Vigilada Mineducación

Aplicación de modelos de inteligencia artificial y aprendizaje automático para la previsión de precios y la optimización de portafolios: un enfoque integrado con datos estructurados y no estructurados con el fin de compararse con el S&P 500 como *benchmark*

Application of Artificial Intelligence and Machine Learning Models for Price Forecasting and Portfolio Optimization: An Integrated Approach with Structured and Unstructured Data for Comparison Against the S&P 500 as a Benchmark

SANTIAGO VÉLEZ GARCÍA

Proyecto de investigación

Asesor, docente:

MSc. Juan Carlos Botero Ramírez

UNIVERSIDAD EAFIT

ESCUELA DE ECONOMÍA Y FINANZAS

MAESTRÍA EN ADMINISTRACIÓN FINANCIERA - MAF

MEDELLÍN

2023

Contenido

Introducción ... *7*

Justificación ... *9*

Objetivos ... *11*

Objetivo general ... **11**

Objetivos específicos ... **11**

Marco conceptual ... *12*

Bolsa de valores ... **13**

Acciones ... **14**

Índices bursátiles ... **16**

Teoría de portafolios ... **17**

Teoría moderna de portafolios ... **18**

***Capital asset pricing model* (CAPM) y *security market line* (SML)** ... **20**

Tolerancia al riesgo ... **23**

Perfiles ... **25**

Medición de riesgos ... **26**

Medidas de desempeño ... 27

Data estructurada y no estructurada ... **29**

Retornos pasados ... 30

Producto interno bruto ... 31

Inflación ... 32

Desempleo ... 35

Tasas de interés ... 36

Confianza del consumidor ... 40

Índices de producción ... 41

Estados financieros ... 42

Data mining ... **44**

Obtención de datos ... 45

Procesamiento de datos ... 46

Modelación ... 47

Análisis de sentimientos ... **60**

Python ... **63**

X (antes Twitter) ... **64**

I. Metodología ... *65*

1. Selección de variables ... **66**

2. Validación de significancia de las variables seleccionadas ... **67**

3. Predicción de precios futuros ... **67**

4. Creación del portafolio óptimo 67

II. Cronograma 68

III. Resultados 68

Análisis y descripción de variables 70

Correlación dinámica 70
Análisis de colinealidad 84
Análisis de estacionalidad 88
Análisis de cointegración 95
Análisis de homocedasticidad y normalidad 98

Análisis de sentimientos 98

Predicción de precios 105

Apple 109
Microsoft 109
Tesla 110
Amazon 111
Google 112

Creación del portafolio 113

Conclusiones 122

Referencias 124

Lista de ilustraciones

Ilustración 1. Comportamiento de las empresas colombianas en la bolsa de valores de Wall Street 15
Ilustración 2. Diversificación 19
Ilustración 3. Capital market line 22
Ilustración 4. CML 22
Ilustración 5. Portafolios apalancados 23
Ilustración 6. Curva de indiferencia 24
Ilustración 7.Perfil de riesgo 26
Ilustración 8. Desviación estándar de un portafolio 26
Ilustración 9. Indicadores económicos 29
Ilustración 10. Análisis de tendencias 31
Ilustración 11. Ciclos de negocios 32
Ilustración 12. S&P 500 y recesiones 32
Ilustración 13. Equilibrio del PIB e inflación 33
Ilustración 14. Inflación y tasas de política monetaria 36
Ilustración 15. Índice de sentimientos del mercado 40
Ilustración 16. Índice de confianza de Estados Unidos 41
Ilustración 17. Índice PMI 41
Ilustración 18. Calidad de los estados financieros 43
Ilustración 19. Data mining 45
Ilustración 20. Representación de regresión logística 48

Ilustración 21. LDA 49
Ilustración 22. LSTM 53
Ilustración 23. LSTM para varios *t*. 55
Ilustración 24. RMSE 58
Ilustración 25. *Tweet* relacionado con Apple 64
Ilustración 26. Base de datos variables 68
Ilustración 27. Market cap 69
Ilustración 28. Escalas de coeficiente de correlación 70
Ilustración 29. Correlación dinámica entre precio y PE_ratio 71
Ilustración 30. Correlación dinámica entre precio y price to books 72
Ilustración 31. Correlación dinámica entre precio y price to CF ratio 72
Ilustración 32. Correlación dinámica entre precio y price to sales ratio 73
Ilustración 33. Correlación dinámica entre precio y ROA 73
Ilustración 34. Correlación dinámica entre precio y price to ROC 74
Ilustración 35. Correlación dinámica entre precio y price to ROE 74
Ilustración 36. Correlación dinámica entre precio y EBITDA 75
Ilustración 37. Correlación dinámica entre precio y gross margin 75
Ilustración 38. Correlación dinámica entre precio y g 76
Ilustración 39. Correlación dinámica entre precio y payout ratio 76
Ilustración 40. Correlación dinámica entre precio y quick ratio 77
Ilustración 41. Correlación dinámica entre precio y PE_current ratio 77
Ilustración 42. Correlación dinámica entre precio y CF 78
Ilustración 43. Correlación dinámica entre precio y retention rate 78
Ilustración 44. Correlación dinámica entre precio y debt-to-EBITDA 79
Ilustración 45. Correlación dinámica entre precio y S&P500 79
Ilustración 46. Correlación dinámica entre precio y deflactor del GDP 80
Ilustración 47. Correlación dinámica entre precio y GDP 80
Ilustración 48. Correlación dinámica entre precio y CPI 81
Ilustración 49. Correlación dinámica entre precio y PE_ratio 81
Ilustración 50. Correlación dinámica entre precio y FED rate 82
Ilustración 51. Correlación dinámica entre precio e índice de confianza 82
Ilustración 52. Colinealidad de Microsoft y demás variables 86
Ilustración 53. Colinealidad de Apple y demás variables 86
Ilustración 54. Colinealidad de Tesla y demás variables 86
Ilustración 55. Colinealidad de Amazon y demás variables 87
Ilustración 56. Colinealidad de Google y demás variables 87
Ilustración 58. Estacionalidad PE_ratio 88
Ilustración 59. Estacionalidad price to book 89
Ilustración 60. Estacionalidad price to CF 89
Ilustración 61. Estacionalidad ROA 90
Ilustración 62. Estacionalidad EBITDA 90
Ilustración 63. Estacionalidad gross margin 91
Ilustración 64. Estacionalidad dividend payout 91
Ilustración 65. Estacionalidad retention rate 92
Ilustración 66. Estacionalidad quick ratio 92
Ilustración 67. Estacionalidad current ratio 93
Ilustración 68. Estacionalidad CPI 93

Ilustración 69. Estacionalidad FED rate 94
Ilustración 70. Estacionalidad índice de confianza 94
Ilustración 71. *Tweets* 99
Ilustración 72. Frecuencia de frases x 100
Ilustración 73. Tabla de frecuencia de sentimientos 102
Ilustración 74. DF tweets con su sentimiento 102
Ilustración 75. Tweets positivos 103
Ilustración 76. Nube de *tweets* negativos 103
Ilustración 77. Nube de tweets positivos 103
Ilustración 78. Optimización de hiperparámetros 106
Ilustración 79. Optimización del modelo 107
Ilustración 80. Predicción de precios para Apple 109
Ilustración 81. Predicción de precios para Microsoft 109
Ilustración 82. Predicción de precios para Tesla 110
Ilustración 83. Predicción de precios para Amazon 111
Ilustración 84. Predicción de precios para Google 112
Ilustración 85. Precios principales de Acciones S&P500 114
Ilustración 86. Retornos principales de Acciones S&P500 114
Ilustración 87. Distribución de retornos principales de Acciones S&P500 115
Ilustración 88. Revisión de la normalidad de retornos 116
Ilustración 89. Frontera eficiente 119
Ilustración 90. Pesos de los diferentes activos 120
Ilustración 91. Performance 120
Ilustración 92. Retornos portafolio Vs S&P 500 120

Resumen

Este estudio presenta un enfoque integrado de modelos de inteligencia artificial y aprendizaje automático, combinando redes neuronales para la previsión de precios y la optimización de portafolios en la industria financiera. Los resultados muestran que el enfoque integrado supera a otros métodos de análisis financiero y proporciona herramientas más efectivas para los profesionales del mercado, en comparación con una estrategia de *buy and hold* representada en el análisis por el S&P500. Los modelos de inteligencia artificial y aprendizaje automático utilizados en este estudio posibilitan la identificación de patrones y tendencias en los datos financieros, lo que ayuda a los inversionistas a tomar decisiones más informadas y precisas. Además, el estudio demuestra que la inclusión de datos no estructurados, como noticias y redes sociales, en el análisis financiero puede mejorar significativamente la precisión de las predicciones de precios, para lograr un R^2 superior al 65 % y la optimización de portafolios.

Abstract

This study presents an integrated approach of artificial intelligence and machine learning models, combining neural networks for price forecasting and portfolio optimization in the financial industry. The results show that the integrated approach outperforms other financial analysis methods and provides more effective tools for market professionals compared to a buy and hold strategy represented in the analysis by the S&P500. The artificial intelligence and machine learning models used in this study enable the identification of patterns and trends in financial data, helping investors make more informed and accurate decisions. Furthermore, the study demonstrates that the inclusion of unstructured data, such as news and social networks, in financial analysis can significantly improve the accuracy of price predictions achieving an R^2greater than 65% and portfolio optimization.

Palabras clave

Inteligencia artificial, aprendizaje automático, optimización de portafolios, datos estructurados, datos no estructurados, modelos predictivos, análisis financiero, estrategias de inversión, algoritmos de *trading,* mercados financieros, gestión de activos.

Introducción

La teoría moderna de portafolios (MPT), desde su introducción por Markowitz en 1952, ha tenido un impacto significativo en la forma en que los inversionistas administran sus portafolios. Esta teoría se basa en la idea de la diversificación, y propone que los inversionistas pueden construir portafolios optimizados que maximicen los rendimientos esperados para un determinado nivel de riesgo de mercado. En esencia, la MPT plantea que la inversión eficiente implica más que simplemente buscar la mayor rentabilidad; se debe considerar la relación de rendimientos entre diferentes activos para reducir el riesgo y mejorar los rendimientos ajustados al riesgo (Markowitz, 1952).

A pesar de su impacto y de su utilidad, el MPT tiene limitaciones que pueden conducir a decisiones de inversión subóptimas. Una de estas limitaciones es su dependencia de la eficiencia del mercado, que supone que los precios de los activos siempre reflejan toda la información disponible. Sin embargo, existe una creciente evidencia de que los mercados no siempre son eficientes (Shiller, 2003). Además, la MPT supone que los rendimientos se distribuyen normalmente y que los inversionistas tienen una aversión constante al riesgo; suposiciones que a menudo se violan en la práctica (Rachev, 2005).

La inteligencia artificial (AI, por sus siglas en inglés) y el aprendizaje automático (ML, por sus siglas en inglés) pueden ofrecer soluciones a algunos de estos problemas. La capacidad de estas tecnologías para analizar grandes cantidades de datos y aprender de patrones complejos puede ayudar a predecir los precios de los activos y optimizar los portafolios de manera más efectiva. Mediante el uso de datos estructurados y no estructurados, pueden capturar relaciones complejas y dinámicas que los modelos basados en MPT podrían no ser capaces de manejar. De esta manera, la AI y el ML pueden ayudar a superar algunas de las limitaciones de MPT y mejorar la gestión de portafolios (Bengio, 2013).

Los **datos estructurados y no estructurados** juegan un papel esencial en la proyección de precios y la gestión de portafolios. Los datos estructurados, entre los cuales se encuentran las variables macroeconómicas y los estados financieros, brindan información crucial sobre la salud económica general y el desempeño financiero de las empresas individuales. Por ejemplo, los indicadores macroeconómicos, como las tasas de interés, el producto interno

bruto (PIB) y la inflación pueden tener un impacto significativo en los precios de los activos (Mishkin, 2007). De igual forma, los estados financieros brindan información detallada sobre la situación financiera, los resultados y los flujos de efectivo de las empresas, lo que puede influir en la valoración de sus acciones (Penman, 2010).

Por otro lado, los datos no estructurados, como las noticias y los *tweets,* también pueden desempeñar un papel importante. Estas fuentes de datos pueden contener información relevante que no se captura en los datos estructurados. Por ejemplo, las noticias pueden contener información sobre eventos inesperados que pueden afectar el desempeño de una empresa o la economía en general. Los *tweets* y otras formas de redes sociales pueden proporcionar información sobre la opinión pública o el sentimiento del mercado, lo que a menudo puede influir en los precios de los activos (Bollen, 2011). La combinación de datos estructurados y no estructurados puede proporcionar una imagen más completa y precisa, lo que puede mejorar la predicción de precios y la gestión de portafolios.

En este estudio nos propusimos desarrollar y validar un modelo de gestión de portafolios de manera activa que integra técnicas de AI y ML. Aspiramos a que este enfoque mejore la predicción de precios y la optimización de los pesos de cada uno de los activos presentes en el portafolio, superando al índice de referencia S&P500. En última instancia, buscamos proporcionar una contribución significativa a la evolución de las prácticas de gestión de inversiones y portafolios.

Justificación

La justificación para llevar a cabo esta investigación es la creciente importancia de integrar técnicas de inteligencia artificial y aprendizaje automático con la teoría moderna de portafolios en el contexto financiero. Esta relevancia se fundamenta en diversos argumentos respaldados por la literatura financiera y por datos bibliográficos pertinentes.

En primer lugar, la gestión tradicional de portafolios se basa principalmente en datos financieros estructurados, como precios de acciones y *ratios* financieros. Sin embargo, en la era digital, la cantidad de datos no estructurados, como noticias financieras, comentarios en redes sociales y otros, ha adquirido gran relevancia. Estos datos no estructurados pueden contener información valiosa que influye en el comportamiento del mercado y que no se captura completamente con enfoques tradicionales (Tsai, 2020).

Además, se ha comprobado que la AI y el ML son herramientas poderosas para analizar grandes conjuntos de datos y descubrir patrones ocultos. Su aplicación en la predicción de precios de activos financieros y la optimización de portafolios ha conducido a mejoras significativas en la toma de decisiones de inversión (Golmohammadi, 2020).

Un objetivo ambicioso de esta investigación es superar de manera consistente el rendimiento del S&P500, un índice de referencia ampliamente utilizado en la industria financiera. Lograrlo tendría un impacto significativo en la gestión de inversiones y podría generar un valor sustancial para los inversionistas (Fama, 2010).

Además, esta tesis tiene la intención de contribuir al avance de la literatura financiera al explorar enfoques más sofisticados y efectivos que integren modelos de AI y ML con la teoría moderna de portafolios (Kumar, 2021).

En última instancia, los resultados de esta investigación tienen implicaciones prácticas en la gestión de activos y la toma de decisiones de inversión en la industria financiera. Los profesionales del mercado, como gestores de fondos, analistas y asesores financieros, pueden beneficiarse de estrategias de inversión más avanzadas y efectivas.

Esta tesis busca responder a una pregunta de investigación crucial en el área de las finanzas: ¿es posible utilizar modelos de inteligencia artificial junto con la teoría moderna de

portafolios, empleando datos paramétricos y no paramétricos, para construir un portafolio que supere el rendimiento del S&P500 como *benchmark*? Esta investigación busca contribuir al conocimiento financiero y proporcionar herramientas más efectivas para los profesionales del mercado.

Objetivos

Objetivo general

Diseñar y desarrollar un modelo de optimización de portafolios y proyección de precios de activos, basado en inteligencia artificial y aprendizaje automático, que incorpore datos estructurados y no estructurados para mejorar la rentabilidad y reducir el riesgo de un portafolio de inversión.

Objetivos específicos

- Revisar la literatura sobre las técnicas de optimización de portafolios y las principales técnicas de AI y ML para comprender el contexto y las oportunidades de mejora en la gestión de portafolios de inversión.
- Identificar y seleccionar las variables relevantes para el análisis y la predicción de precios y volatilidades de los activos subyacentes, incluyendo variables macroeconómicas, microestructurales y técnicas.
- Definir y construir los diferentes algoritmos y herramientas que se utilizarán en la construcción del modelo.
- Evaluar la eficiencia del modelo propuesto utilizando técnicas de *backtesting* y métricas de error.
- Comunicar de manera efectiva los resultados y las conclusiones de la investigación, proporcionando una guía práctica para la implementación y el uso del modelo propuesto en la gestión de portafolios de inversión y la toma de decisiones financieras.

Marco conceptual

El campo de la predicción de precios en el mercado de valores ha sido objeto de estudio durante muchos años. A pesar de que la *hipótesis del mercado eficiente* argumenta que los precios de la bolsa son impredecibles, puesto que involucran de manera eficiente la nueva información disponible en el mercado, los investigadores continúan buscando algoritmos que puedan proporcionar resultados rentables. En este contexto, se distinguen dos enfoques principales de análisis: el análisis técnico y el análisis fundamental (Girija y Attigeri, 2015).

El **análisis técnico** se basa en el estudio de los precios históricos de una acción para predecir su comportamiento futuro. Para este tipo de análisis se utilizan técnicas como la regresión lineal, el análisis discriminante cuadrático y el análisis discriminante lineal (Gareth *et al.,* 2014).

