



FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES
GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

PREDICCIÓN DE PRECIOS DE ACCIONES EN BOLSA A TRAVÉS DE REDES NEURONALES

Autor
César Traseira Linares

Tutora
M^a Lourdes Fernández Rodríguez

Madrid, junio de 2022

Resumen

La predicción de las rentabilidades bursátiles ha sido un tema que ha suscitado la atención de los inversores durante décadas. Este proceso asume que mediante un análisis de la información histórica disponible públicamente sobre los valores que forman el índice que combine variables macroeconómicas, tendencias de la industria, precios de las materias primas e indicadores técnicos relativos al mercado de valores, así como de diversos acontecimientos, noticias financieras y/o opiniones de inversores, se puede establecer una relación predictiva con sus rendimientos futuros.

En este trabajo se ha construido un modelo de predicción del precio de las acciones del Standard & Poor's 500 (S&P 500) mediante una estructura sencilla basada en redes neuronales artificiales. Para proporcionar una evaluación experimental del rendimiento del modelo, se ha llevado a cabo una comparación de las rentabilidades conseguidas por una estrategia de inversión basada en el modelo y otra basada en la compra y mantenimiento de un ETF del S&P 500. Los resultados muestran que, si bien el modelo es capaz de lograr rendimientos positivos relevantes, es al mismo tiempo incapaz de batir al mercado. Se han discutido las posibles limitaciones para la predicción bursátil, como el acceso a toda la información relevante, así como las posibles direcciones de investigación futuras.

Palabras clave

Hipótesis de los mercados eficientes; mercados financieros; inteligencia artificial; redes neuronales; predicción en bolsa.

Abstract

The stock market has been an attractive field for investors to try to create useful predictions to maximize their returns. The idea that an analysis of macroeconomic variables, industry trends, commodities' prices and technical indicators regarding the stock market, as well as the influence of various events, financial news, and/or investors' opinions could be exploited to make reliable predictions and achieve higher profitability has been in the mind of investors for a long time.

In this thesis, a stock price prediction model for the Standard & Poor's 500 (S&P 500) has been built using a simple structure based on artificial neural networks. To provide an experimental evaluation of the model's performance, a comparison of the returns achieved by an investment strategy based on the model and one based on the purchase and holding of an S&P 500 ETF has been carried out. The results show that, while the model can achieve relevant positive returns, it is at the same time unable to beat the market. Possible limitations for stock market prediction, such as access to all the relevant information, as well as possible future research directions have been discussed.

Key words

Efficient Market Hypothesis; financial markets; machine learning; neural networks; stock return forecasting.

Índice

ÍNDICE DE FIGURAS	V
1. INTRODUCCIÓN	1
1.1 OBJETIVOS.....	1
1.2 LIMITACIONES	3
2. REVISIÓN BIBLIOGRÁFICA.....	4
2.1 ANÁLISIS BURSÁTIL	5
2.1.1 <i>¿Por qué el Standard & Poor's 500?</i>	5
2.1.2 <i>La hipótesis de los mercados eficientes y las finanzas del comportamiento</i>	6
2.2 INTELIGENCIA ARTIFICIAL	7
2.2.1 <i>Redes neuronales artificiales</i>	7
2.2.2 <i>Aprendizaje supervisado</i>	9
2.3 APLICACIÓN DE LA INTELIGENCIA ARTIFICIAL A LOS MERCADOS	9
2.4 ELECCIÓN DE VARIABLES.....	11
2.4.1 <i>Bloques</i>	11
3. METODOLOGÍA.....	16
3.1 FUENTES DE DATOS	16
3.1.1 <i>Datos obtenidos de Bloomberg</i>	17
3.1.2 <i>Datos macroeconómicos</i>	17
3.1.3 <i>Datos obtenidos de la web de Aswath Damodaran</i>	18
3.1.4 <i>Otros</i>	19
3.2 ELABORACIÓN	19
3.2.1 <i>Fusión de los datos</i>	19
3.2.2 <i>Tratamiento</i>	20
3.2.3 <i>Cálculo de variables compuestas</i>	20
3.2.4 <i>Normalización</i>	22
3.3 APLICACIÓN DE AZURE MACHINE LEARNING	23
3.3.1 <i>Diseño del diagrama</i>	23
3.3.2 <i>Configuración inicial</i>	25
3.3.3 <i>Entrenamiento y validación</i>	26
4. RESULTADOS.....	28
4.1 EVALUACIÓN.....	28
4.1.1 <i>Evaluación de la precisión</i>	28
4.1.2 <i>Análisis de la curva ROC</i>	29
4.2 RESULTADO SOBRE EL DATASET SIN EL ÚLTIMO TRIMESTRE DE 2021	30

4.3 PREDICCIÓN	32
4.3.1 Resultados.....	33
4.4 OPTIMIZACIÓN DEL MODELO PARA INVERSIÓN	34
4.4.1 Ajuste.....	35
4.4.2 Aplicación al modelo de predicción.....	36
4.5 COMPARACIÓN DE ESTRATEGIAS DE INVERSIÓN	36
4.5.1 Estrategia de comprar y mantener.....	37
4.5.2 Estrategia basada en el modelo de predicción.....	38
4.5.3 Comparativa de rendimiento.....	39
5. DISCUSIÓN.....	41
6. CONCLUSIONES.....	44
7. BIBLIOGRAFÍA.....	46
ANEXOS.....	50
ANEXO 1. CORRESPONDENCIA ENTRE LOS NOMBRES DE LAS VARIABLES.....	50
ANEXO 2. VARIABLES INCORPORADAS Y TIPO DE DATO	51

Índice de figuras

Figura 1. Estructura de una red neuronal simple.....	8
Figura 2. Diagrama del modelo basado en el <i>dataset</i> sin el último trimestre de 2021.....	24
Figura 3. Espacio ROC.....	30
Figura 4. Curva ROC del modelo basado en el <i>dataset</i> sin el último trimestre de 2021.....	31
Figura 5. Estructura del modelo de predicción sobre el último trimestre de 2021.....	32
Figura 6. Curva ROC del modelo de predicción sobre el último trimestre de 2021.....	33
Figura 7. Evolución de la cotización del SPY del 13 de octubre al 12 de noviembre.....	37
Figura 8. Evolución del rendimiento de la estrategia de inversión basada en el modelo.....	38
Figura 9. Comparación del rendimiento de las dos estrategias de inversión.....	39

1. Introducción

En las dos últimas décadas se han producido cambios importantes en el entorno de los mercados financieros. La aparición de nuevas plataformas que han democratizado el acceso a inversiones y el desarrollo de los medios de comunicación han impulsado el interés de los ciudadanos en estos mercados hasta el punto de que actualmente los inversores particulares representan casi una cuarta parte del volumen de negociación, y se prevé que esta tendencia continúe a largo plazo.

En este contexto, cualquier herramienta que permita mejorar la toma de decisiones en el ámbito financiero resulta sumamente atractiva. Cualquier inversor soñaría con disponer de un instrumento que redujese su incertidumbre y al mismo tiempo incrementase los beneficios que obtiene en los mercados. Sin embargo, la predicción bursátil es un problema extremadamente complejo porque, sencillamente, hay demasiados factores que pueden afectar a la cotización de una acción.

La aplicación de inteligencia artificial (IA) a este campo resulta muy interesante porque aborda el problema desde una nueva perspectiva, pero resulta demasiado compleja para una persona sin la formación técnica necesaria. La aparición de plataformas que democratizan el acceso a este tipo de herramientas, sin necesidad de conocer lenguajes de programación, supone una oportunidad muy atractiva para aquellos que quieran adentrarse en este mundo.

1.1 Objetivos

Este Trabajo Fin de Grado (TFG) tiene como objetivo general desarrollar un algoritmo clasificatorio basado en redes neuronales capaz de predecir el comportamiento de las cotizaciones de las acciones que forman parte del índice Standard & Poor's 500 (S&P 500), a través de un modelo que pueda servir como referencia para futuras investigaciones.

Tratar de predecir el comportamiento de los mercados es una tarea extremadamente compleja, puesto que hay demasiados factores que potencialmente pueden afectar a la cotización de un valor.

El desarrollo de las bolsas modernas, entendidas como lugares de reunión en donde se negocia la compra y venta de valores de compañías, fomentó que la iniciativa privada funcionase como impulso al desarrollo económico, permitiendo que aquellas empresas con grandes necesidades de capital pudiesen encontrar financiación. Esta posibilidad, unida a un capital limitado por parte del inversor, trajo consigo un inconveniente: no todas las inversiones eran igual de rentables. Y es este motivo el que ha permitido el desarrollo de las estrategias de inversión. Inicialmente, estas eran totalmente subjetivas, basadas en la intuición de los propios inversores. Pero, con el tiempo, las estrategias han evolucionado hacia métodos en los que prima el análisis estadístico, técnico o fundamental, buscando un aumento de las ganancias y una minimización del riesgo de forma objetiva. Sin embargo, pese a este desarrollo ninguna de estas técnicas ha demostrado suficiente consistencia a la hora de mejorar los rendimientos del mercado.

No obstante, la creciente cantidad de información disponible ha hecho posible la introducción de nuevas técnicas basadas en IA en las estrategias de inversión. Su incorporación al proceso de toma de decisiones podría traducirse en una ventaja competitiva.

De este objetivo general deriva la principal pregunta de investigación del trabajo: ¿es posible desarrollar una estrategia de inversión basada en el uso de IA cuyo rendimiento sea comparable al del mercado? Con este fin, se establecen los siguientes dos objetivos específicos:

- A través de una aproximación sencilla, asequible para personas sin conocimientos técnicos elevados de programación, construir un modelo empleando redes neuronales artificiales capaz de predecir los precios de las acciones cotizadas en el S&P 500, basado en datos históricos de diferentes tipos: características de la compañía, análisis de la industria en la que se encuentra, situación macroeconómica...
- Valorar si una estrategia de inversión basada en el modelo diseñado resultaría más eficiente que otras estrategias alternativas, en concreto la inversión pasiva en un fondo cotizado (ETF, del inglés *Exchange-Traded Fund*) referenciado al S&P 500.

1.2 Limitaciones

Debido a que la propuesta de valor de este trabajo se basa en el desarrollo de un modelo de predicción de precios de acciones cotizadas, algo sobre lo que no existe constancia de que se haya logrado hacer hasta la fecha, la investigación presenta múltiples limitaciones y desafíos.

Muchas de estas limitaciones se producen por intentar predecir el comportamiento de los mercados. De ellas, la principal estará causada por el sobreajuste que pueda tener el modelo. Puesto que la base sobre la que este se construye son datos históricos, es posible que el modelo desarrollado refleje una estrategia de inversión que habría sido exitosa en el pasado pero que, por cambios en la situación económica, social o tecnológica, tanto a nivel regional como a nivel global, no funcione en otro momento del tiempo. Igualmente, la imposibilidad de predecir otros eventos, como conflictos de índole política o militar y sus repercusiones económicas, obstaculiza la creación de un modelo que incluya este tipo de consideraciones.

Asimismo, existe una limitación geográfica que impide que el resultado obtenido por el modelo sea extrapolable a otras regiones, puesto que este ha sido construido en base a la información disponible sobre un índice estadounidense. La condición de Estados Unidos como primera potencia económica mundial y la elevada capitalización de sus empresas refleja un mayor interés en sus mercados por parte de los inversores; en consecuencia, hay una mayor disponibilidad de datos sobre sus empresas. Por esta razón, la replicación de este modelo en índices europeos, aunque no imposible, implicaría el uso de otras fuentes y un trabajo de recogida de datos mucho más tedioso.

2. Revisión bibliográfica

Hasta hace pocas décadas, las finanzas han sido una disciplina caracterizada por la escasez de datos. La mayoría de las bolsas de valores no proporcionaban apenas información, ni se guardaba un historial intradiario más allá de lo exigido por la normativa. En la actualidad, la proliferación de datos y el aumento de la potencia de cálculo han cambiado radicalmente el sector financiero, desde el análisis de carteras hasta la gestión de riesgos (Fang & Zhang, 2016). En este contexto, la predicción de las rentabilidades bursátiles ha sido un tema que ha suscitado la atención de numerosos investigadores. En su esencia, este proceso supone que la información histórica disponible públicamente tanto a escala macroeconómica, como industrial o específica de las empresas tiene alguna relación predictiva con los rendimientos futuros.

Esta línea de pensamiento choca frontalmente con la hipótesis del mercado eficiente (HME) enunciada por el economista estadounidense Eugene Fama en 1965. En términos simples, esta hipótesis establece que «un mercado es eficiente con respecto al conjunto de información θ_t si es imposible obtener beneficios económicos negociando sobre la base del conjunto de información θ_t » (Jensen, 1978, pág. 98). El cumplimiento de esta hipótesis implica que el precio de una acción refleja en todo momento la totalidad de información disponible sobre esta y, en consecuencia, no sería posible realizar predicciones sobre su precio.

