y autorización de solicitudes, el control de tráfico, la aplicación de políticas de seguridad, el versionamiento de servicios y el desacoplamiento entre los consumidores y la lógica interna. Esta capa permite que la plataforma escale de manera horizontal sin afectar la experiencia del usuario.

- **Capa de Servicios de Negocio y Analítica:** Concentra la lógica funcional central del sistema. En ella se implementan los módulos responsables de generar pronósticos financieros y transformar dichos resultados en recomendaciones o decisiones de inversión.

Los servicios de esta capa operan de forma desacoplada y se comunican mediante interfaces bien definidas, lo que permite la evolución independiente de los modelos analíticos, las reglas de decisión y los algoritmos de optimización. Asimismo, esta capa constituye el núcleo donde se integran técnicas de series de tiempo y aprendizaje automático.

- **Capa de Gestión y Persistencia de Datos:** La capa de datos se encarga del almacenamiento estructurado y no estructurado de la información utilizada por el sistema. Incluye repositorios diseñados para conservar grandes volúmenes de datos históricos de mercado, así como bases de datos operativas que mantienen el estado transaccional de la aplicación.

  Su diseño prioriza la trazabilidad, reproducibilidad y auditoría de los datos, aspectos fundamentales en aplicaciones financieras. Esta capa proporciona soporte tanto a los procesos analíticos como a la generación de reportes y al seguimiento de decisiones históricas.

- **Capa de Seguridad y Gobierno de Acceso:** La seguridad se implementa como una capa transversal que abarca todos los componentes del sistema. Su función es garantizar que el acceso a los recursos esté restringido de acuerdo con principios de mínimo privilegio y que la información se encuentre protegida tanto en tránsito como en reposo.

  Esta capa incorpora mecanismos de gestión de identidades, control de roles, protección perimetral, cifrado de datos y administración segura de credenciales. De esta manera, se asegura el cumplimiento de buenas prácticas de ciberseguridad y protección de información sensible.

- **Capa de Observabilidad, Monitoreo y Operación:** Finalmente, la capa de observabilidad proporciona los mecanismos necesarios para el seguimiento continuo del funcionamiento del sistema. A través de la recopilación y análisis de registros, métricas y eventos de seguridad, esta capa permite detectar anomalías, evaluar el desempeño y responder oportunamente a incidentes.

  Su presencia es fundamental para garantizar la estabilidad operativa, la mejora continua del sistema y la confiabilidad de los resultados generados por la plataforma.

A continuación se muestra el nivel de detalle las componentes de la arquitectura.

- Proveedor de Datos de Mercado (Cloud)
  - Rol: fuente externa (broker/market data vendor).
  - Interfaz: WS/REST.

    - Riesgos: límites de rate, latencia, caídas.
    - Controles: reintentos, validación de datos, circuit breaker.

- Conector de Precios en Tiempo Real (ingesta)
    - Rol: adaptador que transforma el feed externo al modelo interno.
    - Hace: autenticación contra el proveedor, normalización (timestamps, zonas horarias, símbolos), control de calidad básico (faltantes/duplicados), envío a almacenamiento (Data Lake)
    - Sale: "datos de mercado" hacia Data Lake
    - Observabilidad: emite logs de ingesta (volumen, lag, errores) a SIEM.
- Data Lake
    - Rol: repositorio de datos histórico para: backtesting, entrenamiento, auditoría y reproducibilidad.

      El data lake es más escalable para histórico masivo y series temporales.
- Servicio de Pronóstico
    - Rol: capa analítica (predicción).
    - Entradas: histórico desde Data Lake, features, indicadores y parámetros.
    - Salidas: pronósticos a motor de recomendaciones, logs a SIEM.
- Motor de Recomendaciones
    - Entradas: pronósticos, estado/estrategia del usuario (BD Operativa)
    - Procesos típicos: reglas (thresholds), filtros de riesgo, asignación (posición objetivo, exposición máxima), optimización de portafolio.
    - Salidas: señales y recomendaciones persistidas en BD Operativa.
- BD Operativa
    - Rol: base de datos transaccional para la app.
    - Guarda: usuarios y preferencias, historial de señales, portafolios simulados, auditoría.
    - Cifrado: aparece conectado a KMS/Secrets.
- API Gateway
    - Rol: "puerta de entrada" de la plataforma.

- Hace: expone endpoints que permiten acceder a las principales funcionalidades del sistema, incluyendo la generación de pronósticos, la obtención de recomendaciones y la consulta de reportes.
- Beneficio: desacopla frontend del backend y mejora seguridad.

- Balanceador de Carga
  - Rol: distribuye tráfico a instancias/servicios detrás (API Gateway o servicios).
  - Beneficio: alta disponibilidad y escalabilidad horizontal.
- WAF / Protección DDoS
  - Rol: seguridad perimetral.
  - Hace: bloquea patrones maliciosos (inyección, bots), mitiga ataques de denegación de servicio. Realiza el control preventivo de ciberseguridad para servicios expuestos.
- IAM / Roles
  - Rol: control de identidades y permisos.
  - Ejemplos de roles: usuario final (solo lectura de sus reportes/señales), admin (acceso a monitoreo/configuración), servicio pronóstico (lectura en Data Lake, escritura de resultados)
  - Beneficio: principio de menor privilegio.
- Firewall / Security Groups
  - Rol: segmentación de red dentro de la VPC.
  - Hace: permitir solo puertos necesarios (HTTPS, etc.), restringir acceso a BD solo desde servicios autorizados.
- Monitoreo de Seguridad / SIEM + Alertas
  - Rol: correlación de logs y alertas.
  - Recibe logs de: conector, pronóstico, recomendaciones, API Gateway/-WAF
  - Alerta por: picos anómalos de tráfico, errores de autenticación, caídas de ingesta, latencias elevadas
- Gestor de Secretos
  - Rol: almacenar credenciales y tokens (proveedor de mercado, BD, keys).

    - Evita: credenciales hardcodeadas.
    - Se integra con: servicios + KMS.

- KMS / Cifrado de Datos
    - Rol: cifrado y llaves administradas.
    - Aplica a: Data Lake (en reposo), BD Operativa (en reposo), secretos (cifrado)
    - En tránsito: HTTPS (ya aparece en tu diagrama).