Por otro lado, el **análisis fundamental** se enfoca en utilizar información relevante, como noticias y datos de redes sociales, para evaluar el valor intrínseco de una acción. En el análisis fundamental se emplean técnicas de análisis de sentimientos para clasificar y analizar información en categorías de sentimientos positivos o negativos (Zhao, 2016). Esto se puede aplicar a blogs, redes sociales como X (antes Twitter) y otras fuentes de información (Rajput y Bobde, 2016).

Además de estos enfoques, se ha investigado el uso de índices de precios de acciones como criterio para evaluar el mercado de valores. Se han desarrollado métodos estadísticos y computacionales para la predicción de series de tiempo financieras, pero se reconoce que es un desafío debido a la naturaleza ruidosa, no estacionaria e irregular de los datos financieros.

En la literatura financiera, los métodos de predicción de precios se clasifican en análisis técnico, análisis fundamental, predicción basada en series de tiempo y aprendizaje automático. En particular, los algoritmos de aprendizaje automático, como las redes neuronales artificiales y las máquinas de vectores de soporte (SVM), han demostrado su eficacia en la predicción de precios de acciones.

En los últimos años, el **aprendizaje profundo** ha surgido como un enfoque prometedor en el campo del aprendizaje automático. Es un subcampo de la inteligencia artificial que utiliza algoritmos para modelar conceptos de alto nivel a través de múltiples capas de aprendizaje.

El aprendizaje profundo ofrece ventajas como el aprendizaje automático de características, la alta precisión y la capacidad de generalización. En conclusión, a pesar de los desafíos inherentes a la predicción de precios en el mercado de valores, los investigadores han explorado diversas técnicas y enfoques, incluido el análisis técnico, el análisis fundamental y el aprendizaje automático (Bengio, Courville y Vincent, 2013; LeCun, Bengio y Hinton, 2015; Schmidhuber, 2015). Estos enfoques están pensados para encontrar patrones en los datos históricos y utilizar información relevante para predecir el comportamiento futuro de los precios de las acciones.

Bolsa de valores

La bolsa de valores representa un mercado en el que las partes interesadas en comprar y vender activos financieros interactúan entre sí. Este entorno se utiliza típicamente para la negociación de dos tipos de activos: renta variable (acciones) y renta fija (deudas) (Roldán, 2020). Las bolsas de valores son en su mayoría organizaciones con fines de lucro, a menudo administradas por entidades privadas, y pueden operar tanto a través de lugares físicos como de plataformas virtuales, previa autorización de una entidad gubernamental (Roldán, 2020).

El papel principal de las bolsas de valores radica en fomentar el desarrollo económico, ya que estas facilitan las transacciones entre compradores y vendedores. Además de esta función fundamental, desempeñan varias otras funciones importantes:

Canalización del ahorro hacia inversiones productivas: Estas permiten a las personas invertir sus recursos financieros en empresas y proyectos que generen crecimiento económico. Esto ayuda a movilizar el capital hacia actividades productivas.

Garantía de seguridad jurídica: Las transacciones realizadas en el mercado bursátil están respaldadas por un marco legal y regulador sólido. Esto proporciona a los inversionistas la seguridad de que sus operaciones se llevarán a cabo de manera justa y conforme a la ley.

Facilitación de la liquidez: Uno de los aspectos clave de las bolsas de valores es que brindan liquidez a los inversionistas.

Tabla 1. Información de bolsas de valores

Nombre	**País**	**Ciudad**	**Número de compañías**	**Capitalización bursátil**
New York Stock Exchange	Estados Unidos	Nueva York	2.304	20.388.427,4
NASDAQ	Estados Unidos	Nueva York	2.900	8.827.942,2
Japan Exchange Group Inc	Japón	Tokio	3.557	5.424.014,0
Shanghái Stock Exchange	China	Shanghái	1.284	4.361.154,6
Euronext	Países Bajos	Ámsterdam	1.281	4.059.354,8
London Stock Exchange Group	Reino Unido, Italia	Londres	2.489	4.047.981,2

Fuente: Elaboración propia.

Acciones

Según Dunham y Singal (2014), una acción representa un porcentaje de propiedad en una compañía y es el tipo de inversión más común en los mercados de renta variable. El valor de una acción está vinculado al desempeño de la empresa emisora.

Los inversionistas que compran acciones obtienen ciertos derechos, tales como el derecho a votar en la elección de miembros de la junta directiva y a participar en la asamblea de accionistas para discutir la estrategia y el desempeño de la empresa. El derecho más significativo es el derecho económico, que les otorga participación en las ganancias de la empresa en forma de dividendos, de acuerdo con las políticas de la junta directiva. Si las perspectivas de la empresa son favorables, los inversionistas pueden obtener ganancias a través de la apreciación del valor de sus acciones.

La compra y la venta de participaciones accionarias se pueden realizar de manera privada o a través de una sociedad comisionista de bolsa en el mercado secundario. Por lo general, las personas adquieren acciones en el mercado secundario después de una oferta pública inicial (OPI) en el mercado primario, donde las empresas emiten acciones a través de bancos de inversión y profesionales. Los inversionistas del mercado primario las venden en el mercado secundario, a menudo esperando obtener ganancias rápidas.

La Bolsa de Valores de Colombia (BVC) administra y establece reglas y mecanismos para la negociación de acciones. Las acciones representan derechos sobre los activos y las ganancias de una o varias compañías (AMV, 2006).

El mercado de acciones funciona de manera similar a un mercado ordinario, donde las partes acuerdan un precio para el intercambio. Al comprar acciones, se adquiere el activo subyacente, lo que conlleva riesgos. Si la empresa se valoriza, el precio de las acciones aumenta, lo que puede generar ganancias en su venta. Por otro lado, si la empresa se devalúa, el precio de las acciones cae, lo que puede resultar en pérdidas (AMV, 2006).

Ilustración 1. Comportamiento de las empresas colombianas en la bolsa de valores de Wall Street

Fuente: Valora Analitik (2019).

Para pronosticar los precios de las acciones, existen dos enfoques importantes en la literatura: el análisis fundamental, que utiliza la información financiera de la empresa, y el análisis técnico, que se basa en las tendencias del mercado. Ambos métodos se han utilizado para analizar el mercado de valores (Chen *et al.,* 2003) (Kimoto, 1990). Sin embargo, factores como las suposiciones sesgadas, el comportamiento irracional de los inversionistas y la selección de datos pueden introducir sesgos en el análisis fundamental. Además, el análisis técnico se basa en tendencias pasadas, y no hay evidencia sólida de que el mercado de valores siga patrones regulares. Por lo tanto, combinar ambos enfoques puede dar lugar a un mejor rendimiento predictivo (Chen, 2003).

Índices bursátiles

Los índices bursátiles son herramientas fundamentales en el mundo de las finanzas, ya que reflejan el comportamiento colectivo de grupos específicos, en este caso de acciones en los mercados financieros. A continuación, se presentan algunos de los principales índices bursátiles de la actualidad:

1. **Promedio Industrial Dow Jones (DJIA):** El DJIA es uno de los índices bursátiles más reconocidos en Estados Unidos. Este índice, propiedad de Dow Jones & Company, sigue el movimiento de precios de 30 grandes empresas estadounidenses que cotizan en el NASDAQ y la Bolsa de Valores de Nueva York (NYSE). El DJIA se considera un indicador clave de las condiciones generales del mercado y la economía estadounidense en su conjunto (Hall, 2021).

2. **S&P 500:** El S&P 500, o el índice Standard & Poor's 500, es un índice ponderado por capitalización de mercado que abarca las 500 empresas más grandes que cotizan en bolsa en Estados Unidos. Aunque no refleja exactamente las 500 principales empresas por capitalización de mercado, es ampliamente aceptado como el mejor indicador de las acciones de gran capitalización en Estados Unidos. El S&P 500 incluye muchas empresas de tecnología y financieras (Hall, 2021).

3. **IBEX 35:** El IBEX 35 es el índice bursátil de referencia de la bolsa española. Este índice mide el comportamiento conjunto de las 35 empresas más negociadas que cotizan en el Sistema de Interconexión Bursátil Electrónico (SIBE) en las cuatro bolsas españolas: Madrid, Barcelona, Bilbao y Valencia. Estas empresas son las que despiertan un mayor interés entre compradores y vendedores (Hall, 2021).

4. **Nikkei:** El Nikkei 225 Stock Average es el índice líder y más respetado de las acciones japonesas. Este índice ponderado por precio está compuesto por las 225 principales empresas de primera línea de Japón que cotizan en la Bolsa de Valores de Tokio. El Nikkei se considera el equivalente japonés del índice Dow Jones Industrial Average (DJIA) en Estados Unidos (Hall, 2021).

Estos índices proporcionan a los inversionistas y analistas una visión general del rendimiento de los mercados financieros y de la economía en general en sus respectivas regiones geográficas. Además, son herramientas útiles para realizar un seguimiento de la evolución de los precios de las acciones y tomar decisiones de inversión informadas.

Teoría de portafolios

A medida que el mercado financiero crece en tamaño, la cantidad de productos e índices negociables que necesitan análisis aumenta significativamente, tanto en el ámbito nacional

como en el internacional. En consecuencia, los inversionistas individuales se encuentran en desventaja en comparación con los inversionistas institucionales debido a sus capacidades limitadas. Para hacer frente a las complejidades y la escala del mercado, muchos recurrieron al poder computacional y las técnicas de aprendizaje automático.

Los métodos de negociación automatizados, que utilizan enfoques de aprendizaje automático como redes neuronales artificiales, máquinas de vectores de soporte, aprendizaje de refuerzo, LSTM y mecanismos de atención basados en análisis técnico, se emplean comúnmente para pronosticar las cotizaciones del mercado. Sin embargo, la naturaleza volátil e impredecible del mercado financiero a menudo dificulta la precisión de tales predicciones. Los datos financieros están plagados de ruido no identificable, lo que hace que el comercio algorítmico sea un desafío.

Si bien las técnicas tradicionales de aprendizaje automático han mostrado resultados prometedores en experimentos académicos, su éxito en las inversiones de la vida real es relativamente limitado en comparación con otros campos. Lograr una alta precisión en las predicciones bursátiles, incluso al 70 % o al 80 %, puede no ser suficiente, ya que una sola estimación incorrecta podría generar pérdidas sustanciales. Factores como el ruido, las dimensiones complejas de los datos y varios elementos socioeconómicos presentan obstáculos en el reconocimiento de patrones.

Para abordar los desafíos planteados por el ruido y la complejidad, los modelos de diversificación y valoración de activos ofrecen soluciones viables. La creación de portafolios y el uso de modelos de valoración de activos ajustados al riesgo proporcionan mejores rendimientos por riesgo y establecen una estrategia de inversión sostenible. Algunos economistas, como Markowitz, Sharpe y Lintner, han realizado extensas investigaciones sobre la medición de los rendimientos con equilibrio de riesgo y la evaluación de los precios de las acciones.

Teoría moderna de portafolios

Todo inversionista racional busca mayores rendimientos, aunque en el mundo de la inversión financiera los mayores rendimientos a menudo traen mayores costos, es decir, mayores riesgos. En términos más simples, si una inversión potencial promete mayores ganancias,

también conlleva una mayor probabilidad de fracaso. Esto significa que no todos los inversionistas buscarán automáticamente altos rendimientos. Cada inversionista tiene su propia tolerancia al riesgo; algunos prefieren inversiones de alto riesgo y alto rendimiento, mientras que otros prefieren opciones estables y menos riesgosas, incluso si eso significa rendimientos potenciales más bajos. El desafío radica en encontrar un producto de inversión que ofrezca ganancias más altas con un riesgo mínimo. Sin embargo, el mercado en constante evolución, influenciado por numerosos factores, hace que sea extremadamente difícil identificar una acción tan singular (Markowitz, 1952).

La teoría moderna de portafolios postula que los inversionistas prefieren inversiones que brinden los rendimientos más altos para cierto nivel de riesgo o activos menos riesgosos si los rendimientos esperados son comparables (Sharpe, 1970).

Según Markowitz, la construcción de un portafolio eficiente compuesto por diversos activos financieros, en lugar de invertir únicamente en un producto básico, permite maximizar los retornos para un nivel dado de volatilidad (Markowitz, 1952). Cada activo del portafolio se selecciona para mitigar la incertidumbre, con la combinación de acciones compensando las características individuales de rendimiento/riesgo y formando un sintético general para todo el portafolio. El rendimiento esperado del portafolio se calcula sumando el rendimiento de las acciones individuales en función de su peso proporcional, mientras que el riesgo se evalúa a través de medidas estadísticas como la varianza y la covarianza.

$$\boldsymbol{Retorno\ portafolio} = \sum Wi * Ri$$

La diversificación, como se propone en la teoría, trata sobre cómo al combinar activos poco correlacionados o con una correlación negativa en un portafolio es posible que el inversionista reduzca el riesgo o la desviación de su inversión.

Ilustración 2. Diversificación

Asset Class	Annual Average Return	Standard Deviation (Risk)
Small-cap stocks	High	High
Large-cap stocks	↓	↓
Long term corporate bonds		
Long term treasury bonds	Low	Low
Treasury bills		

Fuente: Elaboración propia.

***Capital asset pricing model* (CAPM) y *security market line* (SML)**

En finanzas, y especialmente en la teoría moderna de portafolios, existen dos tipos de riesgo: el **riesgo sistemático** y **riesgo no sistemático.** Markowitz y otros investigadores propusieron la idea de reducir la volatilidad mediante la construcción de portafolios diversificados. El riesgo no sistemático, también conocido como riesgo idiosincrásico, se refiere al riesgo propio de cada activo, el cual se puede mitigar a través de la diversificación. Por otro lado, el riesgo sistemático o riesgo no diversificable no puede eliminarse porque es inherente al mercado en general, no específico de activos individuales (Markowitz, 1952).

Ampliando la teoría de portafolios de Markowitz, el modelo CAPM proporciona el rendimiento esperado de un activo en función de su riesgo sistemático. El CAPM está representado por una línea de mercado de valores (*security market line* - SML), graficada en la figura 3, que evalúa el rendimiento esperado para activo contra diferentes niveles de riesgo de mercado en términos del Beta (Markowitz, 1952).

El Beta mide la volatilidad o la sensibilidad de una acción individual en relación con el movimiento de su mercado. Un valor Beta alto superior a 1,0 indica que la acción es más volátil o más riesgosa que el mercado, mientras que un valor Beta bajo inferior a 1,0 sugiere que es más probable que la acción fluctúe menos que el mercado. Un Beta de 1,0 significa que las acciones se mueven en línea con el mercado (Sharpe, 1970).

Al trazar el coeficiente Beta de la acción en el eje **x** y el rendimiento esperado en el eje **y**, la *SML* ilustra la relación riesgo-rendimiento de acuerdo con el modelo CAPM. La

compensación riesgo-rendimiento, o prima de riesgo (que se muestra en la figura 3), representa el exceso de retorno sobre la tasa libre de riesgo que puede ofrecer una inversión y se refleja en el valor de la pendiente de la *SML*.

$$\beta = \frac{Cov(R_i, R_j)}{Var(R_m)}$$

Para calcular el rendimiento esperado de un activo de acuerdo con un riesgo sistemático dado, se puede sumar la tasa libre de riesgo a la prima de riesgo multiplicada por el valor de Beta. Se considera que las acciones ubicadas por encima de la *SML* proporcionan mejores rendimientos debido a su nivel de riesgo, y se supone que están subvaloradas (Sharpe, 1970). Por el contrario, las acciones ubicadas por debajo de la *SML*, donde el rendimiento es inferior a un rendimiento esperado de acuerdo con el modelo CAPM, debido a un nivel de riesgo sistemático, se consideran sobrevaloradas, y los inversionistas deben tener cuidado al incluirlas en sus portafolios de inversión.

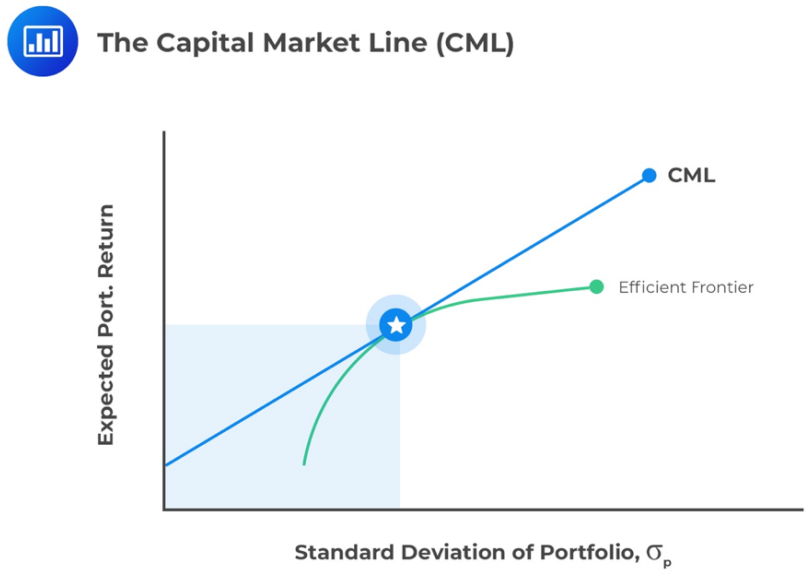
$$R = Real\ Free\ interest\ reate + Inflation\ Premium + Default\ risk\ premium + Liquidity\ premium + Maturity\ premium$$

- ***Real free interest rate:*** Tasa de un activo libre de riesgo si no hubiera inflación.
- ***Inflation premium:*** Compensa la inflación esperada y muestra el promedio de la inflación esperada hasta el vencimiento.
- ***Nominal free interest rate:*** Real free interest rate + Inflation premium. Algunos países usan sus bonos como referencia de este indicador; por ejemplo, los bonos de USA.
- ***Default risk premium:*** Compensa ante la posibilidad de que el prestatario quede en bancarrota.
- ***Liquidity premium:*** Prima que compensa al inversionista por el riesgo en que puede incurrir al querer vender su activo y que deba venderlo más barato. Los bonos no tienen esta prima, ya que son muy líquidos, por lo que los corporativos son más altos.