Sin embargo, desde hace décadas numerosos autores han estudiado escenarios en los que se han generado retornos anormales causados por ineficiencias del mercado, lo que los ha llevado a mostrar reticencia frente a la hipótesis desarrollada por Fama (Jensen, 1978). Este escepticismo unido al desarrollo tecnológico ha promovido una renovación del interés en la IA y su aplicación a las finanzas. En los últimos años, diversos estudios han explorado el uso de técnicas de aprendizaje automático, logrando resultados que apoyan el desafío a la teoría del economista estadounidense.

Desde principios de este siglo existe abundante literatura sobre el uso de IA en el desarrollo de modelos capaces de predecir el comportamiento de los mercados, así como sobre los requerimientos necesarios para construirlos (Walczak, 2001). No obstante, la característica común de estos artículos es que se focalizan en el uso de series temporales de diversas duraciones de precios y volúmenes, características

del análisis técnico, como base sobre la que sostener sus predicciones (Kolarik & Rudorfer, 1997). Pese a ello, existe también un número reducido de artículos que exploran el uso de modelos con varias variables o en los que el peso de la predicción no descansa sobre las series temporales de precios, sino sobre el perfil financiero de una empresa o datos macroeconómicos, lo que se puede clasificar como análisis fundamental (Huang, Capretz, & Ho, 2019).

2.1 Análisis bursátil

2.1.1 ¿Por qué el Standard & Poor's 500?

El S&P 500, junto al Dow Jones Industrial Average (DJIA) y el Nasdaq es uno de los tres principales índices de referencia del mercado norteamericano. Pese a que los tres son algunos de los más citados del mundo y contienen compañías altamente relevantes que podrían servir como base para el desarrollo de este trabajo, la elección del índice S&P 500 se ha llevado a cabo por dos razones:

- Solo incluye compañías con base en los Estados Unidos. De los tres índices mencionados, el Nasdaq Composite es el único que no se limita geográficamente, incluyendo valores de empresas con sede en el extranjero (Chen, 2022). Debido a que algunas de las variables empleadas en el modelo limitan la información proporcionada a Estados Unidos, incluir compañías con sede en el extranjero haría que cualquier medida política, económica o cualquier evento pudiese afectar de manera desigual a las compañías. Pese a que es cierto que muchas de estas empresas operan de forma internacional y existe la posibilidad de que determinadas condiciones en algunos territorios puedan afectarles de manera desigual, se ha tratado de reducir esta variabilidad haciendo que la región de la sede sea uniforme.
- Debido a que el S&P 500 recoge las 500 compañías americanas con mayor capitalización de mercado, el índice contiene todo tipo de industrias (Kenton, 2022). Por el contrario, en el DJIA sólo se incluyen 30 empresas de primera línea que, si bien representan distintos sectores, ofrecen una imagen más limitada y, por lo tanto, menos representativa del panorama industrial estadounidense.

Además, debido a que el S&P 500 es uno de los índices más citados del mundo, también es uno sobre el que más información disponible existe. Al contrario que

otros índices como el europeo STOXX Europe 600, herramientas como Bloomberg computan muchos más índices y recogen más información sobre las empresas norteamericanas, lo que permitirá incluir algunas variables y ratios no disponibles para empresas europeas.

2.1.2 La hipótesis de los mercados eficientes y las finanzas del comportamiento

Como se ha mencionado previamente, el economista estadounidense Eugene Fama presentó en 1963 su hipótesis del mercado eficiente (HME). Esta teoría afirma que, en cualquier momento, el precio de los activos presentes en los mercados refleja plenamente toda la información disponible sobre sus valores futuros y, por lo tanto, es imposible obtener beneficios económicos negociando con dicha información como base. Así, los precios de las acciones realizan un «paseo aleatorio», puesto que cada vez que de forma aleatoria existe nueva información disponible esta se refleja inmediatamente en el precio de las acciones. El precio futuro reflejará esta información. La consecuencia de este «paseo aleatorio» es que no es posible que un inversor gane al mercado, puesto que los precios no son previsible.

Años más tarde, Fama propuso tres modalidades de eficiencia: la forma débil, la semifuerte y la fuerte (1970). La forma débil afirma que la información del precio pasado está totalmente incorporada en el precio actual y, por lo tanto, no tiene ningún poder de predicción, haciendo inservible el análisis técnico. La forma semifuerte amplía este conjunto de información e indica que toda aquella hecha pública está reflejada en los precios. Por último, la forma fuerte asume que toda la información, tanto pública como privada, está reflejada en el precio de cotización y, por lo tanto, incluso si se contase con información que no fuese de dominio público, ningún agente del mercado podría ser capaz de obtener beneficios de forma consistente.

Por lo tanto, puesto que las acciones siempre se negocian a su valor justo en las bolsas, según la HME es imposible que los inversores compren acciones infravaloradas o vendan acciones a precios inflados (Downey, 2021). La implicación de esta hipótesis es que el análisis técnico, fundamental o de series temporales no producirían rendimientos extraordinarios comparados con los del propio mercado.

Aunque la HME contempla la existencia de inversores irracionales, considera que la mayoría son inversores racionales y, por ello, el mercado se comporta de manera racional.

Sin embargo, el hecho de que muchos participantes en el mercado puedan batir sistemáticamente los rendimientos de un índice es un indicio de que la teoría puede no ser cierta en la práctica. En consecuencia, se han desarrollado teorías que abordan el porqué de los comportamientos del mercado desde otra óptica, cuestionando la visión de Fama. Las finanzas del comportamiento se enmarcan en esta crítica y defienden la irracionalidad del ser humano, entendiendo que la toma de decisiones de los agentes del mercado se puede ver influenciada por emociones y sesgos cognitivos, lo que podría explicar por qué los desplomes bursátiles no se basan en información que se produce al azar, sino en momentos de miedo abrumador de los inversores (Barberis & Thaler, 2002).

Puesto que en esencia este TFG pretende «desafiar» la lógica del «paseo aleatorio», se trabajará desde la hipótesis de que los mercados no son eficientes.

2.2 Inteligencia artificial

Hasta hace pocas décadas, la IA, era un concepto que sólo se encontraba las novelas de ciencia ficción, enmarcado en mundos con robots y máquinas que manifestaban un comportamiento semejante al de los seres humanos. Actualmente entendemos este concepto como «la ciencia y la ingeniería de la creación de dispositivos inteligentes, especialmente programas informáticos» (Pothen, 2021). Una de las formas de alcanzar esa «inteligencia» es a través de estructuras que se asemejan a aquellas propias del cuerpo humano.

2.2.1 Redes neuronales artificiales

Una red neuronal es un método de IA que «enseña» a las máquinas a procesar datos tratando de reproducir inorgánicamente la actividad biológica del cerebro humano. En esencia, se imita el comportamiento de las neuronas cerebrales que, interconectadas, «aprenden» en base a la experiencia (Schmueli, Bruce, Yahav, Patel, & Lichtendahl, 2018).

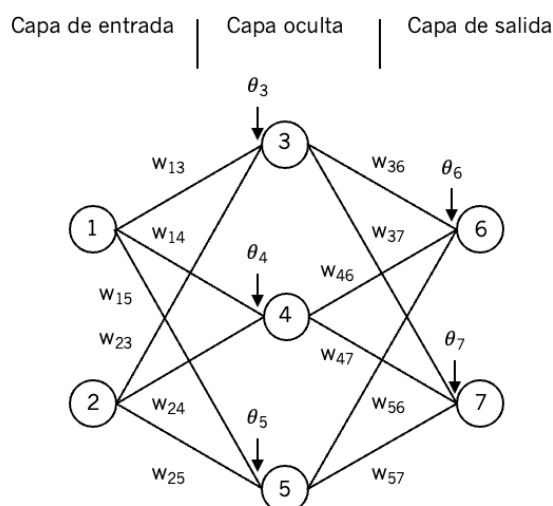
Las redes neuronales artificiales son modelos flexibles que pueden ser usados tanto para clasificación como para realizar predicciones, pues su estructura es capaz de capturar relaciones complejas entre los predictores y el desenlace.

Al contrario que en los modelos de regresión lineal, en donde el usuario especifica directamente esta relación, en los modelos basados en redes neuronales la propia red trata de establecerla en base a los datos que se le proporcionan. Uno de los modelos más populares son las *multilayer feedforward networks*, que cuentan con una estructura formada por una capa de entrada (*input layer*), constituida por nodos que aceptan los valores de los predictores; una capa de salida y una serie de capas ocultas (*hidden layers*) que se sitúan entre la de entrada y la de salida.

Este tipo de estructura tiene conexiones unidireccionales entre los nodos presentes en una capa y todas las neuronas de la capa que se sitúa a continuación. Cada vez que el desenlace de una unidad viaja a través de una conexión a otra unidad, se multiplica por el peso de dicha conexión. Y en cada uno de esos nodos los pesos se suman y se les añade una constante (*bias*), que controla su nivel de contribución, tal y como se muestra en la Figura 1. Una vez todos los *inputs* de cada nodo se han sumado, se aplica una función de activación al resultado (Pizarroso, Portela, & Muñoz, 2021).

Figura 1

Estructura de una red neuronal simple



Nota. Los círculos representan los nodos, las w_{ij} los pesos de las conexiones y θ_j la constante (*bias*). Adaptado de *Data Mining for Business Analytics: Concepts, Techniques and Applications in R* (p. 274) por G. Shmueli et al, 2018, John Wiley & Sons.

En un modelo simple, con una capa de entrada, una única capa oculta y una capa de salida, los nodos de la capa oculta recibirían los *outputs* de la capa de entrada y sus propios *outputs* se calcularían teniendo en cuenta el peso y el *bias* mencionados (DeepAI, 2017).

2.2.2 Aprendizaje supervisado

Es necesario distinguir entre las dos modalidades de aprendizaje automático que se pueden usar en *machine learning*: el aprendizaje supervisado y el no supervisado. Se seleccionará uno u otro dependiendo de los factores que afecten a la toma de decisiones, de las necesidades del problema a solucionar...

El modelo que se pretende construir se basará en aprendizaje supervisado, que se usa cuando la finalidad es clasificar o predecir. Lo característico de este tipo de aprendizaje es que es necesario que el conjunto de datos incluya valores conocidos del desenlace que se pretende estudiar, a los que se denomina datos de entrenamiento. El modelo usará estos datos para «comprender» cómo se relacionan las variables predictoras y el desenlace entre sí. Una vez establecidas estas relaciones, el siguiente paso es aplicarlas a un nuevo conjunto de datos de las mismas características, pero en los cuales el desenlace está oculto. El modelo realizará una clasificación o una predicción para dichos desenlaces y se compararán los resultados con los desenlaces reales.

El otro tipo de aprendizaje automático es el aprendizaje no supervisado. Estos modelos se emplean cuando no hay un desenlace que predecir o clasificar. En este caso, los algoritmos solo se ajustan utilizando datos de entrada y dejando de lado los de salida. Entre los algoritmos más utilizados dentro de este tipo de aprendizaje automático está el agrupamiento (*clustering*), que divide los datos en grupos con características similares (Luna, 2018).

2.3 Aplicación de la inteligencia artificial a los mercados

Los mercados de valores son sistemas muy dinámicos en los que la mayoría de sus componentes se relacionan entre sí de forma no lineal. La capacidad de las redes neuronales para descubrir relaciones de este tipo entre las variables de entrada las convierte en ideales para tratar de modelar este tipo de sistemas (Lawrence, 1997).

Estas variables dependerán del entorno de aprendizaje en el que se vaya a buscar que la red «aprenda». Puesto que estos sistemas de IA buscan predecir cuándo comprar o vender valores en base a información previa disponible en el mercado, en muchos casos el reto se basa en determinar qué variables y datos de entrada serán los más adecuados, así como reunir suficiente información para entrenar el sistema adecuadamente.

Los datos de entrada pueden proporcionar información de todo tipo y, entre otros, pueden tener su origen en el análisis técnico, como ocurriría en el caso de una media móvil; en el análisis fundamental, en el caso de usar ratios como por ejemplo el beneficio por acción; o simplemente basarse en información bruta del mercado, como el precio; y del entorno macroeconómico, como la inflación. Lo importante es que estos datos de entrada deben permitir a la red neuronal generalizar el comportamiento del mercado.