- ***Maturity premium:*** Es la compensación ante la sensibilidad por cambios en la tasa de cambio.

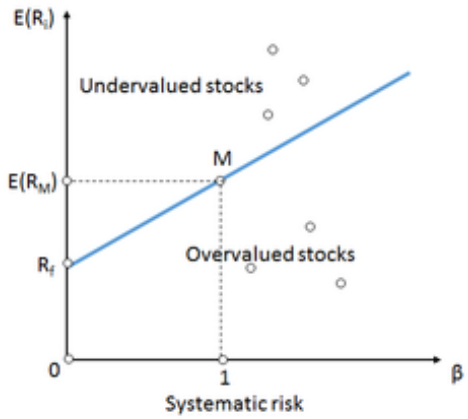
$$E(P\ con\ rf) = Rf + \frac{[E(p) - Rf]}{desviacion\ p} * desviacion\ p$$

Ilustración 3. *Capital market line*



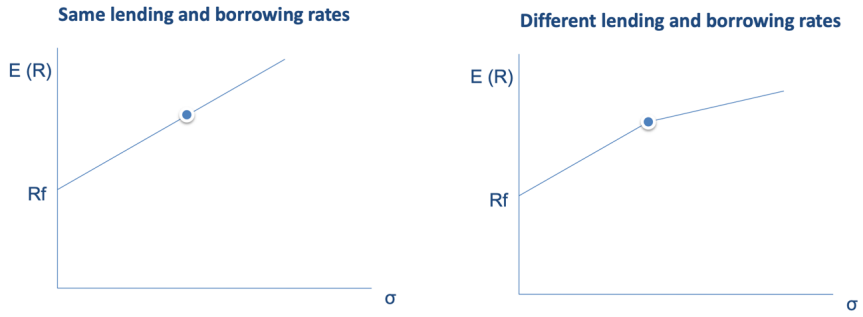
Fuente: Elaboración propia.

Ilustración 4. CML



Fuente: Elaboración propia.

Ilustración 5. Portafolios apalancados



Fuente: Elaboración propia.

Tolerancia al riesgo

La tolerancia al riesgo es un concepto crucial en la teoría de portafolios, que se refiere a la disposición y la capacidad de un individuo o un inversionista para asumir riesgos en busca de rendimientos potenciales. Este refleja la capacidad de una persona para resistir las fluctuaciones en el valor de sus inversiones.

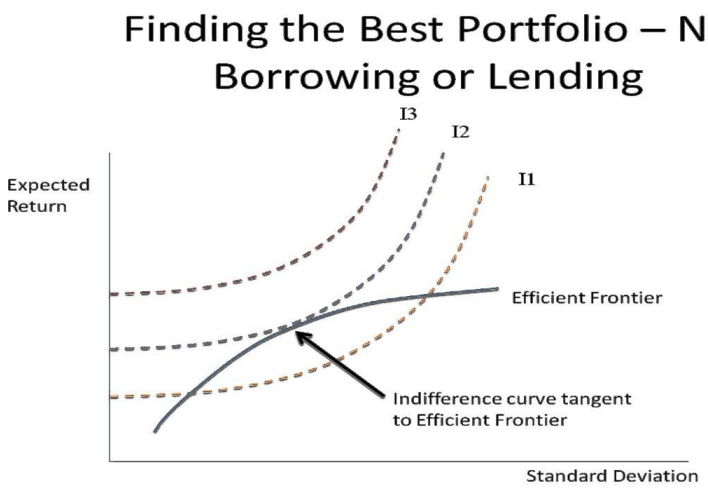
La evaluación de la tolerancia al riesgo es un componente esencial en la teoría de portafolios, ya que ayuda a los inversionistas a determinar la asignación de activos adecuada que se alinea con sus objetivos financieros, su horizonte temporal y su comodidad psicológica. Un inversionista con una alta tolerancia al riesgo puede estar más inclinado a invertir en activos de mayor riesgo, con el objetivo de obtener rendimientos potencialmente más altos. Por otro lado, un inversionista con una menor tolerancia al riesgo puede priorizar la preservación del capital y optar por inversiones más conservadoras y de menor riesgo (Graham, 1949). Al comprender su tolerancia al riesgo, los inversionistas pueden crear portafolios que coincidan con sus preferencias individuales, optimizar los rendimientos potenciales y garantizar una experiencia de inversión más estable y satisfactoria a largo plazo. En los modelos de optimización de media-varianza, la utilidad de un inversionista viene dada por la siguiente expresión:

$$\mathrm{U} = \mathrm{E}(\mathrm{r}) - \frac{1}{2} * A * Varianza$$

- Donde E(r) se refiere al retorno esperado de un activo.
- A es el nivel de aversión al riesgo del inversionista.

Curva de indiferencia: Combinación de riesgo/retorno que un inversionista acepta para una utilidad esperada.

Ilustración 6. Curva de indiferencia



Fuente: Elaboración propia.

Perfiles

En el contexto de la inversión y la gestión de portafolios existen varios perfiles de inversionistas basados en la tolerancia al riesgo. Estos perfiles ayudan a clasificar a los inversionistas según su disposición y su capacidad para asumir riesgos. Estos son algunos de los perfiles de inversionistas:

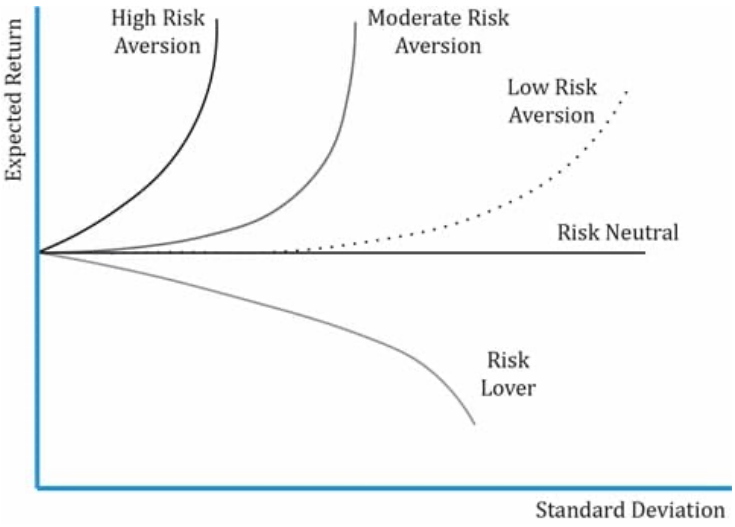
- **Inversionistas conservadores:** Tienen una baja tolerancia al riesgo y priorizan la preservación del capital sobre los altos rendimientos. Prefieren invertir en activos de bajo riesgo, como bonos del Gobierno o equivalentes de efectivo, con el fin de minimizar las posibilidades de perder el capital invertido (Graham, 1949).

- **Inversionistas moderados:** Están dispuestos a asumir un nivel moderado de riesgo para buscar rendimientos razonables. Su portafolio de inversiones puede incluir una combinación de activos conservadores y orientados al crecimiento (Graham, 1949).

- **Inversionistas agresivos:** Tienen una alta tolerancia al riesgo y están dispuestos a asumir un riesgo significativo para buscar mayores rendimientos. Se sienten más cómodos con una asignación alta a acciones y pueden explorar oportunidades de

inversión de mayor riesgo y alto retorno (Graham, 1949).

Ilustración 7.Perfil de riesgo



Fuente: Elaboración propia.

Medición de riesgos

La desviación de los retornos de un portafolio, también conocida como varianza del portafolio o riesgo del portafolio, mide el alcance de las fluctuaciones o la variabilidad en los rendimientos de portafolio de inversiones. Es un concepto crucial en la gestión de portafolios, ya que cuantifica el riesgo global del mismo.

Ilustración 8. Desviación estándar de un portafolio

$$\sigma_p = \sqrt{w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \rho_{1,2} \sigma_1 \sigma_2}$$

$$\sigma_p = \sqrt{w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 Cov_{1,2}}$$

Fuente: Elaboración propia.

En el contexto de la teoría de portafolios, la desviación de este se calcula considerando el riesgo individual de cada activo en el portafolio y sus correlaciones entre sí. La diversificación es una estrategia fundamental para reducir la desviación del portafolio, ya que ayuda a compensar el riesgo no sistemático mediante la inclusión de diferentes activos que no están perfectamente correlacionados (Elton *et al.,* 1965).

Los inversionistas buscan construir portafolios que optimicen los rendimientos mientras gestionan el riesgo de manera eficaz. Al comprender y gestionar la desviación de un portafolio, los inversionistas pueden lograr un equilibrio entre el riesgo y el retorno, adaptando sus estrategias de inversión para alinearse con su tolerancia al riesgo y sus objetivos financieros..

Medidas de desempeño

Sharpe ratio

William F. Sharpe introdujo un modelo simplificado que mejora los aspectos prácticos del enfoque de media-varianza de Markowitz. A diferencia del modelo anterior, que requería cálculos complejos y lentos de matrices de varianza y covarianza para cada acción individual en el portafolio, el modelo de Sharpe evalúa el riesgo total de un portafolio a través de un análisis de regresión sencillo, lo que hace que el proceso de cálculo sea más eficiente (Sharpe, 1970).

Además, Sharpe desarrolló una métrica llamada ***ratio* de Sharpe** para evaluar el retorno de una inversión en relación con el nivel de riesgo asumido. El *ratio* de Sharpe compara el exceso de rendimiento de un activo sobre un punto de referencia (el rendimiento del activo libre de riesgo), con el fin de determinar cuánto retorno se logra para el mismo nivel de

riesgo. Un *Ratio* de Sharpe más alto indica un portafolio con un mejor rendimiento ajustado al riesgo, mientras que los valores negativos indican que el activo libre de riesgo ofrece un rendimiento más alto que las acciones seleccionadas. En esencia, el *ratio* de Sharpe ayuda a los inversionistas a medir qué tan atractiva es una inversión al considerar tanto sus características de rendimiento como de riesgo (Sharpe, 1970).

$$Sharpe\ Ratio = \frac{R_p - R_f}{Volatilidad_p}$$

Alfa de Jansen

El Alfa de Jensen, también conocido como índice de rendimiento de Jensen, es una métrica utilizada en la teoría de portafolios para evaluar el rendimiento de una inversión o de un portafolio, en relación con su rendimiento esperado en función de su nivel de riesgo (Elton *et al.,* 1965).

Alfa de Jensen: $R_p - \left(R_f + Beta * \left(R_m - R_f\right)\right)$

Fue desarrollado por Michael Jensen y complementa el modelo CAPM. Este ayuda a evaluar si una inversión proporciona rendimientos adecuados para su nivel de riesgo sistemático, representado por el coeficiente Beta. Es posible que el CAPM no explique por completo el rendimiento real de una inversión, ya que podría haber factores específicos de la inversión (riesgo idiosincrático) que contribuyan a sus rendimientos y que no son tenidos en cuenta por el CAPM (Elton *et al.,* 1965).

El Alfa de Jensen busca capturar estos rendimientos adicionales, más allá de lo que se esperaría del CAPM. Si el Alfa de una inversión es positivo, significa que la inversión ha superado su rendimiento esperado en función del riesgo del mercado. Por el contrario, un Alfa negativo indica un bajo rendimiento. En la teoría de portafolios, el Alfa de Jensen ayuda a los inversionistas a identificar gestores cualificados que generen rendimientos superiores de forma constante, distinguiendo sus capacidades de las que simplemente se benefician de los movimientos del mercado (Elton *et al.,* 1965).

Treynor ratio

El índice de Treynor, llamado así por Jack L. Treynor, es una medida de rendimiento financiero *utilizada* en la teoría de portafolios para evaluar los rendimientos ajustados al riesgo de una inversión o portafolio. Evalúa el exceso de rendimiento generado por una inversión en relación con su riesgo sistemático, medido por su coeficiente Beta en el modelo CAPM (Treynor, 2014).

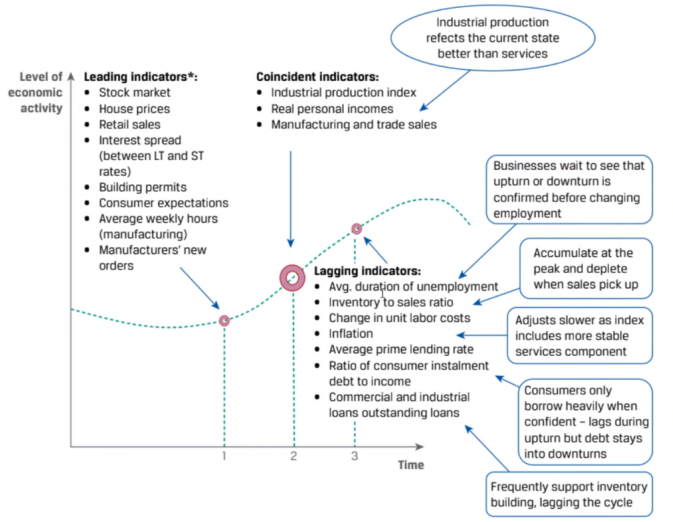
Treynor ratio: $\frac{Rp-Rf}{Beta\ p}$

Un índice de *Treynor* más alto indica un mejor rendimiento ajustado por riesgo, ya que significa que la inversión ha generado mayores rendimientos por unidad de riesgo sistemático asumida (Treynor, 2014).

Data estructurada y no estructurada

Los **datos estructurados** se refieren a información que está organizada y formateada de una manera predefinida, lo que facilita su almacenamiento, su búsqueda y su análisis. Este tipo de datos generalmente se almacena en bases de datos relacionales o formatos de hojas de Excel, donde cada elemento de datos tiene un tipo de datos específico y se organiza en filas y columnas. Los datos estructurados tienen un esquema o un modelo de datos claro, lo que significa que la estructura de los datos está bien definida y es coherente. Los ejemplos de datos estructurados incluyen datos numéricos, fechas, nombres, direcciones y variables categóricas, lo que los hace muy adecuados para el análisis sistemático mediante consultas, filtros y métodos estadísticos.

Ilustración 9. Indicadores económicos



Fuente: Elaboración propia.

Por otro lado, los **datos no estructurados** se refieren a información que no tiene una estructura predefinida o un formato organizado. Los datos no estructurados a menudo se encuentran en forma de texto, imágenes, audio, video, publicaciones en redes sociales, correos electrónicos y otro contenido de formato libre. A diferencia de los datos estructurados, los datos no estructurados carecen de un esquema fijo, lo que dificulta su almacenamiento, su procesamiento y su análisis mediante bases de datos tradicionales. Los datos no estructurados requieren técnicas avanzadas, como el NPL *(natural language processing),* el reconocimiento de imágenes y los algoritmos de aprendizaje automático para obtener conocimientos y dar sentido a la información que contienen.

En la investigación actual, las siguientes variables se utilizarán como entradas para nuestro modelo:

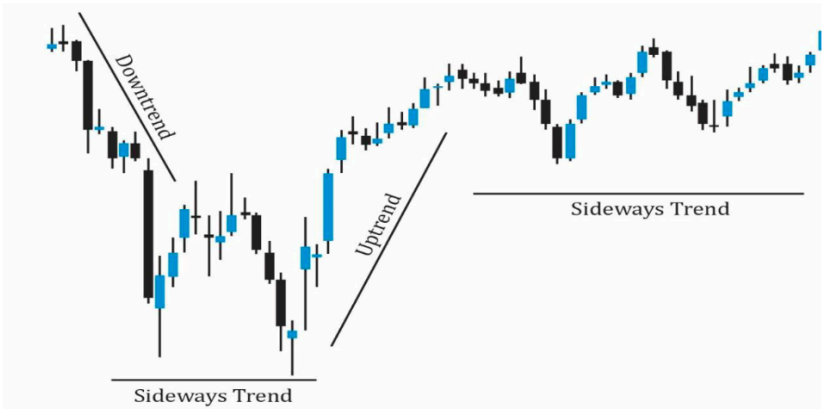
Retornos pasados

Los rendimientos históricos se refieren al rendimiento pasado de un activo financiero o portafolio durante un período específico. Estos sirven como indicadores para la predicción de precios futuros, en este caso de acciones, con base en la hipótesis de los mercados

eficientes y la idea de que el comportamiento pasado puede proporcionar información sobre el comportamiento futuro de estos activos, lo que ocasiona que las tendencias y los patrones se puedan repetir en el futuro (Malkiel, 2003).