Debido a la sencillez con la que una red neuronal es capaz de representar cualquier serie temporal, el uso más común de este tipo de estructuras es precisamente el análisis de series temporales de precios con el fin de predecir las cotizaciones futuras (Kolarik & Rudorfer, 1997). Sin embargo, en la literatura los modelos multivariantes más complejos han demostrado mejor rendimiento frente al mercado. Algunos experimentos, como el llevado a cabo en la década de 1990 que modeló el rendimiento de la Bolsa de Johannesburgo, han sido capaces de predecir la dirección del mercado y refutar la hipótesis de los mercados eficientes (Van Eyden, 1996). Este sistema empleó decenas de indicadores de diversas índoles: rendimiento, tendencias de volumen, índices internacionales de referencia, tipo de cambio, tipos de interés...

La correcta selección de las variables es un aspecto fundamental en la elaboración del modelo, pues el entrenamiento empleando elevados volúmenes de datos históricos requiere mucho tiempo y esfuerzo informático y puede hacer que la red aprenda información no deseada en el conjunto de datos. Esta consideración es especialmente importante a la hora de decidir desde qué momento del tiempo se deben extraer datos, puesto que aquellos muy antiguos pueden llevar a la red a aprender patrones o factores que ya no son importantes o valiosos.

Asimismo, el objetivo de un modelo empleado para predecir precios es batir al mercado o a diferentes índices. Algunos estudios han mostrado que las redes neuronales son capaces de batir al índice, pero el lograrlo de manera consistente,

especialmente en condiciones económicas cambiantes presenta muchas más dificultades.

2.4 Elección de variables

Una de las características que tienen en común la mayoría de las publicaciones sobre la creación de modelos de predicción de cotizaciones es su apuesta por el uso de un número reducido de variables entre las que destacan las series temporales de precios. Sin embargo, la aproximación de este trabajo pretende parecerse más a la llevada a cabo por Van Eyden en su artículo sobre la bolsa de Johannesburgo mencionado en el epígrafe anterior. Para ello, el *dataset* se ha diseñado aunando cuatro bloques de variables en los que estas se encuadran en base a su tipo: propias del análisis fundamental, específicas de la industria, específicas de la compañía y macroeconómicas.

2.4.1 Bloques

Primer bloque: Variables propias del análisis fundamental

Las variables que forman parte de este bloque son en su mayoría ratios financieros, relaciones establecidas entre unidades financieras a través de las cuales es posible analizar la situación de una compañía y compararla con otras de industrias o características similares (Wolkers Kluwer, 2021). Estos ratios se usan con frecuencia al realizar análisis fundamental de valores, un método de análisis a través del cual se evalúa el valor intrínseco de un activo y aquellos factores que podrían influir en su precio en el futuro.

- El ROE (del inglés *Return on Equity*) o rentabilidad de los fondos propios se considera un indicador de la rentabilidad de una empresa y de su eficiencia a la hora de generar beneficios. Se calcula dividiendo los ingresos entre los fondos propios. Se ha incluido en el modelo el ROE reportado por la empresa durante el último trimestre.
- El ratio P/E (del inglés *Price to Earnings*) es la relación entre el precio de una acción y el beneficio por acción de la empresa. Se calcula como el último precio dividido por el beneficio por acción de los últimos 12 meses.

- EV/EBITDA (*Enterprise Value to EBITDA*) y Change in EV/EBITDA son variables que se refieren al ratio calculado dividiendo el valor de la empresa entre el EBITDA (*Earnings Before Intest, Taxes, Depreciation and Amortization*). La característica de este ratio es que tiene en cuenta los niveles de deuda y de efectivo de una empresa, además del precio de sus acciones, y relaciona ese valor con su rentabilidad en efectivo (Hayes, 2022). *Change in EV/EBITDA* representa el cambio en este ratio durante el último trimestre.
- *Change in Net Debt* refleja el cambio en la deuda neta de la empresa durante el último trimestre. La deuda neta se calcula restando el efectivo y sus equivalentes a la deuda de la empresa.
- EPS T12M y BEst EPS BF12M son variables relacionadas con el beneficio por acción (EPS, del inglés *Earnings per Share*). EPS T12M indica el beneficio por acción durante los últimos 12 meses y BEst BF12M es una media de las estimaciones de los analistas para los próximos 12 meses sobre el beneficio ajustado por acción.

Segundo bloque: Variables de la industria

Si bien los ratios son útiles para conocer la situación particular de una empresa, para poder usarlos de cara a establecer comparaciones efectivas es necesario que las compañías tengan características similares e, idealmente, que pertenezcan al mismo sector o industria. Si esto no fuese así y se estableciese una comparativa basada en ratios entre una compañía tecnológica como Apple, Inc. y otra farmacéutica como Pfizer, Inc., esta no proporcionaría información útil ser empresas con una estructura de capital muy diferente. Sin embargo, una comparativa entre Abbvie Inc., y Pfizer Inc, ambas del sector farmacéutico, podría permitir establecer valoraciones útiles entre ambas. Por esta razón, se han incluido variables que proporcionan información sobre los ratios medios de las compañías que forman parte de una industria, con el fin de establecer comparaciones entre los ratios propios de las compañías y la media de la industria que permitan identificar aquellas con un desempeño por encima de la media en determinadas áreas.

- *Industry average P/E ratio*, o el ratio P/E medio de las empresas que forman parte de una industria concreta.

- *Industry average ROE*, la rentabilidad media de los fondos propios que tienen las compañías que forman la industria.
- *Industry average expected growth*, que representa la tasa media de crecimiento esperada en el beneficio por acción en las empresas de la industria a lo largo de los próximos cinco años.

Tercer bloque: Variables específicas de la compañía

Este bloque de variables aporta información sobre el comportamiento de la compañía en los mercados. Representarán *inputs* relevantes para el modelo al mostrar la tendencia que ha seguido a lo largo de los últimos meses, la volatilidad que suele tener en relación al índice, etc.

- El *momentum* es la variación porcentual durante los últimos 6 meses de la media móvil del precio de la acción en relación con un índice de referencia. Refleja la «inercia» que puede tener una tendencia de precios a seguir subiendo o bajando durante un período de tiempo determinado.
- La variable «*Recommendation*» proporciona la opinión de los analistas sobre el valor. La calificación se calcula convirtiendo cada una de las recomendaciones actuales de los analistas en un número de 1 a 5 y tomando la media. Teniendo en cuenta la teoría de las finanzas del comportamiento, sería posible que factores como la recomendación o no de los analistas pudiese tener un impacto positivo o negativo sobre el valor de una acción.
- La beta (*Three-Month Beta*) es una medida que indica cuánto varía en porcentaje el precio de un valor si se produce una variación del uno por ciento en un índice de mercado representativo. En este caso, el valor beta se determina comparando los movimientos de precios del valor y del S&P 500 para los últimos tres meses de datos diarios, al considerarse que una beta basada en movimientos muy lejanos en el tiempo podría ser poco representativa de la situación actual de la empresa.
- La previsión de crecimiento a largo (*BEst Long-Term Growth*) representa el aumento anual previsto de los beneficios de explotación por acción durante el próximo ciclo económico completo de la empresa. En general, estas previsiones se refieren a un periodo de entre tres y cinco años.

- La capitalización de mercado (*Market Cap*) es el valor que tienen en el mercado las acciones de una compañía. Se calcula multiplicando el número de acciones de esta por el precio actual por acción. Se considera importante incluir esta variable puesto que las distintas coyunturas económicas no afectan por igual a las compañías grandes o pequeñas.
- Además, se ha incluido el cambio de precio experimentado por el valor en los últimos 100 (*Price change (%) 100 D*) y 30 (*Price change (%) 30 D*) días, así como el desenlace («*Outcome*»), que funciona como variable dependiente y representa el precio de la acción un mes después del momento de la recogida del resto de datos.

Cuarto bloque: Variables macroeconómicas

Las variables macroeconómicas reflejan de forma numérica la realidad económica que rodea a las empresas y proporcionan información sobre la etapa ciclo en la que se encuentra una economía (López, 2022).

- La inflación («*Inflation*») o deflación representa un aumento o disminución general de los precios y servicios ofertados en una economía de mercado. Este factor es relevante puesto que en función del sector al que pertenezca la empresa el aumento de los precios puede tener diferentes impactos.
- El tipo de cambio («*Exchange Rate*»), en este caso entre el euro y el dólar, y su variación durante los últimos 3 meses («*Exchange Rate Change (%) 3 M*») reflejan cómo ha variado el valor de una divisa en relación con otra. Algunos autores han sugerido que una apreciación de la moneda podría impulsar la bolsa mientras que sucedería lo contrario en caso de depreciación (Dimitrova, 2005).
- Los tipos de interés son otra información fundamental que ha demostrado tener impacto sobre las cotizaciones en la bolsa. En concreto, existe una relación inversa entre tipos de interés y los precios (Siddiqui, 2003). Se ha argumentado que cuando los tipos de interés caen a niveles poco atractivos, los inversores tienen un mayor incentivo para invertir en la bolsa, lo que se traduciría en un aumento de los precios. Se han incluido el valor y el cambio trimestral del bono americano a 10 años («*10Y Bond*» y «*10 Y Bond Change (3 M)*»), a tres meses («*3 M Bond*» y «*3 M Bond Change (3 M)*»), y el cambio mensual en el tipo de interés de los fondos federales, que es el tipo de interés

objetivo fijado por la Reserva Federal al que los bancos comerciales se prestan sus excedentes de reservas hasta el día siguiente («*Fed Funds Rate Change (1M)*»).

- El comportamiento del mercado, en concreto del S&P 500 durante el último mes («*S&P 500 Change (%) 1 M*») es otro factor relevante que debe ser considerado, pues la tendencia general del mercado puede aportar valor al modelo de predicción.
- El VIX es una medida de la volatilidad del S&P 500 durante los últimos 30 días. Valores muy elevados de este indicador reflejan periodos de inestabilidad económica en los que es posible encontrar variaciones de precio más grandes en la bolsa.
- «*WTI Price*» y su variación trimestral («*WTI Price change (%) 3 M*») reflejan los cambios en el precio del barril de West Texas Intermediate (WTI). Junto al Brent son las dos referencias más importantes de los precios del petróleo. Este producto tiene un gran impacto sobre la economía al ser clave en el suministro de mercancías para todo tipo de industrias, por lo que su inclusión puede aportar información relevante al modelo.
- El precio del oro («*Gold*» y «*Gold Price Change (%) 3M*») también puede ser un reflejo de las condiciones económicas. Este metal precioso puede funcionar como reserva de valor en tiempos de crisis económica o de alta inflación.
- El índice de incertidumbre sobre políticas económicas («*EPU*») y su variación («*EPU Change (%)*»), calculado en base al número de menciones de determinados términos en artículos de los 10 periódicos estadounidenses más importantes, es otra métrica relevante que se ha incluido en el modelo. Algunos autores han sugerido que la incertidumbre sobre este tipo de políticas puede elevar la volatilidad de la bolsa y reducir la inversión y el empleo en determinados sectores como defensa o la construcción (Baker, Bloom, & Davis, 2015).

3. Metodología

Este trabajo analiza trimestralmente los valores de varios bloques de variables por cada una de las 500 empresas que forman parte del índice Standard & Poor's 500 entre enero de 2017 y noviembre de 2021, lo que representa los últimos cinco años de datos disponibles. Se ha elegido este periodo para que los datos sean lo más comparables posibles, ya que una comparación entre distintas décadas podría no ser informativa al no representar condiciones económicas similares.

En primer lugar, se ha llevado a cabo una revisión bibliográfica tanto de fuentes primarias como secundarias, sobre la que se ha justificado la elección de los bloques de variables. A continuación, para la recopilación de datos se han empleado diferentes fuentes, desde Bloomberg hasta sitios web gratuitos como Investing.com, Yahoo! Finance o Federal Reserve Bank of St. Louis. El documento en el que se han reunido todos estos datos ha sido una hoja de cálculo de Microsoft Excel que posteriormente se ha convertido a formato CSV para su procesamiento.

Para el diseño del modelo se ha empleado la plataforma Azure Machine Learning, y se ha creado un modelo de clasificación basado en redes neuronales para predecir un desenlace con únicamente dos valores (*two-class neural network*). Dicho desenlace es si ha existido (1, verdadero) o no (0, falso) un incremento en el precio de la acción durante el mes posterior a la fecha de las observaciones.

3.1 Fuentes de datos

La elaboración de un conjunto de datos de estas características precisa de un alto volumen de información, lo que hace imposible su construcción de forma totalmente manual. Por esta razón se ha recurrido a una herramienta como Bloomberg para obtener la información específica de cada una de las empresas, pues estos datos son los que mayor variabilidad presentan. En el caso de la información englobada en el bloque macroeconómico, esta refleja un valor común para todas las empresas en un instante determinado del tiempo. Es decir, si nos situamos en una fecha concreta, aunque la cotización de cada uno de los valores será diferente (información específica de cada empresa), el precio de mercancías como el petróleo o el tipo de interés de las letras del Tesoro de Estados Unidos a tres meses (variables macroeconómicas) será común.