Uno de los principales enfoques para aprovechar este insumo es el análisis técnico, en el que los analistas estudian gráficos de precios históricos y aplican análisis como medias móviles, bandas de Bollinger e índices RSI, para identificar patrones y señales que permitan predecir movimientos futuros (Lo, 2010) .

Ilustración 10. Análisis de tendencias



Fuente: Elaboración propia.

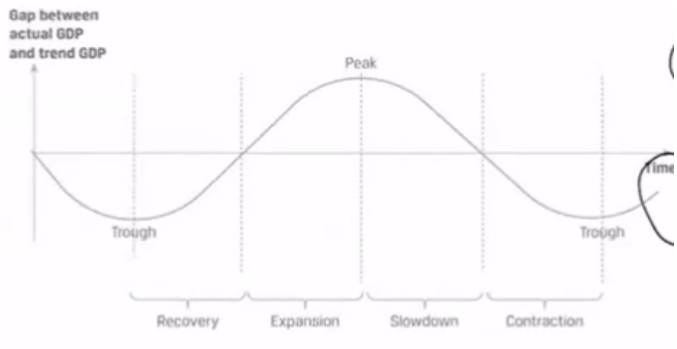
Utilizando datos históricos, se pueden aplicar diversas técnicas cuantitativas y modelos estadísticos, tales como análisis de series temporales, análisis de regresión y algoritmos de aprendizaje automático, para predecir futuros movimientos de precios y estimar los rendimientos potenciales.

Producto interno bruto

El PIB es un indicador económico que refleja la salud general y el crecimiento de una economía. Representa el valor total de bienes y servicios producidos en una economía durante un período específico. Los cambios en el PIB pueden tener impactos significativos

en varios sectores e industrias, lo que a su vez puede influir en el desempeño de los índices bursátiles (Fama, 1989).

Ilustración 11. Ciclos de negocios



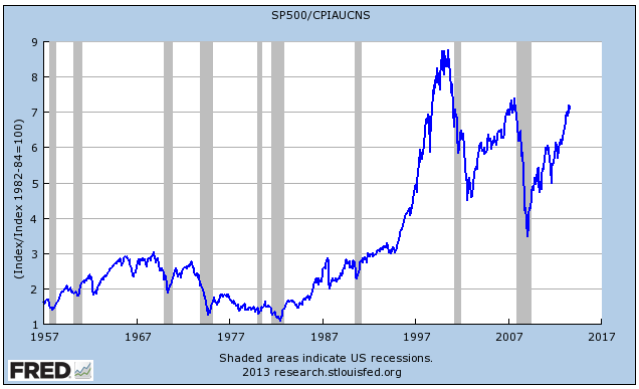
Fuente: Elaboración propia.

Cuando el crecimiento del PIB es sólido y la economía se está expandiendo, generalmente se debe a un aumento de las actividades comerciales, un mayor gasto de los consumidores y mayores ganancias corporativas. Por el contrario, una desaceleración en el crecimiento del PIB o una recesión pueden conducir a una reducción de las actividades comerciales, una disminución del gasto de los consumidores y menores ganancias corporativas. Estas condiciones a menudo dan como resultado una disminución de la confianza de los inversionistas y podrían provocar caídas en los índices bursátiles.

Inflación

La inflación se refiere a la tasa a la que el nivel general de precios de bienes y servicios en una economía aumenta con el tiempo (Rapach, 2005).

Ilustración 12. S&P 500 y recesiones

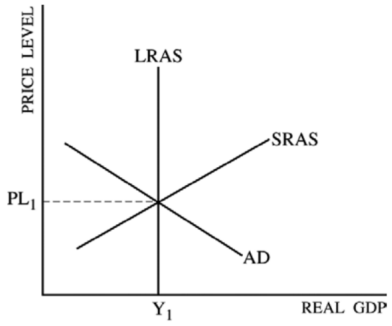


Fuente: Elaboración propia.

- ***CPI:*** *Consumer price index urban.* Se usa en la política económica y para ajustar los contratos.
- ***PCE:*** Consumo personal usando encuestas a empresas.
- ***PPI*****:** Índice de precios al productor. Este influencia el futuro CPI, es más volátil que el CPI y puede variar los pesos según el país.
- ***Core inflation:*** Es la inflación sin energía ni alimentos, que se genera porque estos dos ítems se presentan a corto plazo, no muestran claramente el comportamiento del ciclo económico y se ven altamente afectados por el exterior. Es la que se usa como político-económica.

Cuando la inflación es baja y estable generalmente da cuenta de una economía saludable y puede tener un impacto positivo en el poder adquisitivo del consumidor y la rentabilidad comercial. Sin embargo, si la inflación aumenta rápidamente o se vuelve incontrolable, puede eliminar el poder adquisitivo de los consumidores y reducir las ganancias corporativas. La alta inflación puede generar incertidumbre en los mercados y hacer que los inversionistas busquen activos refugio en lugar de invertir en acciones, lo que podría provocar caídas en los índices bursátiles (Fama, 1989).

Ilustración 13. Equilibrio del PIB e inflación



Fuente: Elaboración propia.

Equilibrio en el corto plazo:

- *Full* empleo en el largo plazo.
- Recesión en el corto plazo.
- Inlfación en el corto plazo.
- Stanflación en el corto plazo.

Equilibrio en el **largo plazo:** GDP= Potential GDP. Capital y empleo están *full.*

- **Inflación por el *cost push:*** Las empresas suben los salarios, lo que hace que estas aumenten sus precios. NAIRU o NARU refleja el potencial de una economía y esta no puede ser observada directamente, por lo que este indicador es muy alto cuando la fuerza laboral no puede satisfacer las necesidades de las empresas. A mayor productividad o costos de las materias primas, aumentan los salarios, pero los precios de los productos no aumentan en la misma proporción.

- **Inflación por la demanda:** Aumenta la demanda, por lo que los precios aumentan y los salarios aumentan. Muestra una relación entre la real y la potencial. Si el bien aumenta la calidad, satisface más y hace que aumente la inflación por este rubro.

Desempleo

La tasa de desempleo se refiere al porcentaje de la fuerza laboral que está desempleada y buscando empleo activamente. Los cambios en la tasa de desempleo pueden brindar información valiosa sobre la salud del mercado laboral y las condiciones económicas generales, lo que a su vez puede afectar los índices bursátiles (Campbell, 1997).

Las bajas tasas de desempleo generalmente dan cuenta de un mercado laboral sólido y un mayor gasto de los consumidores, lo que puede influir positivamente en las ganancias corporativas y la confianza de los inversionistas. Pero el aumento de las tasas de desempleo puede indicar debilidad económica y reducción del gasto de los consumidores, lo que podría conducir a menores ganancias corporativas y una menor confianza de los inversionistas (Fama, 1989).

- **Empleados:** Personas con trabajo.
- **Fuerza laboral:** Personas que tienen o están buscando trabajo.
- **Desempleados:** Personas que están buscando trabajo, pero no tienen uno. Desempleados/fuerza laboral.
    - **Largo plazo:** Que llevan mucho tiempo sin trabajo, pero siguen buscando.
    - **Friccionados:** Que se toman su tiempo para buscar algo que se corresponda con sus gustos. Hace poco terminó su trabajo o están próximos a empezar uno nuevo.
- **Porcentaje de desempleados:** Desempleados/fuerza laboral.
- **Porcentaje de participación:** Fuerza laboral/personas que pueden trabajar (16 a 64 años).
- **Subempleados:** Personas que tienen trabajo, pero podría estar en uno mejor por sus cualidades.

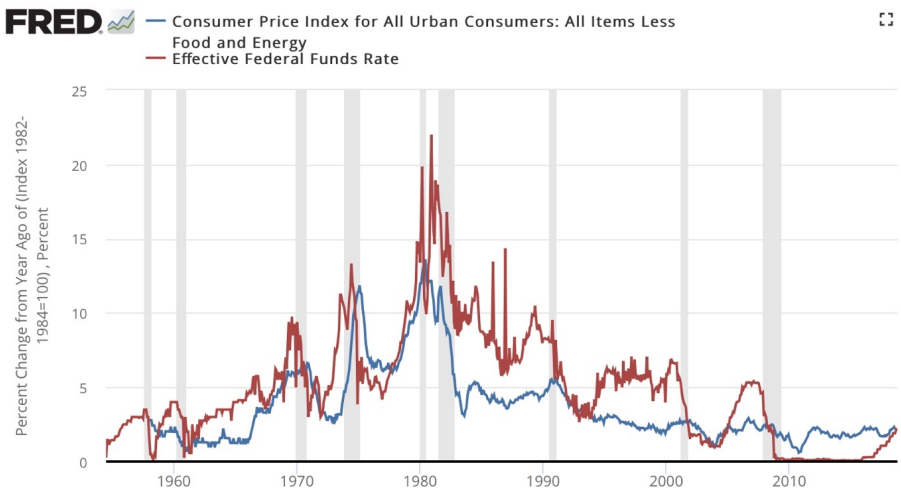
- **Empleados desalentados u ocultos:** Los que dejaron de buscar trabajo. No incluye el voluntario.
- **Desempleados voluntarios:** Que por sí mismos dejaron de trabajar.

Los bancos centrales y los formuladores de políticas monitorean de cerca los datos de desempleo como parte de sus decisiones de política económica. Las tendencias del desempleo pueden influir en la política monetaria, las decisiones sobre tasa de interés y las medidas fiscales, lo que repercutirá aún más en el entorno económico general y los mercados financieros, incluidos los índices bursátiles.

Tasas de interés

Las tasas de interés juegan un papel importante, pues influyen en varios aspectos de la economía y los mercados financieros, incluidos los índices bursátiles.

Ilustración 14. Inflación y tasas de política monetaria



Fuente: Elaboración propia.

Reserva bancaria: Lo que deben de dejar los bancos para respaldar sus préstamos, mientras que el resto del dinero puede ser puesto en circulación. De esta manera se crea dinero, con la siguiente fórmula:

$$\frac{Depósito}{Reserva\ bancaria}$$

$$Multiplicador\ de\ dinero = \frac{1}{Reserva\ bancaria}$$

Dinero en circulación: Efectivo, depósitos bancarios.

- ***Narrow money:*** Notas y dinero en circulación más los depósitos que son líquidos.
- ***Borrow money:*** *Narrow money* + todos los activos líquidos con los que se pueden comprar bienes.

Tasa neutral: No estimula ni hacia arriba ni hacia abajo la economía. Mantiene estables en el largo plazo la inflación y el empleo.

Cantidad teórica del dinero: Existe una relación entre la cantidad de dinero y el nivel de precios. Se presume que la velocidad del dinero es constante. Se usa la ecuación:

$$M * V = P * Y$$

- M, cantidad de dinero.
- V, velocidad de circulación.
- P, precio.
- Y, cantidades, GDP.

Esta teoría plantea que la velocidad del dinero v se mantiene constante, lo que indica que P*Y es proporcional a M. Un amento en M hace que los niveles del precio aumenten, o sea la inflación. La cantidad de dinero en circulación depende del nivel de la actividad económica.

Supuestos:

- El dinero es neutral.

Ahorro del dinero: Cantidad que los hogares guardan para:

- Transacciones futuras.
- Precaución.
- Especulación.
- **Balance de transacciones:** Saldo de dinero que se mantiene para financiar transacciones. Este crece con el GDP.

- **Balance de transacciones de precaución:** Es lo que se guarda en caso de emergencia. Este tiende a ser grande cuando se hacen muchas transacciones, y está directamente relacionado con el GDP.
- **Balance de transacciones especulativas o portafolio de demanda de dinero:** Es lo que se ahorra para posibles oportunidades de inversión o riesgos. Son como pérdidas esperadas en activos. Este portafolio es inversamente proporcional a las ganancias esperadas en activos.

Demanda y oferta de dinero: El precio de equilibrio es la tasa de interés nominal que se recibe por prestar este dinero. Un cambio en la política monetaria hace que en el corto plazo se afecten, pero en el largo plazo se vuelva a un equilibrio, y esto se llama neutralidad del dinero.

- Aumento en la tasa, exceso de oferta de dinero, lo que hace que se compren bonos con ese exceso, que aumente el precio de estos y que la tasa vuelva al equilibrio.
- Disminución en la tasa, exceso de demanda. Salen a vender bonos, lo que hace que bajen de precio y la tasa suba al equilibrio.

Fisher effect: R nominal = R real + Inflación esperada. Demuestra que a lo largo del tiempo la cantidad de dinero no afecta la tasa de interés real y, por ende, en el largo plazo la política monetaria no afecta la economía, pero sí la inflación y, en consecuencia, la tasa nominal. También habla de que los bonos de los Gobiernos tienen intrínseca una inflación esperada en la tasa que se negocia. La tasa nominal está compuesta por los siguientes elemetos:

- Rendimiento esperado.
- Inflación esperada.
- *Risk premium,* porque no hay certeza sobre las variables macro, entonces a mayor incertidumbre es mayor la tasa que se espera.

Cuando las tasas de interés son bajas, los costos de endeudamiento disminuyen, lo que abarata el acceso al crédito para las empresas y los consumidores. Las tasas de interés más bajas pueden estimular la actividad económica, aumentar el gasto de los consumidores y aumentar las ganancias corporativas. Por el contrario, el aumento de las tasas de interés puede conducir a mayores costos de endeudamiento, lo que puede frenar el gasto de los consumidores y afectar la inversión empresarial.

Confianza del consumidor

La confianza del consumidor se basa en encuestas que miden la opinión de las personas sobre las condiciones económicas presentes y futuras. Este indicador refleja la percepción de los consumidores sobre el empleo, los ingresos, la inflación y otros factores económicos clave. Se considera un reflejo importante del sentimiento del mercado y puede influir en el comportamiento de gasto y ahorro de los consumidores (Baker, 2007).

Ilustración 15. Índice de sentimientos del mercado

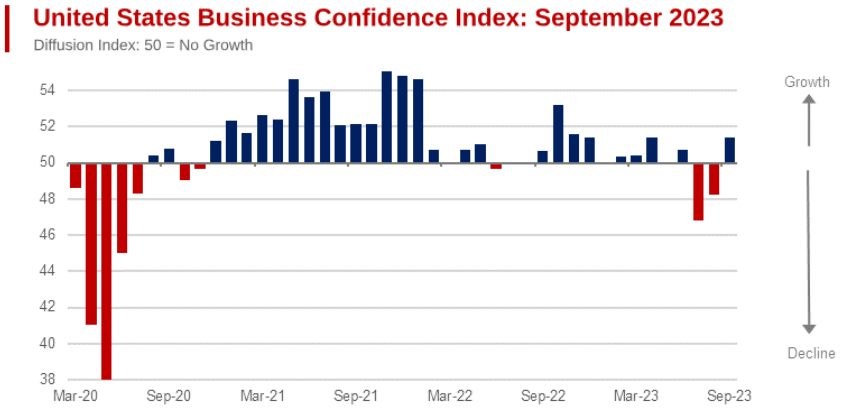
Fuente: Elaboración propia.

La relación entre la confianza del consumidor y los precios de las acciones se basa en varios supuestos clave:

1. **Impacto en el consumo:** Un aumento en la confianza del consumidor a menudo se correlaciona con un mayor gasto del consumidor, lo que puede impulsar los ingresos y las ganancias de las empresas. Esto, a su vez, puede contribuir al aumento de los precios de las acciones.

2. **Indicador adelantado:** La confianza del consumidor puede considerarse un indicador adelantado, ya que refleja las expectativas futuras. Un aumento en la confianza puede sugerir un optimismo sobre la economía y el mercado de valores.

Ilustración 16. Índice de confianza de Estados Unidos

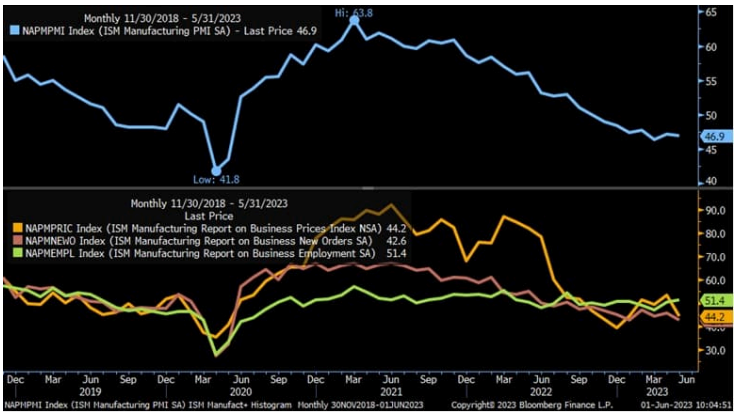


Fuente: Elaboración propia.

Índices de producción

Los índices de producción, como el índice de producción industrial (IPI) y el índice de gerentes de compras (PMI), se utilizan para evaluar el crecimiento económico y la salud de una economía. Estos índices proporcionan una instantánea de la actividad de las empresas manufactureras y de servicios, la producción, el empleo y la demanda de bienes. Históricamente, los inversionistas han considerado estos índices como indicadores líderes de la dirección de una economía, y por lo tanto, del rendimiento futuro de las empresas y de los precios de las acciones (Stock, 2019).

Ilustración 17. Índice PMI



Fuente: Elaboración propia.

Cuando el índice de producción está aumentando, indica una mayor producción industrial y expansión económica. Por el contrario, una disminución en el índice de producción puede indicar una producción industrial reducida y una contracción económica.

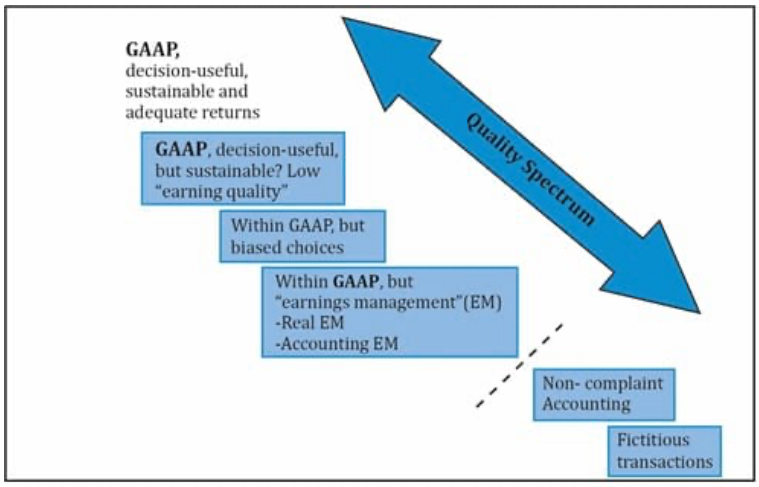
El índice de producción puede ser un indicador esencial para evaluar el ciclo económico y los posibles puntos de inflexión en el ciclo económico, así como el impacto de los índices de producción en sectores y empresas específicas. Los índices de producción se desglosan por sectores, como la manufactura, la construcción y los servicios. Los inversionistas pueden utilizar esta descomposición para identificar oportunidades y riesgos en sectores individuales. Por ejemplo, un aumento en el IPI manufacturero puede ser una señal positiva para las acciones de empresas manufactureras, mientras que un PMI de servicios más débil podría afectar negativamente las acciones de empresas de servicios (Stock, 2019).

Estados financieros

El análisis fundamental es una técnica utilizada por los inversionistas para evaluar la salud financiera de una empresa y su valor intrínseco. Esta técnica se basa en el estudio de los estados financieros de la empresa, que incluyen el balance, el estado de resultados y el flujo de efectivo. El análisis de estos estados proporciona una visión integral de la empresa, su rentabilidad, su solidez financiera y su capacidad para generar flujos de efectivo futuros (Brealey, 2017).

Los estados financieros (balances, estados de ingresos y estados de flujo de efectivo) brindan información detallada sobre el desempeño financiero, la rentabilidad, la liquidez y la salud general de una empresa. Al analizar estos informes, los inversionistas pueden obtener información sobre la salud financiera, el crecimiento de las utilidades y la generación de flujo de caja de una empresa, lo que puede tener un impacto directo en el rendimiento de sus acciones y, a su vez, en los precios futuros del índice general.

Ilustración 18. Calidad de los estados financieros



Fuente: Elaboración propia.

Los índices financieros clave, tales como la relación precio-utilidad (RPG en español o PER en inglés), la relación precio-valor en libros (P/VL en español o P/BV en inglés) y la relación deuda-capital, derivadas de los estados financieros, también se pueden utilizar como indicadores muy importantes a la hora de estimar la valoración, las perspectivas de crecimiento y la salud financiera de una empresa (Brealey, 2017).

Además, los estados financieros pueden revelar eventos y cambios importantes dentro de una empresa, tales como dividendos, fusiones y adquisiciones, y cambios en la alta dirección. Estos eventos pueden tener un impacto significativo en la confianza de los inversionistas y las reacciones del mercado, lo que influye en el rendimiento de los precios futuros del índice.

Data mining

El *data mining* es un proceso que involucra la identificación y la recopilación de datos, el preprocesamiento de datos, la selección de técnicas de *data mining,* el modelado y la evaluación, así como la interpretación y la aplicación de los resultados.

A continuación, se explica y se muestra gráficamente cada uno de los pasos implicados en este proceso:

- **Selección de conjunto de datos:** El proceso de *data mining* comienza con la identificación y la recopilación de datos relevantes para el análisis. Los datos pueden provenir de diversas fuentes, como bases de datos, registros transaccionales, sensores, redes sociales, entre otros. La calidad y la integridad de los datos son fundamentales para obtener resultados significativos.

- **Preprocesamiento de datos:** Antes de aplicar técnicas de *data mining* es necesario realizar tareas de preprocesamiento para limpiar, transformar y preparar los datos. Esto puede incluir la eliminación de valores atípicos, el manejo de datos faltantes, la normalización de variables y la selección de características relevantes. El preprocesamiento garantiza que los datos estén en un formato adecuado para su análisis.

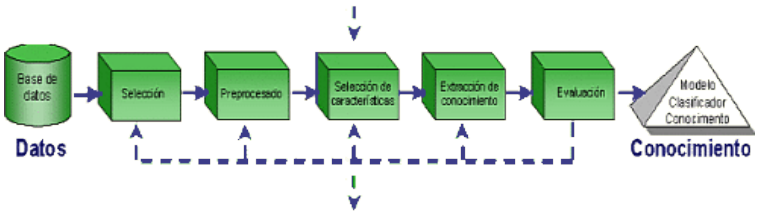
- **Selección de técnicas de *data mining:*** Existen diversas técnicas y algoritmos de *data mining* que se utilizan según los objetivos del análisis y la naturaleza de los datos. Entre las técnicas comunes se encuentran la clasificación, el *clustering,* la regresión, la asociación de reglas, la detección de anomalías y la minería de texto. Cada técnica tiene su enfoque particular y puede revelar diferentes tipos de conocimientos.

- **Modelado y evaluación:** Una vez seleccionada la técnica adecuada, se procede a construir un modelo de *data mining* utilizando los datos preprocesados. Esto implica aplicar el algoritmo elegido al conjunto de datos y generar un modelo que represente los patrones y las relaciones descubiertas. Posteriormente, se evalúan la calidad y el rendimiento del modelo utilizando métricas apropiadas, como la precisión, la sensibilidad, la especificidad o el error cuadrático medio, dependiendo del tipo de

problema.

- **Interpretación y aplicación de resultados:** El último paso en el proceso de *data mining* es interpretar los resultados obtenidos y aplicarlos para tomar decisiones o resolver problemas. Esto implica comprender y comunicar los conocimientos descubiertos de manera clara y significativa. Los resultados del *data mining* pueden utilizarse en una amplia gama de áreas, como el *marketing,* las finanzas, la salud y la seguridad, entre otras, para mejorar la toma de decisiones y obtener ventajas competitivas.

Ilustración 19. *Data mining*



Fuente: Antut (s. f.).

Obtención de datos

Existen dos tipos principales de datos que se pueden obtener:

- **Datos en lote o *batch:*** Estos son datos que ya están almacenados y a los que se puede acceder en su totalidad. Los datos en lote o *batch* pueden provenir de diversas fuentes, como archivos en hojas electrónicas (como Excel) o archivos CSV *(comma separated values)* que contienen datos estructurados. Además, se pueden encontrar datos semiestructurados, como archivos en formato JSON (JavaScript Object Notation) o XML (eXtensible Markup Language), que contienen información organizada en una estructura jerárquica. Asimismo, los datos en formato de texto plano, como noticias, publicaciones o documentos, pueden ser considerados datos en lote.

- **Datos en *streaming:*** Se generan y transmiten en tiempo real a medida que ocurren. Los datos en *streaming* pueden provenir de diversas fuentes, como sensores,

dispositivos IoT *(internet of things),* redes sociales, transmisiones en vivo, entre otros. Estos datos son continuos y fluyen en tiempo real, lo que requiere técnicas y herramientas especiales para su captura y procesamiento en tiempo real.

La obtención de datos en lote implica acceder y recopilar los datos almacenados de manera completa antes de comenzar el análisis. Por otro lado, la obtención de datos en *streaming* implica la captura continua de datos en tiempo real a medida que se generan.

Es importante destacar que, independientemente del tipo de datos que se esté obteniendo, es esencial considerar la calidad de los datos, la privacidad y la seguridad, así como asegurarse de que se cumplan los requisitos legales y éticos en cuanto a la recopilación y el uso de los datos.

Una vez que se han obtenido los datos, se pueden realizar tareas de preprocesamiento, como la limpieza de datos, la normalización, la eliminación de valores atípicos y la integración de diferentes fuentes de datos, para prepararlos adecuadamente antes de aplicar las técnicas de *data mining.*

Procesamiento de datos

Una de las etapas más importantes es el procesamiento de datos. Esto implica la transformación y la limpieza de los datos antes de realizar cualquier análisis o modelado. En esta etapa es común encontrarse con datos vacíos, omisos o no relevantes que pueden afectar la calidad y la validez de los resultados.

Respecto a los datos omisos o nulos, hay diferentes enfoques y opiniones sobre cómo tratarlos. Algunos autores sugieren eliminar los registros que contienen datos faltantes, ya que esto puede evitar introducir sesgos o distorsiones en los análisis. Al eliminar estos registros, se garantiza que solo se utilicen datos completos y confiables (Roiger, 2017).

Por otro lado, otros autores argumentan que es posible llenar los valores faltantes utilizando técnicas de imputación. Una opción común es reemplazar los datos faltantes con la media de los valores existentes en la misma variable. Esta técnica es útil cuando la variable en cuestión tiene una distribución aproximadamente normal y los datos faltantes son aleatorios.

Sin embargo, es importante tener en cuenta que la imputación de datos introduce cierto grado de incertidumbre y puede afectar los resultados del análisis. La elección de la estrategia de manejo de datos faltantes depende del contexto específico y de la naturaleza de los datos.

Modelación

En la modelación de datos se distinguen dos tipos principales: supervisada y no supervisada. En la **modelación supervisada** se cuenta con un conjunto de datos de entrenamiento que incluye atributos de entrada y de salida. El objetivo es entrenar un modelo para que aprenda la relación entre los atributos de entrada y las salidas conocidas, de tal manera que pueda predecir las salidas para nuevos datos de entrada. Por otro lado, en la **modelación no supervisada** solo se dispone de atributos de entrada y no hay salidas conocidas. El propósito es descubrir patrones o estructuras ocultas en los datos, sin ninguna guía externa. En cuanto a la diferencia entre clasificación y regresión, esta radica en el tipo de salida que se espera del modelo. En la clasificación se asigna una etiqueta o una categoría a las muestras de datos, mientras que en la regresión se busca predecir un valor numérico continuo.

A continuación, se mostrarán algunos algoritmos comúnmente utilizados en la modelación de datos, como los utilizados en la presente investigación.

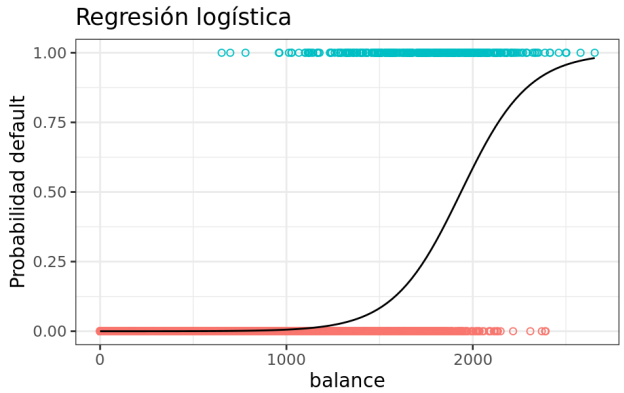
Regresión logística

La **regresión logística simple,** desarrollada por David Cox en 1958, es un método estadístico utilizado para modelar la relación entre una variable cualitativa binaria y una variable cuantitativa. Se utiliza para estimar la probabilidad de que la variable cualitativa tome un valor específico en función de los valores de la variable cuantitativa.

Una de las aplicaciones más comunes de la regresión logística es la clasificación binaria. En este caso, las observaciones se clasifican en dos grupos distintos con base en la variable predictora. Por ejemplo, se puede utilizar la regresión logística para clasificar a un individuo desconocido como hombre o mujer en función del tamaño de la mandíbula. La regresión logística utiliza una función logística para modelar la probabilidad de pertenencia a cada una de las categorías. El resultado de la regresión logística es una estimación de la probabilidad de pertenencia a la categoría positiva, lo que permite realizar la clasificación (Rodrigo, 2016).

Es importante tener en cuenta que la regresión logística no se limita únicamente a problemas de clasificación binaria, también se puede extender a casos de clasificación multiclase. Además, la regresión logística puede utilizarse con múltiples variables predictoras, lo que se conoce como regresión logística múltiple. En resumen, la regresión logística es una técnica versátil y ampliamente utilizada en el campo de la estadística y el aprendizaje automático para abordar problemas de clasificación binaria y estimar probabilidades de pertenencia a una categoría determinada.

Ilustración 20. Representación de regresión logística



Fuente: Rodrigo (2016).

Análisis discriminante lineal

El análisis discriminante lineal (LDA por sus siglas en inglés, *linear discriminant analysis*) es un método de clasificación supervisado utilizado para categorizar variables cualitativas. Con este método se conocen de antemano dos o más grupos y se desea clasificar nuevas observaciones en uno de estos grupos en función de sus características.

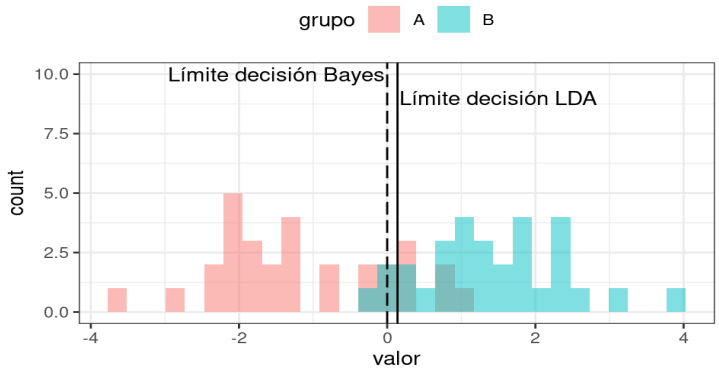
El LDA utiliza el teorema de Bayes para estimar la probabilidad de que una observación pertenezca a cada una de las clases de la variable cualitativa, dado un determinado valor de los predictores. Es decir, se calcula la probabilidad condicional de que la observación

pertenezca a cada grupo con base en sus características. Posteriormente, se asigna la observación al grupo para el cual la probabilidad predicha es mayor.

El objetivo principal del LDA es encontrar una combinación lineal de las variables predictoras que maximice la separación entre los grupos conocidos. Esto se logra al maximizar la razón de las varianzas entre grupos y minimizar la varianza dentro de los grupos. De esta manera, se busca encontrar una función discriminante que permita clasificar correctamente las observaciones en los grupos conocidos.

El LDA es ampliamente utilizado en diversas áreas, como el reconocimiento de patrones, el análisis de imágenes y la bioinformática, entre otras. Es una técnica efectiva cuando los grupos están bien separados y las variables predictoras cumplen con ciertas suposiciones, como una distribución normal multivariada y una matriz de covarianza común entre los grupos (Rodrigo, 2016).

Ilustración 21. LDA



Fuente: Rodrigo (2016).

Support vector machines

El método de clasificación-regresión máquinas de vector soporte (SVM, por sus siglas en inglés) es un algoritmo ampliamente utilizado en el campo del aprendizaje automático y la ciencia de datos. Fue desarrollado a finales de la década de los noventa y originalmente se

concibió como un método de clasificación binaria, pero luego se extendió para abordar problemas de clasificación múltiple y regresión (Rodrigo, 2017).

Las SVM se consideran uno de los mejores clasificadores disponibles para una amplia variedad de situaciones, y son ampliamente reconocidas en el ámbito del aprendizaje estadístico y el ML (Das y Padhy, 2012).

La base teórica de las SVM se encuentra en el *maximal margin classifier,* que a su vez se basa en el concepto de hiperplano. En resumen, las SVM están pensadas para encontrar el hiperplano que mejor separa las clases en un espacio de características de alta dimensionalidad. El objetivo es maximizar el margen entre las clases, lo que se traduce en una mejor capacidad de generalización del modelo (Wei, 2012).

En el contexto de los mercados financieros, las SVM también han demostrado ser eficaces. Se han utilizado para predecir series de tiempo en situaciones en las que los métodos clásicos no son adecuados debido a la falta de estacionariedad de las variables o la complejidad de las series de tiempo.

En el ámbito financiero, las SVM se clasifican en SVM y SVR. Las SVM se utilizan para resolver problemas de clasificación, mientras que las SVR *(support vector regression)* son un tipo de SVM que se utiliza para predecir valores numéricos, como precios futuros.