3.1.1 Datos obtenidos de Bloomberg

El terminal Bloomberg es un sistema informático que permite a sus usuarios acceder a la base de datos de Bloomberg, a través de la cual es posible analizar información sobre los mercados en tiempo real. Esta herramienta ha resultado esencial en la construcción de nuestro *dataset*, puesto que nos permite situarnos en una fecha determinada del tiempo y obtener datos relativos a ese instante temporal.

Debido a que el S&P 500 es un índice que engloba a las 500 compañías más grandes de los Estados Unidos en base a su capitalización bursátil, sus componentes están sujetos a cambios. Se pueden producir entradas y salidas a medida que surgen nuevas empresas que crecen en capitalización y reemplazan a otras compañías. Si se emplea la actual lista de componentes y se busca su información en los distintos momentos que queremos estudiar, estaríamos teniendo en cuenta únicamente aquellos valores que han «sobrevivido» en el índice desde 2017, además de otros nuevos que se hayan incorporado y que no estuviesen presentes en ese momento. Sin embargo, valores que formasen parte del S&P 500 en 2017 que no sean ya constituyentes no se habrían incluido, por lo que nuestro *dataset* no estaría reflejando el comportamiento real de los mercados: esto se conoce como sesgo de supervivencia. Usando el terminal Bloomberg evitamos este problema, puesto que si nos situamos en una fecha concreta la información que nos proporciona incluye únicamente los constituyentes en el momento señalado en el tiempo.

De esta fuente se ha obtenido información sobre 12 variables en cada uno de los trimestres estudiados. Algunas de ellas han formado parte del *dataset* final y otras han sido utilizadas para calcular otras variables compuestas, tal y como se explicará en el epígrafe 3.2.3. Las variables y su correspondencia con la información procedente de Bloomberg figuran en el Anexo 1.

3.1.2 Datos macroeconómicos

Debido a la amplia variedad de variables que se engloban dentro del apartado macroeconómico se ha recurrido a diversas fuentes para la obtención de sus valores. En concreto, se han utilizado cuatro sitios web: Yahoo! Finance, Investing.com, inflation.eu y Federal Reserve Bank of St. Louis:

Tabla 1*Variables macroeconómicas y su procedencia*

Variable	Dato más reciente	Nota	Procedencia
<i>Inflation</i>	Mes anterior	Variación interanual	Inflation.eu
<i>VIX</i>	Mismo día		Yahoo! Finance
<i>S&P 500</i>	Mismo día		Investing.com
<i>3-month bond</i>	Mismo día	Media semanal	Investing.com
<i>10-year bond</i>	Mismo día	Media semanal	Investing.com
<i>Federal Funds rate</i>	Mismo día	Variación mensual	Federal Reserve Bank of St. Louis
<i>Economic policy uncertainty</i>	Mes anterior		Federal Reserve Bank of St. Louis
<i>Exchange rate</i>	Mismo día	Media semanal	Federal Reserve Bank of St. Louis
<i>WTI</i>	Mismo día		Investing.com
<i>Gold</i>	Mismo día	Media semanal	Investing.com

3.1.3 Datos obtenidos de la web de Aswath Damodaran

Aswath Damodaran, profesor de Finanzas en la Stern School of Business de la Universidad de Nueva York, dispone de una base de datos propia con información sobre las diferentes industrias que actualiza anualmente, en base a los últimos doce meses de datos disponibles al comienzo de cada año. Puesto que las empresas publican su información de manera trimestral, la actualización de enero de 2017, momento en el que comienza la descarga de los datos utilizados en este trabajo, refleja la información disponible desde octubre de 2015 hasta septiembre de 2016 (Damodaran, 2022). El profesor Damodaran especifica que la información con la que trabaja proviene de distintas fuentes como Bloomberg o Morningstar, pero calcula todas las estadísticas que presenta empleando su propia metodología.

Debido a que estos datos se actualizan anualmente y nuestros puntos de datos tienen carácter trimestral la situación no es ideal, ya que la comparación será más adecuada durante los primeros trimestres de cada año que durante los últimos.

Las variables que se han utilizado en este trabajo procedentes de esta fuente son:

- *Industry average P/E ratio.*
- *Industry average ROE.*
- *Industry average expected growth.*

3.1.4 Otros

Para facilitar la fusión de los distintos bloques de datos en un único *dataset* se han añadido varias columnas de carácter identificativo: año, trimestre y un código de industria basado en la clasificación establecida por el profesor Damodaran. De ellas, sólo el código identificativo de cada industria ha formado parte del conjunto de datos final, debido a que se ha considerado que el tipo de industria en el que se enmarca cada compañía podría ser una variable relevante para llevar a cabo la predicción.

3.2 Elaboración

El modelo contaba inicialmente con 41 columnas agrupadas en los cuatro bloques explicados en la sección 2.4: información relativa al precio, macroeconómica, de la industria y específica de la compañía. Este número se ha reducido a 31 tras haber añadido nuevas variables «compuestas» (calculadas a partir de otras del conjunto de datos) y descartado tanto aquellas empleadas para su cálculo como otras de carácter meramente identificativo como el nombre de la empresa o de la industria.

3.2.1 Fusión de los datos

Una vez completada la descarga se ha fusionado la información obtenida de las distintas fuentes en un único conjunto de datos.

Al finalizar esta fase existen un total de 10.100 filas, conteniendo cada una información de 41 variables diferentes, lo que equivale a un *dataset* con 414.100 celdas.

3.2.2 Tratamiento

A continuación, se ha realizado un análisis para comprobar si las variables incorporadas al modelo contenían celdas con valores ausentes. Para un correcto funcionamiento del modelo es importante evitar su presencia, pues esta puede dar lugar a estimaciones incorrectas.

Dentro del conjunto de datos obtenido del terminal Bloomberg, algunas variables presentaban este tipo de valores. Debido a su alto número, se ha optado por eliminar aquellas filas que carecían de la información correspondiente con alguna de las variables obtenidas por esta vía. Se ha optado por esta decisión en lugar de otras alternativas, como inferir los datos incompletos, para no introducir en el modelo factores externos que puedan confundir al modelo y afectar a su capacidad predictiva. En consecuencia, se eliminaron del modelo 2.705 filas, dejando el total en 7.395.

Sin embargo, aunque entre los datos sobre la industria del profesor Damodaran también existían algunos valores ausentes, en este caso se ha procedido de manera diferente. De un total de 1.410 celdas, sólo 10 carecían de información. Se ha considerado que un impacto de este calibre sería reducido y, por ello, se ha optado por inferir los datos repitiendo el último disponible.

3.2.3 Cálculo de variables compuestas

En esta fase se han empleado algunas variables presentes en el *dataset* para calcular otras que se han denominado «compuestas». Estas nuevas variables tienen un carácter más informativo y representan una fuente de información más interesante para el modelo, puesto que ya desde su concepción establecen una comparación entre métricas de cada compañía con respecto a su industria. Estas variables son:

Earnings forecast

Esta variable se ha definido como el incremento que Bloomberg estima sobre los beneficios de la empresa durante los próximos 12 meses (*BEst EPS Blended Forward 12M*), en comparación con los últimos 12 (*EPS Trailing 12M*).

$$\text{Earnings forecast} = \frac{(\text{Best EPS Blended Forward 12M} - \text{EPS Trailing 12M})}{\text{EPS Trailing 12M}}$$

ROE compared to industry

Esta variable compara el ROE reportado por la empresa durante el último trimestre con el ROE medio de la industria. En base al cálculo, aquellas empresas que obtengan un valor superior a la unidad tendrán un ROE superior a la media, mientras que las que obtengan un valor inferior a la unidad se situarán por debajo de la media.

$$\text{ROE compared to industry} = \frac{\text{ROE: Q}}{\text{Industry average ROE}}$$

P/E compared to industry

De manera similar al caso anterior, esta variable compara la ratio precio-beneficio (P/E) de la empresa con el de su industria en el instante en el que se han obtenido los datos. La ratio precio-beneficio indica cuántas veces está cotizando una empresa por encima de su beneficio por acción. Tras el cálculo, un valor superior a uno nos indicará que la empresa tiene una ratio superior a la media de su sector. Teniendo en cuenta otros valores como su tamaño o su industria, entre otros, un valor muy elevado podría sugerir que la empresa estudiada está «sobrevalorada», lo cual podría ser negativo en empresas maduras, pero positivo en aquellas en plena fase de crecimiento.

$$\text{P/E compared to industry} = \frac{\text{P/E}}{\text{Industry average P/E}}$$

3.2.4 Normalización

Debido a los diferentes órdenes de magnitud de las variables independientes, se ha procedido a su normalización para mejorar el rendimiento del modelo de predicción. Este tipo de transformación es una práctica común en un conjunto de datos para su uso en procesos de *machine learning* (Huang, Capretz, & Ho, 2019). En este caso, se ha optado por la normalización mínimo-máximo, que permite escalar todos los datos a un rango establecido entre cero y uno. La fórmula empleada es la siguiente:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Donde X es el valor sin normalizar, X_{min} el valor mínimo, X_{max} el máximo y X' el valor normalizado. Debido a las características de este tipo de normalización, la existencia de valores extremos (*outliers*) puede provocar que las variables normalizadas no sean informativas. Puesto que para este TFG se ha decidido establecer los valores con un máximo de tres decimales, una vez realizada la normalización dos variables presentaban valores poco informativos, como se puede ver en la Tabla 2 a través de su mediana normalizada:

Tabla 2

Valores antes de realizar ajustes

Variable	<i>Change in EV/EBITDA: Q</i>	<i>Earnings Forecast</i>
Máximo sin normalizar	2.551,901	156.428.024,457
Mínimo sin normalizar	-98,048	-199,368
Media normalizada	0,040	0,000
Mediana normalizada	0,038	0,000

Para corregir esta situación, se han calculado tanto el percentil 1 como el percentil 99 de los valores sin normalizar. A continuación, se ha sustituido el 1% superior por el valor del percentil 99 en las dos variables. Además, en la variable *Earnings*

Forecast también se ha sustituido el 1% inferior por valor del percentil 1, tal y como se presenta a continuación:

Tabla 3

Valores ajustados

Variable	<i>Change in EV/EBITDA: Q</i>	<i>Earnings Forecast</i>
Máximo sin normalizar	135,569	12,001
Mínimo sin normalizar	-98,048	-9,777
Media normalizada	0,449	0,463
Mediana normalizada	0,434	0,458

El resultado de todos estos procesos es un *dataset* con 7.395 filas y 31 columnas. De ellas:

- Una columna cuenta con valores comprendidos entre el 1 y el 94, correspondiente con el código de industria.
- Un total de 29 columnas se encuentran normalizadas y todos sus valores están en el rango comprendido entre 0 y 1.
- Una columna, el desenlace u *outcome*, sólo tiene como posibles valores el 0 y el 1.

3.3 Aplicación de Azure Machine Learning

La herramienta que se ha empleado para elaborar el modelo es Azure Machine Learning. El proceso consta de 3 partes: diseño del diagrama, configuración y entrenamiento y validación.

3.3.1 Diseño del diagrama

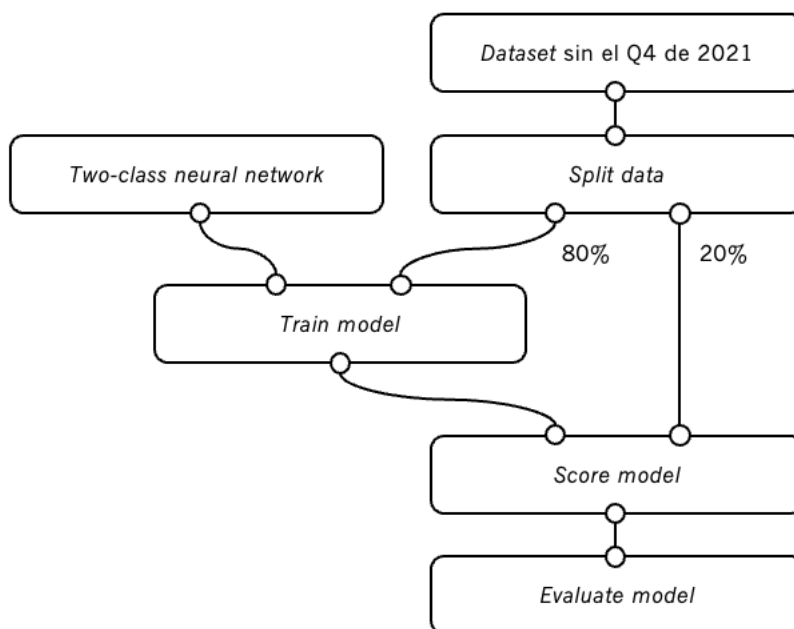
En primer lugar, se ha cargado en la plataforma el primer *dataset* construido en formato CSV. Se especifica que es un archivo en un formato delimitado por punto y coma.