La ventaja de las SVM en el proceso de predicción es su capacidad para eliminar datos irrelevantes y dispersos, lo que puede mejorar la precisión de la predicción. Esto se logra mediante la identificación y el uso de vectores de soporte, que son los puntos de datos más relevantes para la separación de clases.

En términos de implementación, las SVM resuelven el problema que implica encontrar el hiperplano óptimo utilizando métodos de programación cuadrática, que son técnicas conocidas para resolver problemas restrictivos. Antes de aplicar la clasificación lineal, los datos se transforman mediante una función *phi* a un espacio de características de mayor dimensionalidad, lo que permite a la SVM clasificar datos altamente complejos (Han, Pei y Kamber, 2011).

Red neuronal artificial

El concepto de redes neuronales artificiales surgió por primera vez en 1943, con la propuesta de McCulloch y Pitts. Sin embargo, fue con el desarrollo del algoritmo de retropropagación por parte de McClelland, Rumelhart y Hinton en 1986 que las redes neuronales experimentaron un avance significativo. El algoritmo de retropropagación posibilitó el desarrollo de redes neuronales *feedforward* multicapa, como el perceptrón multicapa (MLP).

Las redes neuronales tienen una amplia gama de aplicaciones en áreas financieras y de inversión. Se utilizan en la predicción de quiebras, la toma de decisiones y la planificación financiera, entre otros procedimientos (Haykin, 1999).

En una red neuronal *feedforward,* los nodos se organizan en capas consecutivas con una relación unidireccional. Cuando se presenta un patrón de entrada a la red, la primera capa calcula su salida y la transmite a la siguiente capa. Este proceso continúa capa por capa hasta llegar a la capa de salida. El algoritmo de retropropagación o *backpropagation* es utilizado para ajustar los parámetros de la red de manera que se minimice la diferencia entre la salida de la red y el valor real.

El aprendizaje en las redes neuronales se logra equilibrando las ponderaciones de las conexiones entre las neuronas con el objetivo de minimizar una función de error. Esto implica ajustar las ponderaciones iniciales de manera iterativa a medida que se presentan los datos de entrenamiento. El equilibrio de las ponderaciones permite a la red mejorar su capacidad de hacer predicciones más precisas.

En la mayoría de las redes neuronales las conexiones entre las neuronas están estructuradas de tal manera que las neuronas en capas intermedias reciben entradas de todas las neuronas en la capa anterior. Esto permite que las señales se propaguen desde la capa de entrada hacia las capas superiores, hasta llegar a la capa de salida, en un proceso conocido como alimentación hacia adelante o *feedforward* (Lawrence, 1997; Tan, 2009).

1. **Redes neuronales artificiales (ANN):** Técnica de ML basada en la estructura y el funcionamiento del cerebro humano. Las redes neuronales profundas (DNN) son una

extensión de las ANN que utilizan múltiples capas ocultas para aprender características más complejas y abstractas de los datos.

2. **Redes neuronales recurrentes (RNN):** Variante de ANN diseñada para procesar secuencias temporales de datos. Las RNN son especialmente útiles para predecir series temporales como los precios y la volatilidad de activos financieros. La memoria a largo plazo (LSTM) y las redes de memoria a corto y largo plazo (GRU) son dos tipos populares de RNN que han demostrado ser efectivas en tareas de predicción de series temporales.

3. **Redes neuronales convolucionales (CNN):** Aunque las CNN se utilizan comúnmente para el análisis de imágenes, también pueden aplicarse a series temporales y datos financieros. Las CNN pueden capturar patrones locales y relaciones temporales en los datos, lo que las convierte en una opción viable para predecir los precios y la volatilidad.

4. **Redes neuronales LSTM:** A mediados de la década de los noventa, los investigadores alemanes Sepp Hochreiter y Juergen Schmidhuber propusieron una variación de la red recurrente con las llamadas unidades de memoria a corto y largo plazo, o LSTM, como una solución al problema del gradiente de desaparición.

LSTM

Todos sabemos que una red neuronal usa un algoritmo llamado *backpropagation* (BP) para actualizar los pesos de la red. Primero calcula los gradientes del error usando la regla de la cadena del cálculo, para luego actualizar los pesos (descenso de gradiente).

Debido a que el BP comienza desde la capa de salida hasta la capa de entrada, es posible que en una red neuronal simple no tengamos problemas para actualizar los pesos, pero en una red neuronal profunda podríamos enfrentar algunos problemas.

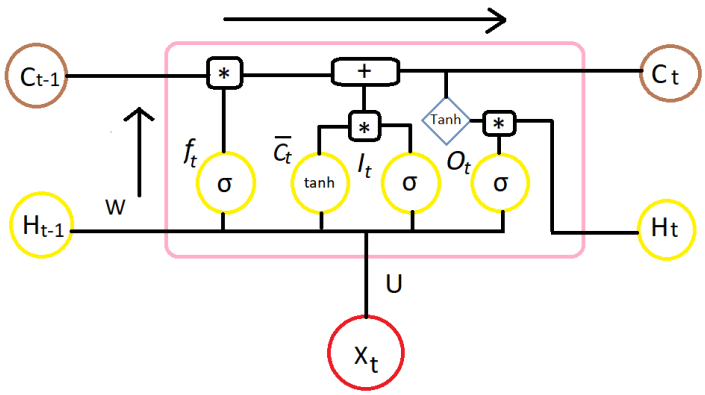
Las redes LSTM *(long short-term memory)* se denominan redes neuronales recurrentes de lujo con algunas características adicionales.

Entonces, el modelo LSTM contiene los siguientes componentes:

1. *Forget gate* ***"f"*** *(a neural network with sigmoid).*
2. *Candidate layer* ***"$\bar{C}$"*** *(a NN with tanh).*
3. *Input gate* ***"I"*** *(a NN with sigmoid).*
4. *Output gate* ***"O"*** *(a NN with sigmoid).*
5. *Hidden state* ***"H"*** *(a vector).*
6. *Memory state* ***"C"*** *(a vector).*

Aquí está el diagrama de la celda LSTM en el paso de tiempo *t:*

Ilustración 22. LSTM



Fuente: Zhuge *et al.* (2009).

Las entradas del modelo LSTM son las siguientes:

$$x_t = Entrada\ actual$$

$$H_{t-1} = Salida\ del\ modelo\ anterior$$

$$C_{t-1} = Memoria\ del\ modelo\ anterior$$

Las salidas del modelo son las siguientes:

$$H_t = Salida\ del\ modelo\ actual$$

$$C_t = Memoria\ del\ modelo\ actual$$

Ecuaciones:

$f_t = \sigma(x_t * U_f + H_{t-1} * W_f)$

$\bar{C}_t = tanh(x_t * U_c + H_{t-1} * W_c)$

$I_t = \sigma(x_t * U_i + H_{t-1} * W_i)$

$O_t = \sigma(x_t * U_o + H_{t-1} * W_o)$

$C_t = f_t * C_{t-1} + I_t * \bar{C}_t$

$$H_t = O_t * \tanh(C_{t)}$$

Para *forget gate (f), memory state (c), input gate (I)* y *output gate (O)* se producen vectores (entre 0 y 1 para Sigmoide, -1 a 1 para *tanh*) por lo que obtenemos cuatro vectores *f, $\bar{C}$, I, O* para cada paso de tiempo.

El modelo LSTM toma el estado de memoria anterior C_{t-1} y hace una multiplicación de elementos con *forget gate (f).*

$$C_t = f_t * C_{t-1}$$

- Si el valor de f_t es 0, el C_{t-1} se olvida por completo.
- Si el valor de f_t es 1, el C_{t-1} se pasa por completo al modelo.

Ahora, con el estado actual de la memoria C_t, calculamos el nuevo estado de la memoria a partir del estado de entrada y la capa C.

$$C_t = C_t + (I_t * \bar{C}_t)$$

C_t es el estado actual de la memoria en el paso de tiempo *t* y pasa al siguiente paso de tiempo.

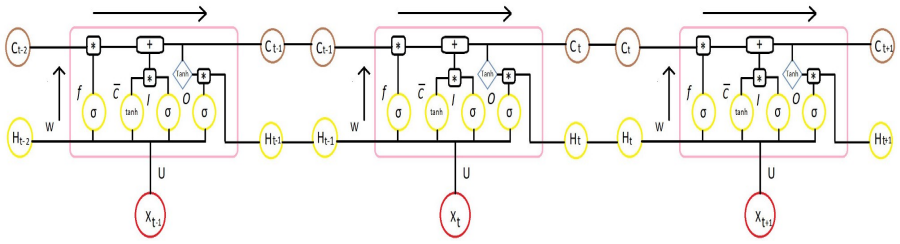
Finalmente, necesitamos calcular lo que vamos a producir. Esta salida se basará en nuestro estado de celda C_t, pero será una versión filtrada, así que aplicamos tanh a C_t, y luego hacemos una multiplicación de elementos con la puerta de salida O. Ese será nuestro estado oculto actual H_t.

$$H_t = \tanh(C_t)$$

Pasamos estos dos C_t y H_t al siguiente paso de tiempo y repetimos el mismo proceso.

Aquí podemos ver un diagrama que muestra el procedimiento del modelo LSTM para diferentes pasos de tiempo *t*.

Ilustración 23. LSTM para varios *t*.



Fuente: Zhuge *et al.* (2009).

5. **Evaluación e interpretación de los resultados**

En la etapa final de un modelo predictivo es importante evaluar su capacidad para predecir la variable objetivo. Para ello, se utilizan diferentes elementos de validación que nos permiten medir la calidad del modelo y su capacidad de generalización. A continuación se definen algunos de estos elementos (Rodrigo, 2020):

- ***Set* de entrenamiento:** Es el conjunto de datos u observaciones que se utiliza para entrenar y construir el modelo estadístico. Este conjunto de datos contiene tanto las variables predictoras como la variable respuesta. Se utiliza para ajustar los parámetros

del modelo y encontrar la mejor configuración (Rodrigo, 2020).

- ***Set* de validación:** Es un conjunto de datos separado del *set* de entrenamiento, que tiene las mismas características y la misma distribución, pero no se utiliza durante el proceso de entrenamiento del modelo. Estos datos son "nuevos" para el modelo, ya que no ha sido expuesto a ellos previamente. El *set* de validación se utiliza para evaluar el rendimiento del modelo y medir su capacidad de generalización (Rodrigo, 2020).

- ***Training error rate:*** Error promedio que se obtiene al utilizar el modelo para predecir las observaciones del *set* de entrenamiento. Es una medida del ajuste del modelo a los datos de entrenamiento. Sin embargo, el *training error rate* puede no ser una estimación precisa del rendimiento del modelo en datos no vistos, ya que el modelo ha sido diseñado específicamente para ajustarse a estos datos durante el entrenamiento (Rodrigo, 2020).

- ***Test error rate:*** Error promedio que se obtiene al utilizar el modelo para predecir nuevas observaciones que no formaron parte del *set* de entrenamiento. Estas observaciones son consideradas datos "no vistos" por el modelo. El *test error rate* es una medida más confiable del rendimiento del modelo en situaciones de uso real, ya que refleja su capacidad para generalizar y hacer predicciones precisas en datos no vistos (Rodrigo, 2020).

Las técnicas de validación cruzada, o *cross-validation,* son ampliamente utilizadas para estimar el error de prueba de un modelo y evaluar su capacidad predictiva. Algunas de las técnicas más comunes de validación cruzada son las siguientes:

- **Validación simple:** Es el método más sencillo y consiste en dividir aleatoriamente las observaciones disponibles en dos grupos: uno se utiliza para entrenar el modelo y el otro se utiliza para evaluarlo (Kulesa, 2015). El principal problema de este método es que el error de prueba puede ser altamente variable, ya que depende de la cantidad de observaciones asignadas a los conjuntos de entrenamiento y prueba.

- **LOOCV** ***(leave-one-out cross-alidation):*** Este método se basa en un enfoque iterativo en el que se excluye una observación a la vez y se utiliza el resto de las observaciones para entrenar el modelo. Luego, se calcula el error utilizando la observación excluida. Este proceso se repite para cada observación y el error de prueba estimado por LOOCV es el promedio de todos los errores calculados (Rodrigo, 2020). La principal desventaja de este método es su alto costo computacional, ya que requiere ajustar y evaluar el modelo tantas veces como observaciones haya en el conjunto de datos (Kulesa, 2015).

- ***k-fold cross-validation:*** Este método también es iterativo y se basa en dividir los datos en k grupos de tamaños similares. Luego, se entrena el modelo utilizando k-1 grupos como conjunto de entrenamiento, y se evalúa utilizando el grupo restante como conjunto de prueba. Este proceso se repite k veces, utilizando un grupo diferente como conjunto de prueba en cada iteración. El error de prueba estimado por *k-fold cross-validation* es el promedio de los errores obtenidos en cada iteración. Este método es menos costoso computacionalmente que LOOCV, y se considera una técnica de validación cruzada comúnmente utilizada (Kulesa, 2015).

El RMSE es una medida de dispersión que indica cuán cerca están las predicciones del modelo de los valores reales. Al elevar al cuadrado los errores antes de promediarlos, se da más peso a los errores grandes, lo que hace que el RMSE sea más sensible a los valores atípicos. Al tomar la raíz cuadrada, la métrica vuelve a la misma escala que los datos originales, lo que facilita la interpretación.

El RMSE se utiliza en una variedad de aplicaciones, como pronósticos de series de tiempo, predicción de precios, evaluación de modelos de regresión, entre otros. Proporciona una medida de qué tan bien se ajusta el modelo a los datos y es útil para comparar diferentes modelos o para evaluar la mejora de un modelo en relación con otro.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(S_i - O_i)^2}$$

Desde una perspectiva matemática, el RMSE se asemeja a la fórmula de la distancia euclidiana entre dos vectores en un espacio euclidiano de dos dimensiones (R^2). La fórmula del RMSE calcula la raíz cuadrada de la media de los errores al cuadrado, lo cual es similar al cálculo de la distancia euclidiana entre dos puntos (Kenney, 1962).

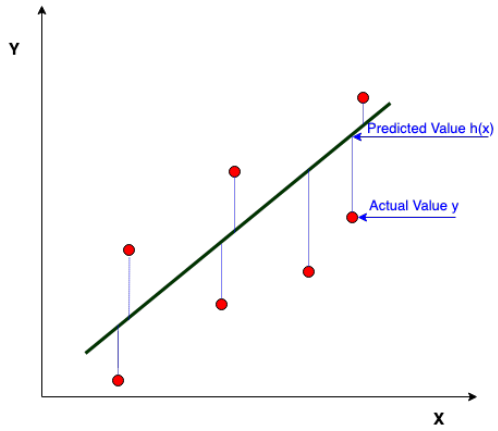
Si consideramos las observaciones reales como un vector *O* y los valores predichos como un vector *S*, podemos interpretar el RMSE como una medida de la distancia entre estos dos vectores en un espacio de dos dimensiones. Al elevar al cuadrado los errores y luego promediarlos, estamos enfatizando la magnitud de los errores y su influencia en la distancia entre los dos vectores.

Sin embargo, es importante tener en cuenta que esta interpretación matemática del RMSE tiene sus limitaciones. El RMSE es sensible a la escala de los datos, lo que significa que si los datos tienen diferentes escalas, el RMSE puede estar sesgado hacia las variables con mayor escala. Por lo tanto, al comparar errores entre diferentes variables, es recomendable utilizar otras medidas de error que sean más adecuadas para tener en cuenta las diferencias de escala, como el error porcentual absoluto medio (MAPE) o el error absoluto medio (MAE).

$$distancia(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

Esto indica heurísticamente que RMSE puede considerarse como una especie de distancia (normalizada) entre el vector de valores predichos y el vector de valores observados.

Ilustración 24. RMSE



Fuente: Kenney (1962).

En el aprendizaje automático es útil tener una medida única para evaluar el rendimiento de un modelo. El error cuadrático medio (RMSE) es una de las medidas más utilizadas y populares en este sentido. Es una regla de puntuación adecuada que proporciona una forma intuitiva de comprender el rendimiento del modelo, y es compatible con supuestos estadísticos comunes.

El RMSE calcula la raíz cuadrada de la media de los errores al cuadrado, lo que significa que penaliza de manera más significativa los errores grandes. Esto es útil en muchos escenarios, ya que los errores grandes pueden tener un impacto desproporcionado en la precisión general del modelo.

Además, el RMSE es compatible con algunos supuestos estadísticos comunes, como el supuesto de normalidad de los errores. Si se asume que los errores siguen una distribución normal, el RMSE puede ser interpretado como una medida de dispersión alrededor de los valores predichos. Esto puede ser útil para evaluar si el modelo se ajusta adecuadamente a los datos y si los errores se distribuyen de manera razonable.

Sin embargo, es importante tener en cuenta que el RMSE también tiene limitaciones. Como se mencionó anteriormente, es sensible a la escala de los datos, lo que puede dificultar la

comparación de errores entre variables con diferentes escalas. Además, el RMSE puede verse afectado por valores atípicos en los datos.