El siguiente paso en el proceso de carga consiste en detallar el tipo de datos que contiene cada una de las 30 variables independientes de las que consta el modelo, así como de la variable dependiente. Esta información se muestra en la tabla que figura en el Anexo 2.

Una vez que el *dataset* se encuentra cargado en la plataforma se puede añadir a la estructura del diagrama. Este se crea en Azure Machine Learning colocando y uniendo distintos componentes sobre un lienzo en blanco, tal y como se puede ver a continuación:

Figura 2

Diagrama del modelo basado en el dataset sin el último trimestre de 2021



Nota. Adaptación del diagrama creado por la plataforma Azure Machine Learning.

Los componentes son:

- El conjunto de datos que contiene la información sobre la que se basará el entrenamiento del modelo. Puesto que la predicción se llevará a cabo sobre los datos del último trimestre de 2021, este *dataset* contendrá todos los trimestres entre el primero de 2017 y el tercero de 2021.
- Un componente de transformación, denominado *Split data*, que divide el *dataset* dando lugar a un conjunto de entrenamiento y otro de validación. Se ha establecido una división 80-20, de manera que el 80% de las filas,

elegidas aleatoriamente, se han usado para el entrenamiento del modelo y el 20% restante para su validación.

- El componente de redes neuronales, que se explicará en la siguiente sección.
- *Train model* y *Score model*, dedicados a llevar a cabo el entrenamiento del modelo usando el conjunto de entrenamiento y la validación usando el conjunto de validación previamente creados.
- El *output* del componente anterior se conectará con el componente *Evaluate model*, que llevará a cabo una evaluación de la precisión del modelo a partir de distintas métricas.

Elección del componente de redes neuronales

Se ha optado por utilizar el componente de Azure Machine Learning denominado *Two-class neural network*, que crea un modelo de redes neuronales que se puede utilizar para predecir un objetivo con únicamente dos valores. Se ha seleccionado este componente puesto que el desenlace propuesto sólo tiene dos soluciones posibles (1 = verdadero o 0 = falso).

Se está creando, por lo tanto, un método de aprendizaje supervisado que establece una clasificación basada en redes neuronales. Como se ha mencionado previamente, en el aprendizaje supervisado se «muestra» a la red neuronal una respuesta esperada, para comparar después dicha respuesta con la elaborada por la propia red. A continuación, se realizan ajustes para que ambas respuestas se acerquen cada vez más.

3.3.2 Configuración inicial

Se ha creado un clasificador de base con una arquitectura inicial sencilla, definida utilizando los valores más populares descritos en la literatura, que posteriormente se ha revisado con el fin de optimizar los parámetros que lo componen.

En la especificación inicial:

- El número de nodos en la capa de entrada es igual al número de variables en el conjunto de entrenamiento.
- Se han incluido dos capas ocultas con 12 y 6 nodos.

- El número de nodos en la capa de salida equivale al número de clases que puede tener el desenlace. Puesto que en este caso son dos, la capa de salida contará con dos nodos.

Por lo tanto, la capa de entrada contará con 30 nodos y la de salida con dos.

Para asegurar la capacidad de generalización del modelo y evitar el sobreajuste, el número de nodos en las capas ocultas no debe ser excesivamente elevado. Si este exceso existe, la red neuronal se convertiría en un banco de memoria capaz de recordar el conjunto de entrenamiento a la perfección, pero no realizaría bien predicciones sobre muestras que no formen parte del conjunto de entrenamiento.

La tasa de aprendizaje (*learning rate*) es un parámetro que controla la rapidez con la que el modelo «se adapta» al problema. Suele tener un valor positivo, a menudo en el rango entre 0,0 y 1,0. Si se usa una tasa de aprendizaje pequeña, los cambios en cada actualización son pequeños y el modelo tardará más tiempo en llegar a la solución óptima, mientras que si esta es grande (0,2-0,3) puede dar lugar a cambios más rápidos, y el modelo convergerá antes (Brownlee, 2019). Debido a la complejidad del modelo, se ha elegido una tasa de 0,1 para evitar una convergencia demasiado rápida en una solución subóptima.

Se ha establecido un máximo de 200 iteraciones y un valor de 0,9 para el peso que se aplica durante el proceso de aprendizaje a las iteraciones previas (Karim, 2018).

3.3.3 Entrenamiento y validación

Tal y como se ha descrito previamente, se ha llevado a cabo una división aleatoria del conjunto de datos en dos subconjuntos: entrenamiento y validación. Esta división aleatoria de las filas que componen el conjunto de datos, si bien evita «premiar» el sobreajuste, presenta un inconveniente: el resultado que alcanzará el modelo dependerá de cómo se realice la partición de los datos. Dos modelos basados en el mismo conjunto de datos, pero entrenados sobre distintos subconjuntos pueden dar lugar a resultados diferentes (Águila, 2017).

Esto ocurre porque, durante la fase de entrenamiento, el algoritmo trata de relacionar un grupo de variables de salida con otro grupo de variables de entrada mediante un sistema de aprendizaje supervisado. Si las filas que se usan para «informar» al modelo son diferentes en un caso u otro, durante la fase de validación

el algoritmo predecirá en base a estas relaciones establecidas en la fase anterior un desenlace distinto.

4. Resultados

4.1 Evaluación

Una vez concluidos el diseño del modelo de clasificación y su puesta en funcionamiento, es necesario analizar si el modelo es lo suficientemente válido para el fin para el que ha sido construido.

Para ello se han empleado dos métricas clásicas: la evaluación de la precisión y el análisis de la curva ROC.

4.1.1 Evaluación de la precisión

La primera métrica de evaluación que se abordará está basada en el porcentaje de error en la clasificación predicha por el modelo con respecto al desenlace real.

Para poder entender este método, primero es necesario explicar que la aplicación de un clasificador a un conjunto de casos puede producir cuatro resultados diferentes (Águila, 2017):

- Si el desenlace es positivo y se clasifica como positivo por el modelo, se denomina verdadero positivo (TP, del inglés *True Positive*).
- Si el desenlace es negativo y se clasifica como positivo se denomina falso positivo (FP, del inglés *False Positive*), también conocido como error de tipo I.
- Si el desenlace es negativo y se clasifica como negativo se denomina verdadero negativo (TN, del inglés *True Negative*).
- Si el desenlace es positivo y se clasifica como negativo se denomina falso negativo (FN, del inglés *False Negative*), también llamado error de tipo II.

Empleando las cuatro denominaciones mencionadas es posible construir una tabla de contingencia, que representará cómo se distribuyen los desenlaces en base a su clasificación. Puesto que en este caso se tratará de una predicción con dos desenlaces (verdadero o falso), se ha adaptado la tabla teniendo esto en cuenta, tal y como se representa a continuación:

Tabla 4

Tabla de contingencia

		Desenlace	
		Verdadero	Falso
Predicción	Verdadero	TP	FP
	Falso	FN	TN

La precisión (*accuracy*, en inglés) será el porcentaje de casos clasificados correctamente, y su fórmula será:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

4.1.2 Análisis de la curva ROC

Una alternativa a la evaluación basada en la medida de la precisión es el uso de un análisis gráfico, pues permite una mejor visualización de la multidimensionalidad del problema que estamos evaluando.

En el análisis de la curva ROC (del inglés *Receiver Operating Characteristic*) se emplean los valores normalizados de la tabla de contingencia para construir un gráfico. En el eje de ordenadas se representa la tasa de verdaderos positivos (TPR) y en el de abscisas la tasa de falsos positivos (FPR).

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN}$$

El espacio creado se denomina espacio ROC, y en él cada modelo de clasificación se representa con un punto (Prati, Monard, & Batista, 2008).

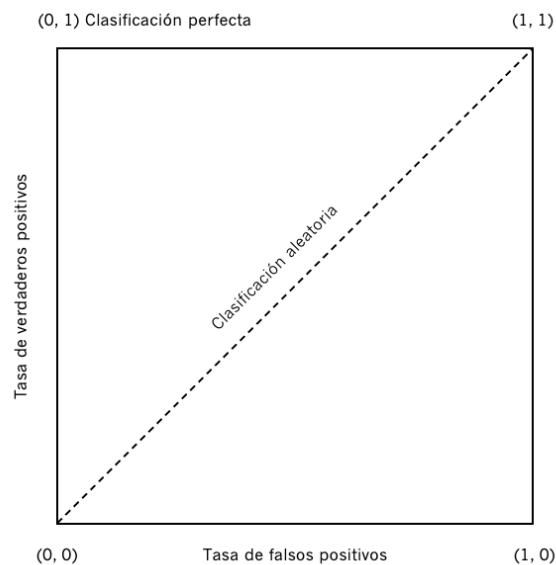
Los vértices de este espacio representan distintos clasificadores:

- El vértice inferior izquierdo, correspondiente a las coordenadas (0, 0), representa la estrategia de nunca clasificar algo como positivo. No habrá falsos positivos, pero tampoco verdaderos positivos.

- El vértice superior derecho, correspondiente a las coordenadas (1, 1), representa la estrategia inversa.
- El vértice inferior derecho (1, 0), representaría un modelo que siempre hace predicciones erróneas.
- El vértice superior izquierdo (0, 1) representa la clasificación perfecta.

Figura 3

Espacio ROC



La diagonal que discurre entre el punto (0, 0) y el punto (1, 1) y divide el espacio ROC en dos triángulos representa la clasificación aleatoria. Los modelos que se sitúen en el triángulo localizado a la izquierda de esta línea serán clasificadores útiles y, por lo tanto, se buscará que el modelo se sitúe en dicho espacio.

4.2 Resultado sobre el *dataset* sin el último trimestre de 2021

Como se ha mencionado previamente, en primer lugar, se ha tratado de valorar la capacidad de predicción del modelo sobre un *dataset* con los datos de los trimestres comprendidos entre el primero de 2017 y el tercero de 2021. El 80% de los datos se emplean para entrenar el modelo y el 20% restante para realizar predicciones y valorarlas. Empleando la configuración inicial descrita en el punto 3.3.2, la precisión (*accuracy*) del modelo alcanza un valor máximo del 73% cuando el umbral se ajusta

a una puntuación de 0,52. Esto significa que, cuando el valor calculado por el modelo es inferior a 0,52, este clasifica el desenlace como negativo (0, falso) y, cuando el valor calculado por el modelo es igual o superior a 0,52, el modelo clasifica el desenlace como positivo (1, verdadero).

Tabla 5

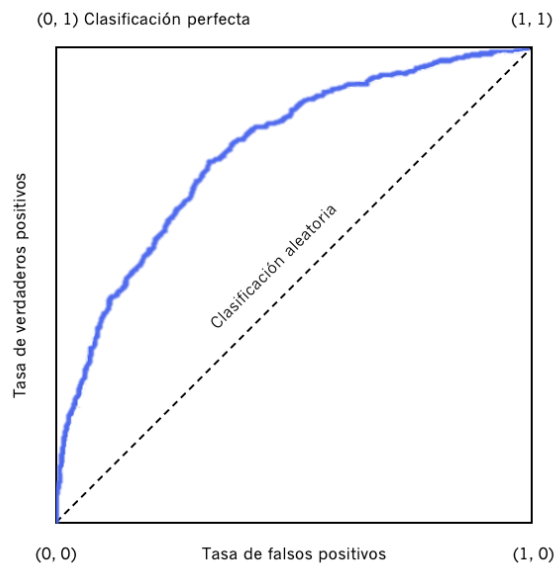
Tabla de contingencia del modelo basado en el dataset sin el último trimestre de 2021 maximizando la precisión

		Desenlace	
		Verdadero	Falso
Predicción	Verdadero	691	243
	Falso	137	335

El valor de la precisión se calcula a través de la fórmula descrita en el apartado 4.1.1 utilizando los valores de la Tabla 5.

Figura 4

Curva ROC del modelo basado en el dataset sin el último trimestre de 2021



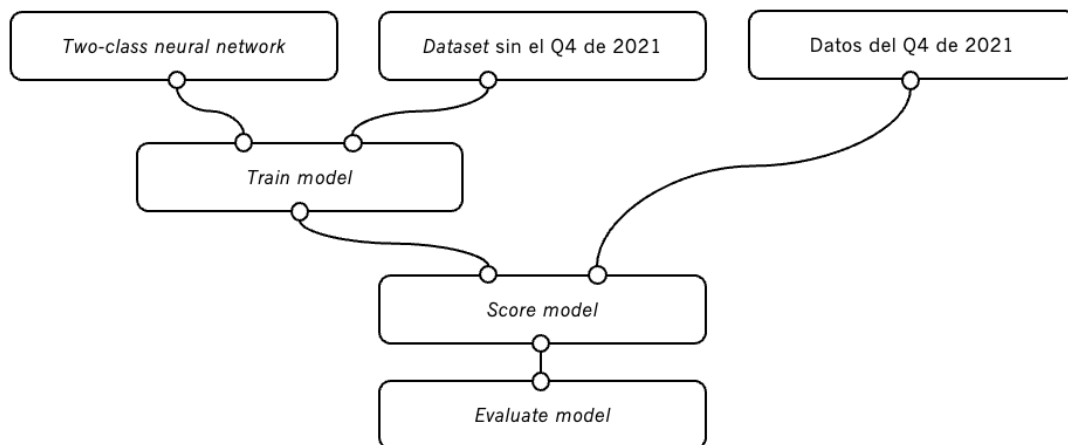
Por otra parte, el análisis en el espacio ROC (Figura 4) muestra una curva que se sitúa en el triángulo de la izquierda, lo que representa un clasificador útil, pues mejora aquella clasificación llevada a cabo si se asignaran valores al azar.

4.3 Predicción

Una vez validada la estructura del modelo se ha procedido a aplicarlo sobre un *dataset* diferente; aquel que contiene datos sobre el último trimestre de 2021. En el caso descrito anteriormente, se había dividido el conjunto de datos en dos subconjuntos de forma aleatoria, por lo que parte de los datos empleados durante el proceso de entrenamiento del modelo coincidían en el tiempo con aquellos sobre los que se iba a llevar a cabo la predicción. Es decir, como la división era aleatoria, a la hora de predecir si un valor subiría o no, el modelo podía tomar como referencia otros valores en el mismo momento del tiempo. Este factor no se ajusta a la realidad, puesto que, si un inversor quiere saber si el precio de una acción de Apple, Inc. subirá durante el próximo mes, no dispone en el momento de tomar la decisión de información sobre si el valor de otro componente del índice, por ejemplo, Microsoft, ha aumentado su cotización durante ese mismo mes, ya que el desenlace todavía no ha ocurrido.

Figura 5

Estructura del modelo de predicción sobre los datos del último trimestre de 2021



Por ello, el siguiente paso ha sido aplicar el modelo entrenado en la totalidad del conjunto de datos previo a un nuevo instante temporal: el último trimestre de 2021, tal y como se puede ver en la Figura 5.

En este caso el modelo utiliza datos históricos para entrenarse y realiza predicciones hacia el futuro. Esta situación sí se corresponde con la que rodea a un inversor en el momento de tomar sus decisiones, ya que sólo dispondrá de los datos históricos sobre ese valor. Por ejemplo, si ese inversor está decidiendo si comprar

o no acciones de Apple, Inc., sólo dispondrá de datos históricos sobre la cotización de ese valor y del resto de los miembros del índice.

4.3.1 Resultados

Usando el mismo umbral que en el caso anterior, la precisión de las predicciones del modelo creado es del 68%, tal y como se puede calcular a partir de los datos de la Tabla 6.

Tabla 6

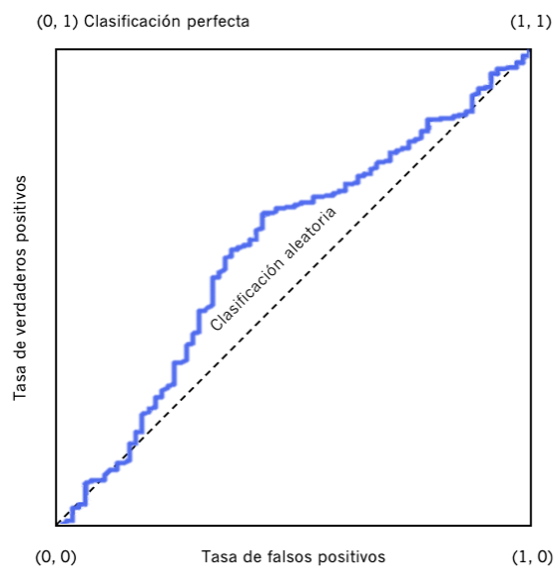
Tabla de contingencia del modelo basado en los datos del último trimestre de 2021 maximizando la precisión

		Desenlace	
		Verdadero	Falso
Predicción	Verdadero	228	54
	Falso	61	20

Sin embargo, en el espacio ROC se dibuja una curva que de nuevo se sitúa en el triángulo izquierdo, por lo que se trata de un clasificador útil.

Figura 6

Curva ROC del modelo de predicción sobre los datos del último trimestre de 2021



4.4 Optimización del modelo para inversión

Debido a que el objetivo de este trabajo es la creación de un modelo capaz de discernir qué valores van a aumentar su cotización en bolsa, para optimizar el modelo se ha tratado de minimizar el error de tipo I. Tal y como se ha descrito en la sección 4.1.1, el error de tipo I se produce cuando un desenlace negativo se clasifica como positivo. Pero... ¿por qué se ha elegido este tipo de optimización?

A la hora de realizar una estrategia de inversión en bolsa basada en los valores clasificados como positivos se puede proceder de diferentes formas. Debido a que el modelo tal y como está diseñado sólo especifica si un valor va a subir, pero no cuánto va a subir, no es posible asignar una mayor cantidad de dinero al valor con la mayor subida estimada. La estrategia más sencilla consistiría, por lo tanto, en dividir la cantidad de dinero que se desea invertir de manera equitativa entre todos los valores sobre los que el modelo ha predicho que van a incrementar su cotización. Para que esta estrategia funcione es fundamental que los valores clasificados como positivos sean efectivamente positivos. Sin embargo, las medidas previamente descritas como la precisión o la curva ROC representan más dimensiones del modelo de las necesarias para lograr este fin. Una optimización basada en estas métricas, si bien permite valorar hasta qué punto es el modelo capaz de predecir los diferentes desenlaces, no tiene por qué traducirse en la mejor estrategia de inversión.

Así, en el modelo optimizado explicado en el epígrafe 4.2, en el que se buscaba maximizar la precisión (*accuracy*), se ha tratado de que el porcentaje que representa la suma de verdaderos positivos y verdaderos negativos sobre el total de desenlaces sea máxima. Si se desarrollase una estrategia de inversión basada en dicha métrica, al invertir en los 282 valores clasificados como positivos por el modelo, en 54 casos la cotización del valor habría descendido y, por lo tanto, se habría perdido dinero.

Sin embargo, minimizando el error de tipo I esto no sucede, pues se reduce al máximo la cantidad de falsos positivos. Pese a que en este caso la cantidad de valores clasificados como positivos es menor, el número de falsos positivos se reduce notablemente. Una estrategia de inversión basada en esta última opción asignará a cada uno de los valores una cantidad de dinero y sólo en un porcentaje reducido de casos el valor disminuirá, lo que en comparación con el caso anterior se traduce en un mayor rendimiento económico.

4.4.1 Ajuste

Ajustando el umbral hasta 0,94 con la intención de minimizar el error de tipo I en el modelo inicial, este es capaz de conseguir un 100% de acierto en las predicciones sobre los valores clasificados como positivos dentro del conjunto de datos utilizado. Los resultados se presentan a continuación:

Tabla 7

Tabla de contingencia del modelo basado en el dataset sin el último trimestre de 2021 minimizando el error de tipo I

		Desenlace	
		Verdadero	Falso
Predicción	Verdadero	37	0
	Falso	791	578

Sin embargo, un umbral tan alto pese a conseguir un rendimiento elevado únicamente es capaz de identificar el 2,6% de los datos verdaderos, lo cual en este caso proporciona 37 resultados. Puesto que el *dataset* sobre el que se harán las predicciones tiene una cantidad más reducida de valores (363), se ha considerado que usar dicho umbral podría arrojar un número muy pequeño de resultados clasificados como verdaderos, por lo que se ha reducido un poco el umbral, de 0,94 a 0,92, de forma que el número de valores clasificados como positivo fuese mayor. En este nuevo caso, la tabla de contingencia del modelo 1 cuenta con nuevos valores y, si bien la tasa de verdaderos positivos se reduce del 100% al 95,6%, el número de valores clasificados como positivos se eleva hasta 90 (6,5%).

Tabla 8

Tabla de contingencia del modelo basado en el dataset sin el último trimestre de 2021 con el umbral situado en 0,92

		Desenlace	
		Verdadero	Falso
Predicción	Verdadero	86	4
	Falso	742	574

En consecuencia, será ese umbral de 0,92 el que se empleará cuando el modelo entrenado se testeé contra los datos del último trimestre del año para valorar su capacidad predictiva.

4.4.2 Aplicación al modelo de predicción

El ajuste del umbral en el modelo de predicción acorde con el razonamiento explicado en el apartado 4.4.1 muestra que de un total de 11 veces en las que el modelo ha predicho que el valor va a ser verdadero (es decir, que la acción subirá de valor), en 9 ocasiones este ha sido el desenlace real, lo que equivale a una tasa de verdaderos positivos del 81,8%.

Tabla 9

Tabla de contingencia del modelo basado en los datos del último trimestre de 2021 con el umbral situado en 0,92

		Desenlace	
		Verdadero	Falso
Predicción	Verdadero	9	2
	Falso	280	72

Pese a que con este umbral el modelo sólo tiene una precisión del 23%, al exigir que este valor sea más alto también se están seleccionando aquellas acciones que el modelo ha calculado que con mayor seguridad deberían tener un desenlace positivo. Este tipo de optimización provoca que la precisión del modelo se haya visto mermada hasta ser de únicamente un 22,3%, pero a la hora de llevar a cabo una estrategia de inversión muestra un mayor rendimiento.

4.5 Comparación de estrategias de inversión

Para valorar correctamente una estrategia de inversión no es suficiente con seleccionar un conjunto de valores que proporcionen un rendimiento positivo. Los inversores se fijan en los índices de referencia, como el S&P 500, para comparar su rendimiento con el de sus propias estrategias de inversión.

Puesto que existen las denominadas estrategias «pasivas» de inversión, en las cuales los inversores invierten en un índice sin necesidad de realizar análisis individuales de ningún tipo sobre las acciones, se ha establecido que este debe ser el punto de referencia sobre el que se debe establecer la comparativa de nuestro modelo. En concreto, se comparará con una estrategia de comprar y mantener (*buy and hold*) un fondo cotizado del S&P 500.

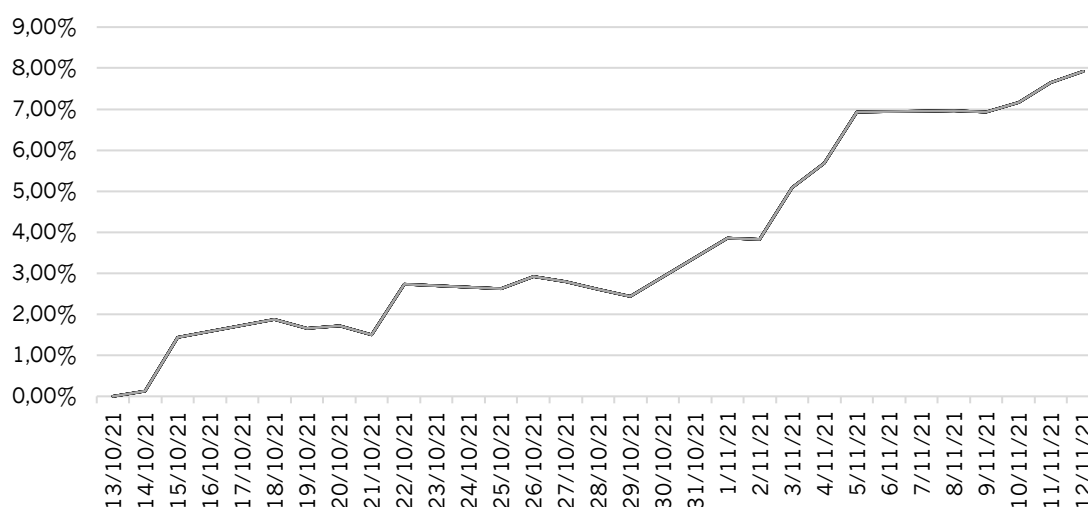
4.5.1 Estrategia de comprar y mantener

El fondo cotizado que se ha seleccionado para establecer la comparativa es el SPDR S&P 500 ETF (SPY), uno de los que cuentan con un mayor volumen de negocio y que replica completamente el índice S&P 500, manteniendo todos los miembros del índice subyacente en sus ponderaciones objetivo (Nickolas, 2022).

Tal y como se puede observar en la Figura 7, entre el 13 de octubre de 2021 y el 12 de noviembre del 2021, el SPY creció un 7,9%. El índice al que replica, el S&P 500, alcanzó su máximo histórico durante el mes de noviembre de 2021, catapultado por un escenario macroeconómico caracterizado por bajos tipos de interés. Así, si se hubiese invertido 1€ en el índice al principio del periodo y asumiendo que no existen otros costes que puedan afectar a la rentabilidad, al final del periodo dispondríamos de casi 1,08€.