Análisis de sentimientos

Esta metodología tuvo sus inicios hace mucho tiempo (Salton y McGill, 1983), aunque la categorización basada en sentimientos es un campo más reciente, introducido por Das y Chen en el 2001, Morinaga *et al.* En el 2002, Pang *et al.* en el 2002, Tong en el 2001, Turney en el 2002 y Wiebe en el 2000.

El enfoque tradicional para representar texto, conocido como el método de *bag of words* (BoW), se originó con Salton y McGill en 1983. Según este modelo BoW, un documento se representa como un vector en el espacio euclidiano, donde cada palabra se considera independiente de las demás. Este conjunto de palabras individuales se llama comúnmente colección de unigramas. El enfoque BoW es fácil de comprender y ha demostrado un alto rendimiento. Por ejemplo, los mejores resultados en la categorización de múltiples etiquetas para el conjunto de datos Reuters-21578 se obtuvieron utilizando el enfoque BoW, según Dumais *et al.* en 1998 y Weiss *et al.* en 1999.

Los dos principales métodos de análisis de sentimientos, el método basado en el léxico (un enfoque no supervisado) y el método basado en el aprendizaje automático (un enfoque supervisado) se basan en la Bolsa de Palabras. En el método supervisado de aprendizaje automático, los clasificadores utilizan los unigramas o sus combinaciones (N-gramas) como características. En el método basado en el léxico, los unigramas presentes en el léxico reciben una puntuación de polaridad, y la puntuación de polaridad general del texto se calcula sumando las polaridades de los unigramas.

Al considerar qué elementos del léxico de un mensaje deben incluirse en el análisis de los sentimientos, se han investigado diversas partes del discurso (Pak y Paroubek, 2010; Kouloumpis *et al.,* 2011). Benamara y sus colegas propusieron un enfoque de *combinaciones adverbio-adjetivo* (AAC) que utiliza adverbios y adjetivos para detectar la polaridad de los sentimientos (Benamara *et al.,* 2007). En los últimos años, se ha estudiado el papel de los emoticonos (Pozzi *et al.,* 2013a; Hogenboom *et al.,* 2013; Liu *et al.,* 2012; Zhao *et al.,* 2012). En su investigación más reciente, Fersini y su equipo exploraron el uso de adjetivos,

emoticonos, expresiones enfáticas y onomatopéyicas, así como el alargamiento expresivo, como señales expresivas en el análisis de sentimientos de microblogs. Demostraron que estas señales pueden enriquecer el conjunto de características y mejorar la calidad de la clasificación de sentimientos.

Se han desarrollado algoritmos avanzados para el análisis de sentimientos que consideran no solo el contenido del mensaje, sino también el contexto en el que se publica, el autor del mensaje, los amigos del autor y la estructura subyacente de la red. Por ejemplo, Hu y su equipo investigaron cómo las relaciones sociales pueden mejorar el análisis de sentimientos a través de un enfoque sociológico para manejar textos ruidosos y cortos (SANT) en el 2013. Zhu *et al.* demostraron que la calidad de la agrupación de sentimientos en X (anteriormente llamada Twitter) se puede mejorar al agrupar *tweets,* usuarios y funciones de manera conjunta en el 2014. En el trabajo de Pozzi y sus colegas en el 2013, se analizaron las conexiones de amistad y se estimaron las polaridades de los usuarios sobre un tema mediante la integración de los contenidos de las publicaciones con las relaciones de aprobación. You y Luo mejoraron la precisión de la clasificación de sentimientos al agregar contenido visual además de información textual en el 2013. Aisopos *et al.* aumentaron significativamente la precisión de la clasificación de sentimientos mediante el uso de características basadas en contenido junto con características basadas en el contexto en el 2012. Saiff *et al.* lograron mejoras al aumentar el espacio de características con características semánticas en el 2012.

Aunque muchos trabajos de investigación se centraron en identificar las mejores características, también se realizaron esfuerzos para explorar nuevos métodos de clasificación de sentimientos. Wang *et al.* evaluaron el rendimiento de métodos de conjunto como Bagging, Boosting y Random Subspace, y demostraron empíricamente que los modelos de conjunto pueden superar a los aprendices individuales en el 2014. Fersini *et al.* propusieron el uso del método de conjunto de promediado del modelo bayesiano, que superó tanto la clasificación tradicional como los métodos de conjunto en el 2014. Carvalho *et al.* emplearon algoritmos genéticos para encontrar subconjuntos de palabras paradigmáticas que mejoraron la precisión de la clasificación en el 2014.

Hoy en día, se habla mucho sobre el análisis de redes sociales, que implica la extracción de información de plataformas de medios sociales, como texto, redes, acciones, enlaces y

ubicación (Khan, 2015). El análisis de opiniones es un subcampo del análisis de redes sociales que se centra en el estudio de las opiniones y actitudes de las personas frente a productos, empresas, servicios, personas, entre otros (Liu, 2015).

Las opiniones constan de cinco componentes:

1. **Objetivo:** El sujeto o tema de la opinión.
2. **Atributo:** Lo que se menciona sobre el objetivo.
3. **Sentimiento:** Si la opinión es positiva, negativa o neutral.
4. **Poseedor:** La persona que emite la opinión.
5. **Tiempo:** El momento en que se emite la opinión.

Por ejemplo, en el siguiente *tweet* del 6 de enero del 2021, @Boavenossa escribió: "Están subiendo un 20 % las acciones de Smith & Wesson, es oportunidad de comprar o no vender". En este caso, el objetivo es Smith & Wesson, el atributo es el precio de las acciones, el sentimiento es positivo, el poseedor es Boavenossa Chacinated @Boavenossa, y el tiempo es el 6 de enero del 2021.

El análisis de sentimientos categoriza las opiniones en rangos de -1 a 1, representando negativo, neutral y positivo respectivamente, para cuantificar la intensidad de la polarización (Pawar, 2016). Aunque identificar estos componentes en una opinión puede ser fácil para las personas, resulta desafiante para los algoritmos debido a consideraciones como la polisemia, el sarcasmo, la negación, los sinónimos y la subjetividad.

Las principales herramientas para el análisis de sentimientos son Freeling y OpenNPL, que favorecen tareas como la segmentación de oraciones, el etiquetado y la extracción de entidades (Padró, 2012) (Community, 2011). En el ámbito de las redes sociales, la herramienta más utilizada en la actualidad es VADER, que ha demostrado producir excelentes resultados en X (Hutto, 2014).

Python

Python es un lenguaje de programación conocido por su simplicidad y por la facilidad para el aprendizaje, pero a la vez es poderoso y versátil. Se utiliza para procesar diversos tipos de datos, ya sean textuales o numéricos. Una de las ventajas más destacadas de Python es su licencia de código abierto, que permite a los desarrolladores crear y compartir bibliotecas con la comunidad. En el ámbito del desarrollo de aplicaciones de aprendizaje automático se utilizan varias bibliotecas de Python que desempeñan un papel fundamental. Las siguientes son algunas de estas bibliotecas clave:

1. **Scikit-Learn:** Es una biblioteca ampliamente utilizada en el aprendizaje automático. Ofrece herramientas para el análisis y la minería de datos, y se basa en otras bibliotecas de Python, como NumPy, SciPy y Matplotlib. Proporciona una amplia gama de algoritmos de aprendizaje supervisado y no supervisado, lo que la convierte en una elección popular para desarrolladores y científicos de datos (InfoWorld, 2017).

2. **NLTK *(natural language toolkit):*** NLTK es una biblioteca especializada en procesamiento de lenguaje natural. Se utiliza para crear vocabularios, procesar texto, etiquetar palabras y analizar oraciones. Es ampliamente utilizada en tareas de análisis de sentimientos y procesamiento de texto en el aprendizaje automático (NLTK, 2017).

3. **NumPy:** Es una biblioteca esencial para el procesamiento de matrices y operaciones matemáticas de alto rendimiento. Es fundamental para tareas como el álgebra lineal, las transformadas de Fourier y la manipulación de datos multidimensionales. También se utiliza como base para otras bibliotecas de aprendizaje automático, como TensorFlow (InfoWorld, 2017).

4. **SciPy:** Es una biblioteca que se utiliza para la computación científica y el análisis de datos. Ofrece módulos para una amplia variedad de tareas, como el álgebra lineal, la optimización y el procesamiento de señales. SciPy se basa en NumPy y proporciona herramientas adicionales para tareas numéricas y científicas (InfoWorld, 2017).

5. **TensorFlow:** Es una biblioteca desarrollada por Google que se utiliza para crear modelos de aprendizaje automático, especialmente redes neuronales. Puede aprovechar la potencia de las unidades de procesamiento de gráficos (GPU) y las unidades de procesamiento tensorial (TPU) para acelerar el entrenamiento de modelos. Es ampliamente utilizada en el campo del aprendizaje profundo (InfoWorld, 2017).

6. **Keras:** Es una biblioteca de alto nivel para la construcción y el entrenamiento de redes neuronales. Se integra bien con TensorFlow y otras bibliotecas de aprendizaje automático. Keras es conocida por la facilidad de uso y su capacidad para crear modelos de manera rápida y sencilla (InfoWorld, 2017).

X (antes Twitter)

X es una plataforma de redes sociales que posibilita la comunicación bidireccional, lo que significa que puedes interactuar con otros usuarios compartiendo información de manera rápida y gratuita. Con más de 300 millones de usuarios activos en todo el mundo, X se utiliza tanto a través de su aplicación móvil como en su sitio web (Twitter, s. f.).

Registrarse en X es un proceso gratuito y es el primer paso para empezar a publicar *tweets.* Una vez que tienes una cuenta, puedes seguir a otros usuarios y, a la vez, ser seguido por ellos. Es importante destacar que tus publicaciones solo son visibles para las personas que te siguen. Además de los usuarios individuales, en X también pueden registrarse empresas, entidades gubernamentales, marcas y productos, entre otros.

Cuando creas tu cuenta, debes elegir un nombre de usuario que te representará en la red social, precedido por el símbolo "@" (por ejemplo, @username). Para llegar a un público más amplio, puedes utilizar *hashtags* relacionados con temas específicos. Los *tweets* pueden contener texto plano, enlaces, menciones y otros elementos.

En el ámbito del mercado bursátil se utilizan *hashtags* que corresponden a los símbolos utilizados por las compañías en la bolsa de valores, precedidos por el símbolo "$". Por ejemplo, $APPL representa a Apple.

Ilustración 25. *Tweet* relacionado con Apple

Fuente: Elaboración propia.

X proporciona herramientas que permiten buscar *tweets* bajo ciertos parámetros, como usuario, fecha, ubicación y *hashtags*, entre otros. Estas herramientas incluyen la API REST, que recupera *tweets* publicados en el pasado, y la API Streaming, que permite acceder a *tweets* en tiempo real (Twitter, s. f.). Otra herramienta comúnmente utilizada es TwiQuery, que ofrece la posibilidad de descargar información utilizando los mismos parámetros que las API de Twitter. TwiQuery se destaca por su capacidad para identificar emoticonos, lo que facilita el análisis de los sentimientos expresados en la red social (Jabreel *et al.,* 2016).

La herramienta TwiQuery desempeñará un papel fundamental en el desarrollo del proyecto, ya que se utilizará para obtener datos relacionados con las compañías estudiadas en el mercado bursátil. Además, se planea agregar una nueva variable llamada *sentimiento* a los modelos de aprendizaje automático, con el objetivo de capturar los sentimientos expresados por los usuarios de X con respecto a estas compañías.

I. Metodología

Para el desarrollo del modelo propuesto, se pone en práctica la siguiente metodología:

1. **Selección de variables**

   - **Datos de mercado:**
     - Retornos pasados.
     - Curvas de mercado de derivados sobre estos.
     - Índice de la volatilidad del S&P500 (VIX).
   - **Datos macro:**
     - Producto interno bruto (PIB).
     - Inflación.
     - Desempleo.
     - Indicadores de confianza empresarial y del consumidor.
     - Índices de producción.
   - **Información financiera:**
     - Balances.
     - Estados de resultados.
     - Flujos de efectivo y datos de deuda. Meterlos como múltiplos.
     - También puede incluir eventos corporativos como dividendos, fusiones y adquisiciones, y cambios en la dirección.
   - **Data no estructurada**: Como *tweets* para realizarles análisis de sentimientos. Se sacará la información de la API de X.

2. Validación de significancia de las variables seleccionadas

Una vez recolectada la información para cada variable según la periodicidad y el lapso propuesto, se aplicarán metodologías estadísticas para validar la significancia con respecto a la variable objetivo. Para esto último se evaluarán la colinealidad, la cointegración, la correlación y la estacionalidad, entre otros aspectos.

- Revisar la correlación dinámica.
- Cointegración.
- Colinealidad.
- Estacionalidad.
- Normalidad de los residuos.
- Homocedasticidad.

3. Predicción de precios futuros

Una vez se tengan corroboradas las variables, se procederá a entrenar y validar el modelo, mediante la combinación de diferentes tipos de redes, ya que a veces puede producir mejores resultados. Se usará una CNN para extraer características de las variables y luego introducir esas características en un LSTM para hacer la predicción final del precio.

Se utilizarán herramientas como Keras Tuner, Optuna o GridSearchCV y RandomizedSearchCV de *sklearn,* que pueden ayudar a automatizar el proceso de ajuste de los hiperparámetros del modelo.

4. Creación del portafolio óptimo

Una vez tengamos los precios futuros estimados, se procederá a calcular los rendimientos esperados y la matriz de varianza-covarianza para las acciones, para luego utilizarlos en el cálculo de las ponderaciones óptimas de los portafolios, mediante la utilización de metodologías de teoría moderna de portafolios y modelos de AI.

II. Cronograma

Actividad	Semanas																			
	Agosto				Septiembre				Octubre				Noviembre				Diciembre			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Realizar una revisión de la literatura sobre técnicas de optimización de portafolios, principales técnicas de inteligencia artificial y machine learning para comprender el contexto y las oportunidades de mejora en la gestión de carteras de inversión.																				
Investigacion en Fuentes secundarias																				
Diseño Estado del arte																				
Identificar y seleccionar las variables relevantes para el análisis y la predicción de precios y volatilidades de los activos subyacentes, incluyendo variables macroeconómicas, microestructurales y técnicas.																				
Obtencion base de datos de historicos																				
Analitica descreptivida de los datos.																				
Validacion de los datos																				
Definir y construir los diferentes algoritmos y herramientas que se utilizaran en la construcción del modelo.																				
Revision diferentes lenguajes y su selección																				
Definicion de idea general del modelo																				
Evaluar la eficiencia del modelo de predicción.																				
Validacion del modelo																				
Optimizacion del modelo																				
Comunicar de manera efectiva los resultados y conclusiones de la investigación, proporcionando una guía práctica para la implementación y el uso del modelo propuesto en la gestión de carteras de inversión y la toma de decisiones financieras.																				
Elaboración de informe final																				
Entrega de informe final																				

III. Resultados

Para el desarrollo del análisis, se tomaron datos diarios de Bloomberg, de 5 de las principales acciones que componen el S&P500, sus múltiplos y algunas variables macroeconómicas definidas anteriormente, desde el 2015 hasta el 2020.

Ilustración 26. Base de datos variables

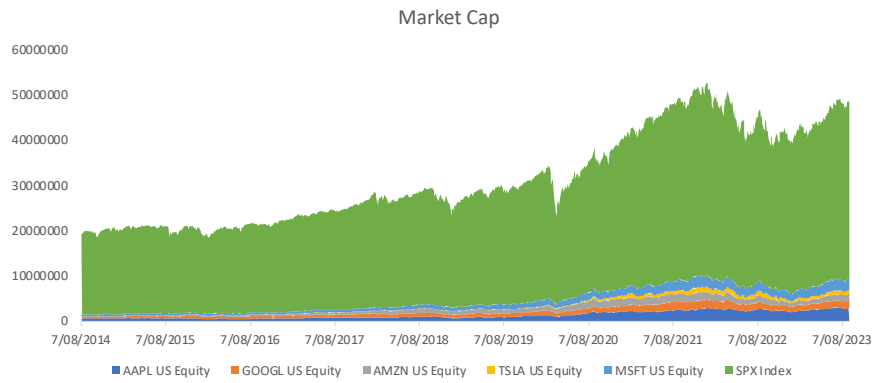
	Dates	PX_LAST	PE_RATIO	PX_TO_BOOK_RATIO	PX_TO_SALES_RATIO	PX_TO_CASH_FLOW	RETURN_ON_ASSET	RETURN_ON_CAP	RETURN_COM_EQY	EBITDA
0	2014-08-05	43.08	16.3112	3.9532	4.1164	11.0836	14.0235	21.7881	26.1652	8224
1	2014-08-06	42.74	16.1824	3.9220	4.0840	10.9961	14.0235	21.7881	26.1652	8224
2	2014-08-07	43.23	16.3680	3.9670	4.1308	11.1222	14.0235	21.7881	26.1652	8224
3	2014-08-08	43.20	16.3566	3.9642	4.1279	11.1145	14.0235	21.7881	26.1652	8224
4	2014-08-11	43.20	16.3566	3.9642	4.1279	11.1145	14.0235	21.7881	26.1652	8224

Fuente: Elaboración propia.