Figura 7

Evolución de la cotización del SPY entre el 13 de octubre y el 12 de noviembre



4.5.2 Estrategia basada en el modelo de predicción

El modelo de predicción diseñado, tal y como se explicó en el apartado 4.4.2, clasificó como verdaderos 11 valores, que se muestran en la Tabla 9. De ellos, 2 resultaron ser falsos positivos.

Tabla 9

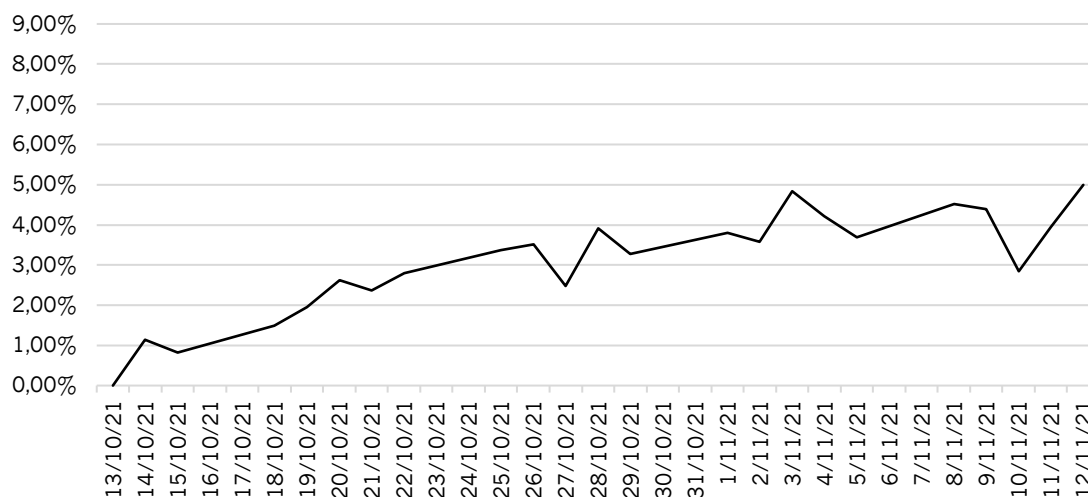
Valores seleccionados por el modelo

Compañía	Desenlace predicho	Desenlace real	Rendimiento
Coterra Energy, Inc.	Verdadero	Falso	-1,44%
Thermo Fisher Scientific, Inc.	Verdadero	Verdadero	10,76%
CF Industries Holdings, Inc.	Verdadero	Verdadero	6,99%
The Mosaic Company	Verdadero	Falso	-8,20%
Netflix, Inc.	Verdadero	Verdadero	8,39%
Diamondback Energy, Inc.	Verdadero	Verdadero	1,63%
Eastman Chemical Company	Verdadero	Verdadero	11,60%
General Mills, Inc.	Verdadero	Verdadero	2,23%
Apple, Inc.	Verdadero	Verdadero	6,44%
Microsoft Corporation	Verdadero	Verdadero	13,64%
The Procter & Gamble Company	Verdadero	Verdadero	2,89%
Total			4,99%

En el caso de haber llevado a cabo una estrategia de inversión que asignase el mismo peso a cada uno de los valores, el rendimiento acumulado durante el mes habría sido del 4,99% (Figura 8).

Figura 8

Evolución del rendimiento de la estrategia de inversión basada en el modelo



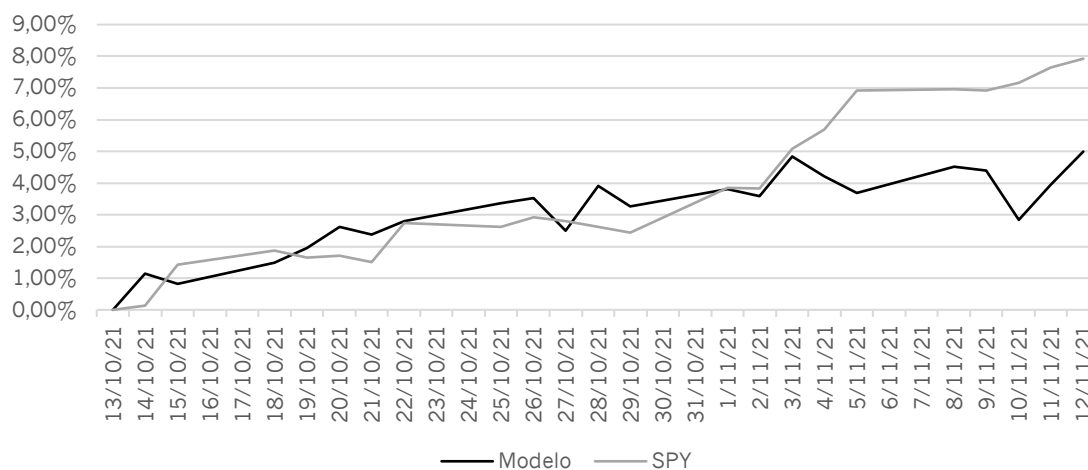
4.5.3 Comparativa de rendimiento

Si se combinan ambos gráficos, tal y como se puede ver en la Figura 9, se observa cómo durante el mes de octubre de 2021 la estrategia basada en el modelo de inversión alcanzó unos rendimientos ligeramente superiores a los del SPY hasta el inicio de noviembre, momento en el que noticias concretas sobre determinados valores provocaron una caída de su cotización que afectó a la actuación global del modelo. En concreto:

- Las acciones de Netflix bajaron un 5,2% entre los días 1 y 5 de noviembre debido a la mala recepción del mercado al lanzamiento de un sistema de videojuegos incluido en su plataforma de *streaming* para dispositivos con Android (Golum, 2021).
- La presentación de resultados trimestrales de The Mosaic Company resultó en una caída que alcanzó un máximo cercano al 10% el día 2 de noviembre, pese a haber anunciado un crecimiento en ingresos del 44% respecto al mismo periodo del año anterior (Chamaria, 2021).

Figura 9

Comparación del rendimiento de las dos estrategias de inversión



Sólo cuatro de los valores seleccionados por el modelo superaron el rendimiento ofrecido por el SPY, cinco obtuvieron un rendimiento positivo pero inferior al índice y dos obtuvieron un rendimiento negativo.

Las carteras de valores soportan dos tipos principales de riesgo: el riesgo sistemático, provocado por las interdependencias entre los factores económicos, monetarios, políticos y sociales, y el no sistemático, que afecta de forma específica

a cada empresa y se reduce diversificando la inversión en un mayor número de valores (Titman & Martin, 2016). En este caso, ambas estrategias de inversión presentan un riesgo sistemático parecido, pero debido a que el modelo sólo ha seleccionado 11 valores, frente a los 500 que componen el S&P 500, el riesgo no sistemático que soporta esta cartera de valores es muy elevado, muestra de ello es el impacto que han tenido los dos hechos comentados previamente sobre el rendimiento global de la cartera.

5. Discusión

Por definición, la construcción de este modelo de predicción representa un desafío a la hipótesis de los mercados eficientes. Si bien es argumentable que, en base al perfil de datos utilizados en el primer modelo, que permitían que este se entrenase usando información de la que no dispondría en una situación real, el desafío es cuestionable, el hecho de que sobre un conjunto nuevo de datos el modelo haya sido capaz de llevar a cabo predicciones útiles atenta directamente contra una teoría que defiende que el aumento o reducción del valor de las cotizaciones ocurre de forma aleatoria. Pese a ser un modelo con una precisión reducida, en este caso del 68% cuando el valor del umbral es 0,52, la posibilidad de escoger aquellos valores a los que el modelo asigna mayores probabilidades de subir, tal y como se hizo elevando el umbral para reducir los falsos positivos, es suficiente para desacreditar la teoría enunciada por Fama en 1965. Si un modelo es capaz de asegurar que un número determinado de valores, por pequeño que sea, aumentarán su cotización, entonces su movimiento no podría ser aleatorio.

Asimismo, conviene destacar que, como se mencionó en el apartado 4.5.1, la bolsa durante el periodo estudiado atravesó una etapa de ascenso muy importante, en la que la mayoría de los valores del S&P 500 ganaron valor. Como consecuencia del impacto causado por la pandemia de COVID-19 en la economía, en Estados Unidos se llevaron a cabo políticas públicas de ayuda a los ciudadanos que, unidas al recorte de los tipos de interés permitieron inundar de dinero el mercado, factores que durante los últimos meses de 2020 y hasta finales de 2021 promovieron un crecimiento significativo de la bolsa. Este contexto añade dificultad a las predicciones que debe llevar a cabo el modelo, pues en su mayor parte los datos empleados para su entrenamiento provienen de una situación macroeconómica diferente. En concreto, durante este periodo cerca del 80% de los valores analizados incrementaron su cotización, lo cual sólo ocurre en momentos de intenso crecimiento. Debido a las circunstancias mencionadas previamente, que también provocaron el auge de los inversores particulares, muchos de ellos sin experiencia previa, hasta representar un tercio de las operaciones del mercado, es posible que se haya desencadenado una divergencia de los criterios de inversión seguidos en épocas previas que el modelo no habría sido capaz de captar (Reddel, 2022).

Otra consecuencia de este periodo de bonanza en la bolsa es que el modelo no ha podido testarse en una situación de crisis económica. Siguiendo la misma filosofía

de una optimización basada en la reducción del error de tipo I, si el modelo fuese capaz de identificar aquellas características de una empresa que en un contexto macroeconómico desfavorable la hacen más proclive a conservar su valor, podría ser de utilidad para diseñar estrategias que protejan a los inversores ante estas tesituras.

De igual forma, existen dudas sobre la capacidad del modelo para llevar a cabo una estrategia de inversión exitosa. El problema de base es claro: con el diseño actual no es posible cuantificar cuánto se prevé que crezca la cotización de la acción, sino que sólo se ofrece información sobre si subirá o no. Si un inversor pudiese conocer cuánto se incrementará el valor de la acción, a la hora de construir su cartera priorizaría la inversión en aquellas que experimentarían un crecimiento más alto. El problema del desenlace binario reside en que el inversor no dispone de esta información y, por lo tanto, la forma de actuar implica repartir equitativamente los recursos entre todos los valores seleccionados por el algoritmo. Si bien el modelo se ha construido de esta forma por razones de simplicidad, explorar un desenlace que no fuese binario ofrecería mayor utilidad.

La consecuencia de esta estructura es que, si bien ha habido cuatro valores (Netflix, Inc., Thermo Fisher Scientific, Inc., Microsoft Corporation y Eastman Chemical Company) que han superado el rendimiento del SPY, la distribución de los recursos en partes iguales ha mermado su rendimiento. Un modelo de predicción que hubiese ofrecido un mayor valor esperado de subida habría permitido focalizar las inversiones en esos cuatro valores e hipotéticamente se habría podido superar el rendimiento ofrecido por el fondo cotizado.

Del mismo modo, es importante tener en cuenta que el crecimiento experimentado por el S&P 500 en un periodo de tan solo un mes de duración fue cercano al 8%, un valor muy elevado. Si bien el modelo podría haber seleccionado mejor aquellos valores que fuesen a experimentar el mayor crecimiento del índice, conviene recordar que:

- Debido a que no todos los valores disponían de información concreta, algunos fueron eliminados durante el procesamiento de los datos llevado a cabo en las fases iniciales. Existe la posibilidad de que si se hubiesen incluido el modelo los hubiese seleccionado, por lo que no es posible descartar que si se dispusiese de mejores datos de entrada el modelo habría sido capaz de generar mayor rendimiento.

- El periodo estudiado se enmarca en un contexto de circunstancias excepcionales sobre las que, de nuevo, los datos de entrenamiento son reducidos al no ser una situación común a lo largo de la historia. Por ello, es posible que en base a la información histórica disponible no fuese posible esperar un crecimiento tan elevado de algunos valores. Este sobreajuste, mencionado en el epígrafe 1.2 como una de las principales limitaciones de este trabajo, implica que la capacidad del modelo para generalizar y adaptarse a todo tipo de contextos económicos es limitada, pero no descarta un mejor funcionamiento en circunstancias menos excepcionales.

Además, debido a que no se han realizado análisis concretos para medir la importancia de las variables en el modelo, se desconoce cuáles son las que tienen más relevancia a la hora de llevar a cabo la predicción, así como cuáles son irrelevantes. Abordar esta cuestión permitiría sustituir aquella información con menos aporte por nuevas variables cuya relación con el desenlace se podría testar, lo que debería de traducirse en un mayor rendimiento del modelo al entrenarse sobre datos mejor seleccionados. Este es un punto sobre el que no se ha podido profundizar debido a la limitación en espacio de este trabajo.

6. Conclusiones

Los objetivos de este Trabajo Fin de Grado eran dos: construir un modelo capaz de predecir si la cotización de un valor aumentaría o no en un periodo de 30 días y valorar si una estrategia de inversión basada en dicho modelo superaría o no al rendimiento del S&P 500.

Pese a la dificultad que implica la realización de predicciones en el área de las finanzas, debido a la inmensa cantidad de factores que pueden afectar a la cotización de un valor, se puede afirmar que existe un componente de predictibilidad que el modelo ha sido capaz de abordar en mayor o menor medida. En primer lugar, si los comportamientos en bolsa de las acciones fuesen aleatorios, el modelo, cuando se entrenó y testó sobre el mismo conjunto de datos en una de las primeras fases, no debería de haber sido capaz de alcanzar un 73% de precisión. El hecho de que se redujese esta cifra en el paso posterior al verse limitada la información usada para llevar a cabo las predicciones no habla tampoco a favor de la hipótesis de los mercados eficientes, sino que refleja la necesidad de obtener mejores variables que sean capaces de anticipar con más claridad hechos relevantes relacionados con los precios de las materias primas o a nivel macroeconómico.

El factor más limitante para alcanzar este objetivo es la disponibilidad de los datos. Si bien al haberse elegido el S&P 500 como índice objetivo del trabajo los volúmenes de información disponible son relativamente altos, algunas variables como puntuaciones en el ámbito de la responsabilidad social corporativa o sobre el *rating* de las empresas no han podido ser incorporadas debido a que dicha información no estaba disponible en todas las compañías para su uso. De la misma manera, debido a que este Trabajo Fin de Grado es un proyecto no financiado, no existe la posibilidad de acceder a otras bases de datos de pago que podrían ofrecer información ya estructurada que fuese relevante, más allá del uso de la licencia de Bloomberg a nivel educativo.

El segundo objetivo era valorar si el modelo sería capaz de vencer al rendimiento del S&P 500, que se ha aproximado al del fondo cotizado SPY. La conclusión en este caso es que el modelo no es capaz de vencer al mercado tal y como está diseñado actualmente, pero no se puede afirmar que en el caso de que el desenlace

fuese no binario, ofreciendo una estimación del incremento del precio, este objetivo no se pudiese alcanzar.

Dentro de los objetivos de este trabajo también se enmarcaba realizar el modelo a través de una aproximación sencilla a las redes neuronales, «asequible para personas sin conocimientos técnicos elevados de programación». Si bien la construcción de este modelo requiere conocimientos previos sobre *machine learning* para poder enfocar de forma correcta la estructura del modelo y sobre finanzas para una correcta elección de variables, conviene reconocer que con el uso de la plataforma Azure no ha sido necesario escribir código en lenguaje informático, lo que permitiría a un estudiante de Administración y Dirección de Empresas sin conocimientos especializados abordar este tipo de tareas si contase simplemente con unos nociones teóricas básicas.

Asimismo, y pese a sus limitaciones, este trabajo puede ser un punto de partida para la creación de modelos de predicción sencillos basados en el uso de redes neuronales artificiales. Mediante el abordaje de los principales problemas descritos a lo largo de las secciones anteriores y un acceso a más datos más actualizados debería ser posible incrementar la precisión del modelo, así como reducir el error de tipo I en mayor medida.

7. Bibliografía

- Águila, J. (2017). Aprendizaje supervisado en conjuntos de datos no balanceados con redes neuronales artificiales. Métodos de mejora de rendimiento para modelos de clasificación binaria en diagnóstico médico. *Universitat Rovira i Virgili*.
- Arias, M. (11 de Noviembre de 2020). *Historia y orígenes de la bolsa*. Obtenido de Rankia: <https://www.rankia.com/blog/bolsa-desde-cero/3349076-historia-origenes-bolsa>
- Baker, S., Bloom, N., & Davis, S. (2015). Measuring economic policy uncertainty. *National Bureau of Economic Research*.
- Barberis, N., & Thaler, R. (2002). A survey of behavioral finance. *National Bureau of Economic Research*.
- Brownlee, J. (2019). *Understand the impact of learning rate on neural network performance*. Obtenido de Machine Learning Mastery: <https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/>
- Chamaria, N. (2 de noviembre de 2021). *Why Mosaic Stock Crashed Despite a 50% Dividend Hike*. Obtenido de The Motley Fool: <https://www.fool.com/investing/2021/11/02/why-mosaic-stock-crashed-despite-50-dividend-hike/>
- Chen, J. (5 de mayo de 2022). *Investopedia*. Obtenido de Nasdaq Composite Index: <https://www.investopedia.com/terms/n/nasdaqcompositeindex.asp>
- Ciner, C. (2018). Do industry returns predict the stock market? A reprise using the random forest. *The Quarterly Review of Economics and Finance*.
- Damodaran, A. (2022). *Data: History & sharing*. Obtenido de Damodaran Online: https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datahistory.html#history
- DeepAI. (2017). *What is Weight?* Obtenido de DeepAI: <https://deepai.org/machine-learning-glossary-and-terms/weight-artificial-neural-network>

- Dimitrova, D. (2005). The Relationship between Exchange Rates and Stock Prices: Studied in a Multivariate Model. *Issues in Political Economy*.
- Downey, L. (31 de diciembre de 2021). *Efficient Market Hypothesis* . Obtenido de Investopedia:
<https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp>
- Enke, D. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, 29(4), 927-940.
- Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*.
- Fang, B., & Zhang, P. (2016). *Big data concepts, theories, and applications*. Cham: Springer.
- Fitriyaningsih, I. (2019). Implementation of Artificial Neural Network to Predict S&P 500 Stock Closing Price. *Journal of Physics Conf. Series*.
- Golum, R. (2 de noviembre de 2021). *Netflix Launches Into Video Games for Android With Titles Including 'Stranger Things'*. Obtenido de Bloomberg:
<https://www.bloomberg.com/news/articles/2021-11-02/netflix-launches-foray-into-video-games-with-titles-for-android>
- Goncalves-Pinto, L., Grundy, B., Hameed, A., van der Heijden, T., & Zhu, Y. (2019). Why Do Option Prices Predict Stock Returns? The Role of Price Pressure in the Stock Market. *Management Science*, 1-24.
- Hayes, A. (9 de abril de 2022). *Enterprise Multiple*. Obtenido de Investopedia:
<https://www.investopedia.com/terms/e/ev-ebitda.asp>
- Hiemstra, Y. (1996). Linear regression versus backpropagation networks to predict quarterly stock market excess returns. *Computational Economics*(9), 67-76.
- Huang, Y., Capretz, L., & Ho, D. (2019). Neural network models for stock selection based on fundamental analysis. *Canadian Conference on Electrical & Computer Engineering*, (págs. 1-4). Edmonton.
- Jensen, M. (1978). Some anomalous evidence regarding market efficiency. *Journal of Financial Economics*, 95-101.
- Karim, R. (22 de Noviembre de 2018). *Stochastic gradient descent optimisation algorithms*. Obtenido de Towards Data Science:

<https://towardsdatascience.com/10-gradient-descent-optimisation-algorithms-86989510b5e9>

Kenton, W. (15 de febrero de 2022). *Investopedia*. Obtenido de The S&P 500 Index: <https://www.investopedia.com/terms/s/sp500.asp>

Kolarik, T., & Rudorfer, G. (1997). Time series forecasting using neural networks. *Vienna University of Economics and Business Administration*, 2-6.

Lawrence, R. (1997). Using neural networks to forecast stock market prices. *University of Manitoba*, 333, 2006-2013.

Luna, J. (8 de febrero de 2018). *Tipos de aprendizaje automático*. Obtenido de Medium: <https://medium.com/soldai/tipos-de-aprendizaje-automático-6413e3c615e2>

López, J. (29 de mayo de 2022). *Variable macroeconómica*. Obtenido de Economipedia: <https://economipedia.com/definiciones/variable-macroeconomica.html>

Nickolas, S. (3 de abril de 2022). *SPDR S&P 500 ETF Trust*. Obtenido de Investopedia: <https://www.investopedia.com/articles/investing/122215/spy-spdr-sp-500-trust-etf.asp>

Pizarroso, J., Portela, J., & Muñoz, A. (2021). Neursens: Sensitivity Analysis of Neural Networks. *Journal of Statistical Software*.

Pothen, A. (2021). Artificial Intelligence and its increasing importance. En J. Karthikeyan, & S. Hie, *Learning outcomes of classroom research*. Routledge: L'Ordine Nuovo Publication.

Prati, R., Monard, M., & Batista, G. (2008). Evaluating classifiers using ROC curves. *IEEE Latin America Transactions*, 6(2), 215-222.

Reddel, S. (6 de enero de 2022). *Three thoughts on the rise of retail investors*. Obtenido de Accenture: <https://capitalmarketsblog.accenture.com/three-thoughts-rise-retail-investors>

Schmueli, G., Bruce, P., Yahav, Y., Patel, N., & Lichtendahl, K. (2018). Neural Nets. En G. Schmueli, P. Bruce, Y. Yahav, N. Patel, & K. Lichtendahl, *Data Mining for Business Analytics: Concepts, Techniques and Applications in R* (págs. 271-292). Hoboken: John Wiley & Sons.

- Siddiqui, S. (2003). Can interest rate changes help predict future stock price movements? Evidence from the German market. *Applied Economics Letters*, 10(4), 209-211.
- Titman, S., & Martin, J. (2016). *Valuation. The art and science of corporate investment decisions*. Austin: Pearson.
- Van Eyden, R. (1996). *The Application of Neural Networks in the Forecasting of Share Prices*. Johannesburgo: Finance and Technology Publishing.
- Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of Management Information Systems*, 17(4), 203-222.
- Wolkers Kluwer. (18 de noviembre de 2021). *Los ratios financieros: cuáles son y cómo se calculan*. Obtenido de Wolkers Kluwer: https://www.wolkersklower.com/es-es/expert-insights/ratios-financieros-cuales-son-como-se-calculan#anchor_1637226030375
- Yao, J., & Tan, C. L. (2001). Guidelines for financial forecasting with neural networks. *International Conference on Neural Information Processing*. Shanghai.

Anexos

Anexo 1. Correspondencia entre los nombres de las variables

Nombre de la variable	Correspondencia en Bloomberg
<i>ROE: Q</i>	ROE: Q
<i>P/E</i>	P/E
<i>EV/EBITDA Adj</i>	EV/EBITDA Adj
<i>Change in EV/EBITDA: Q</i>	Gr PoP of EV/EBITDA Adj: Q
<i>Change in Net Debt: Q</i>	Gr PoP of Net Debt: Q
<i>EPS T12M</i>	EPS T12M
<i>Best EPS BF12M</i>	Best EPS BF12M
<i>Momentum</i>	Relative Share Price Momentum
<i>Recommendation</i>	Rec Consensus: D-1
<i>3-Month Beta</i>	Raw Beta: M-3
<i>Best Long-Term Growth</i>	Best Est Long Term Growth
<i>Market Cap</i>	Market Cap

Anexo 2. Variables incorporadas y tipo de dato

Variable	Tipo de dato
<i>Industry code</i>	Número entero
<i>P/E compared to industry</i>	Número decimal
<i>ROE compared to industry</i>	Número decimal
<i>Expected growth</i>	Número decimal
<i>Market cap</i>	Número decimal
<i>3-month beta</i>	Número decimal
<i>EV/EBITDA adjusted</i>	Número decimal
<i>Change in EV/EBITDA: Q</i>	Número decimal
<i>Change in net debt: Q</i>	Número decimal
<i>Earnings forecast</i>	Número decimal
<i>BEst long-term growth</i>	Número decimal
<i>Momentum</i>	Número decimal
<i>Recommendation</i>	Número decimal
<i>Inflation</i>	Número decimal
<i>Economic policy uncertainty</i>	Número decimal
<i>Economic policy uncertainty change (%)</i>	Número decimal
<i>Exchange rate</i>	Número decimal
<i>Exchange rate change (%) 3 M</i>	Número decimal
<i>10-year bond</i>	Número decimal
<i>10-year bond change (%) 3 M</i>	Número decimal
<i>3-month bond</i>	Número decimal

<i>3-month bond change (%) 3 M</i>	Número decimal
<i>Federal Funds rate change (%) 1 M</i>	Número decimal
<i>S&P 500 1-month change (%)</i>	Número decimal
<i>VIX</i>	Número decimal
<i>WTI price</i>	Número decimal
<i>WTI change (%) 30 D</i>	Número decimal
<i>Gold price change (%) 3 M</i>	Número decimal
<i>100-day stock price change (%)</i>	Número decimal
<i>30-day stock price change (%)</i>	Número decimal
<i>Outcome</i>	Booleano (1 = Verdadero, 0 = Falso)