Para el análisis se tomaron 5 compañías que a lo largo de los últimos años han tenido un *market cap* conjunto en promedio del 25 % del total del correspondiente al S&P500. Estas son:

- Apple.
- Google.
- Amazon.
- Tesla.
- Microsoft.

Ilustración 27. *Market cap*



Fuente: Elaboración propia.

Análisis y descripción de variables

Correlación dinámica

Para el análisis de cada una de las variables predictoras, se compararán con la variable dependiente, que en este caso es el precio de la acción con un total de 20 rezagos *(lags),* para así validar la fuerza de la correlación. En nuestro caso usaremos como punto mínimo de correlación un coeficiente de 50 % que nos represente una correlación moderada.

Ilustración 28. Escalas de coeficiente de correlación

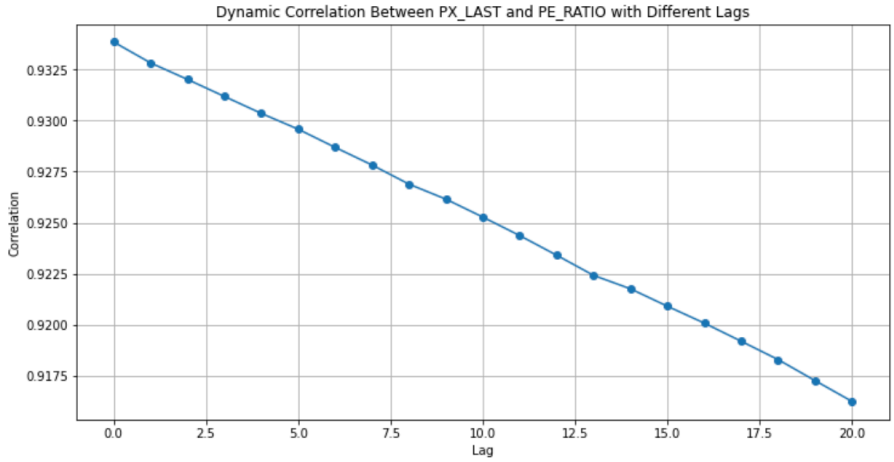
Fuente: Elaboración propia.

Como se observa en las siguientes gráficas correspondientes a Apple, los precios de las diferentes acciones tienen un elemento común en cuanto a las correlaciones con las diferentes variables analizadas.

Podemos concluir que el precio y los diferentes múltiplos tienen una correlación fuerte temporal por encima del 90 % en promedio para las diferentes acciones y los múltiplos analizados. En cuanto a las variables macroeconómicas, observamos un comportamiento mixto, con correlaciones temporales fuertes por encima del 85 % y correlaciones atemporales con diferentes *lags;* encontramos diferentes variables en las que la máxima correlación se encuentra en el *lag* 20 y podría seguir aumentando.

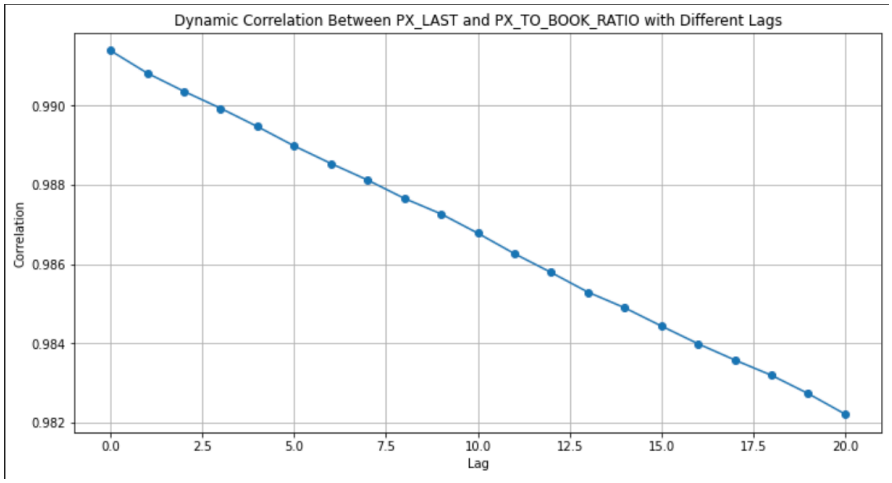
Lo que se propone para el análisis será descartar primeramente las variables con correlaciones inferiores a 40 %, encontrar para cada una de las variables el *lag* que maximice esta correlación y definir esta como el punto de partida para los diferentes análisis descriptivos que se les realizan a las variables.

Ilustración 29. Correlación dinámica entre precio y PE_ratio



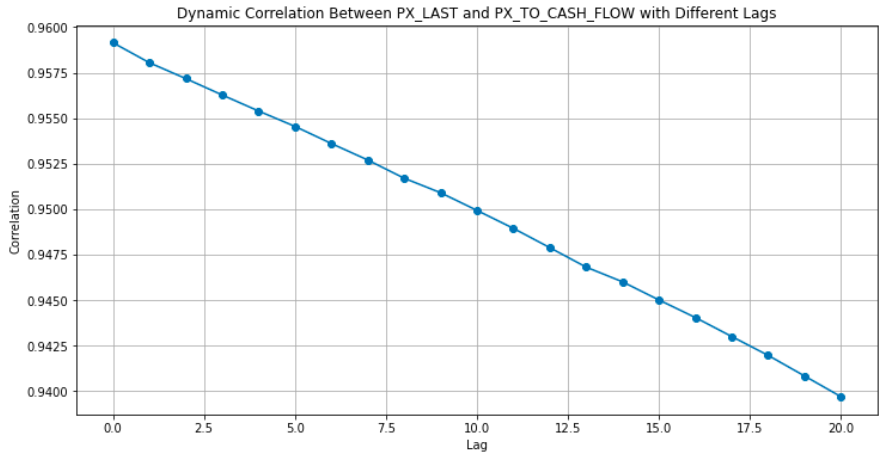
Fuente: Elaboración propia.

Ilustración 30. Correlación dinámica entre precio y *price to books*



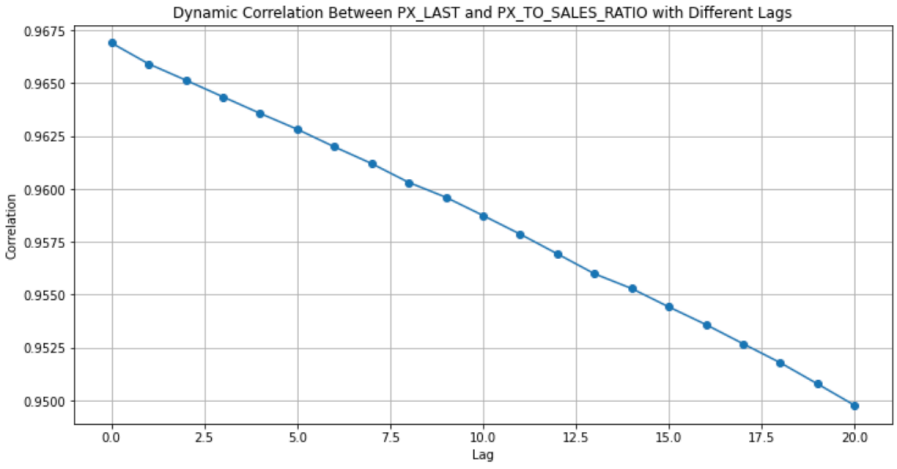
Fuente: Elaboración propia.

Ilustración 31. Correlación dinámica entre precio y *price to* CF *ratio*



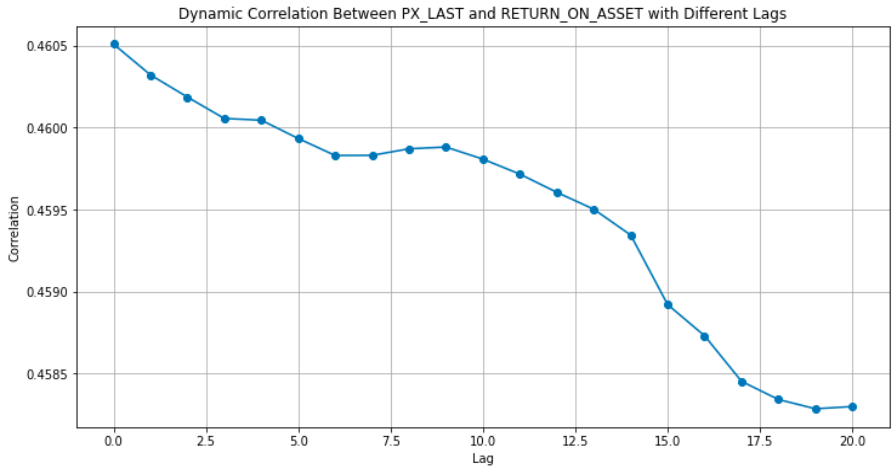
Fuente: Elaboración propia.

Ilustración 32. Correlación dinámica entre precio y *price to sales ratio*



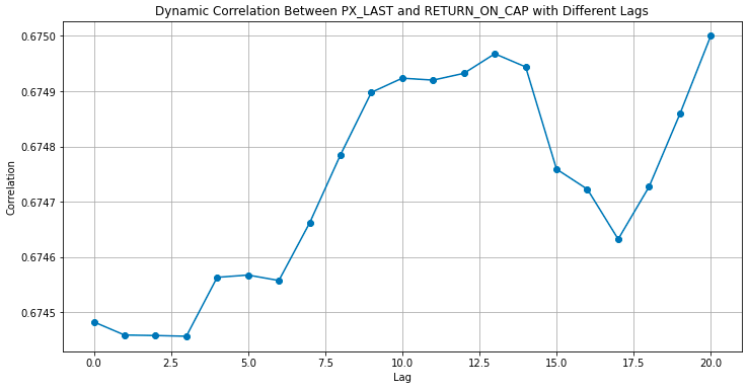
Fuente: Elaboración propia.

Ilustración 33. Correlación dinámica entre precio y ROA



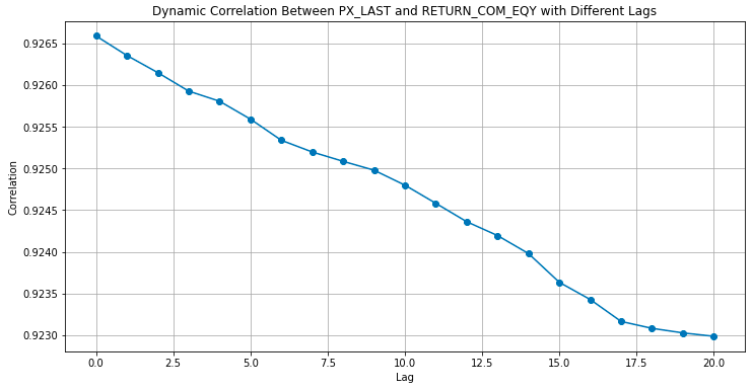
Fuente: Elaboración propia.

Ilustración 34. Correlación dinámica entre precio y *price to* ROC



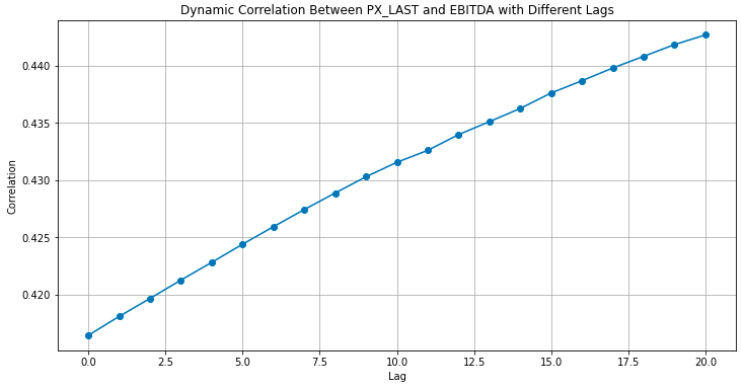
Fuente: Elaboración propia.

Ilustración 35. Correlación dinámica entre precio y *price to* ROE



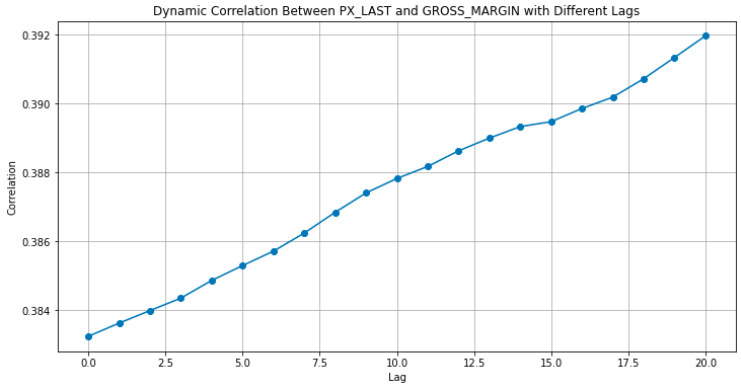
Fuente: Elaboración propia.

Ilustración 36. Correlación dinámica entre precio y EBITDA



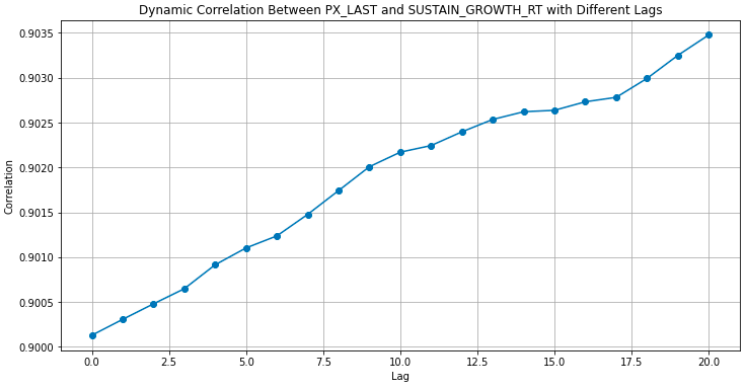
Fuente: Elaboración propia.

Ilustración 37. Correlación dinámica entre precio y *gross margin*



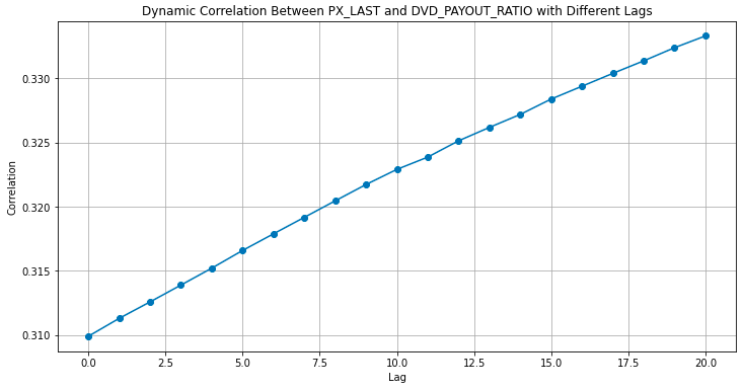
Fuente: Elaboración propia.

Ilustración 38. Correlación dinámica entre precio y *g*



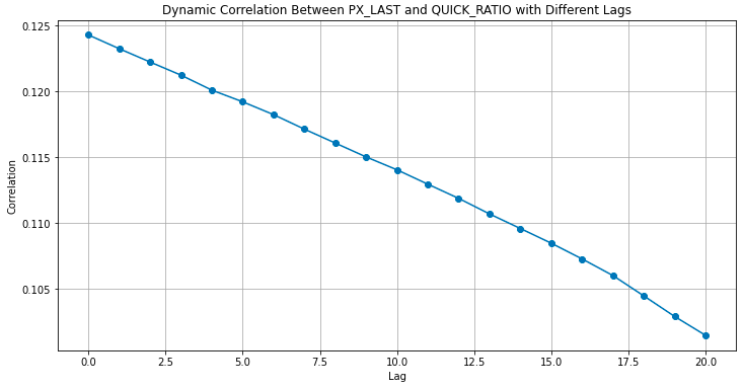
Fuente: Elaboración propia.

Ilustración 39. Correlación dinámica entre precio y *payout ratio*



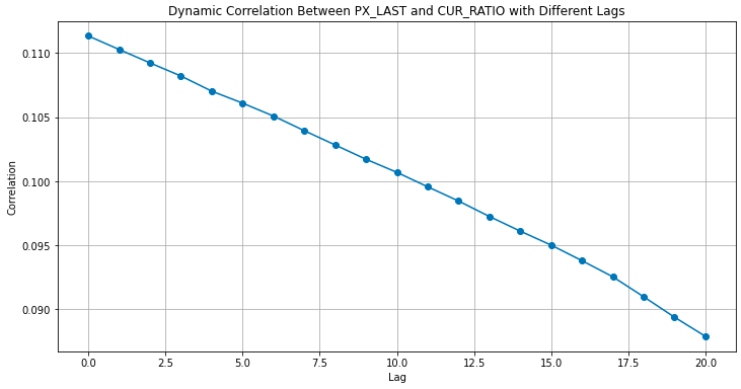
Fuente: Elaboración propia.

Ilustración 40. Correlación dinámica entre precio y *quick ratio*



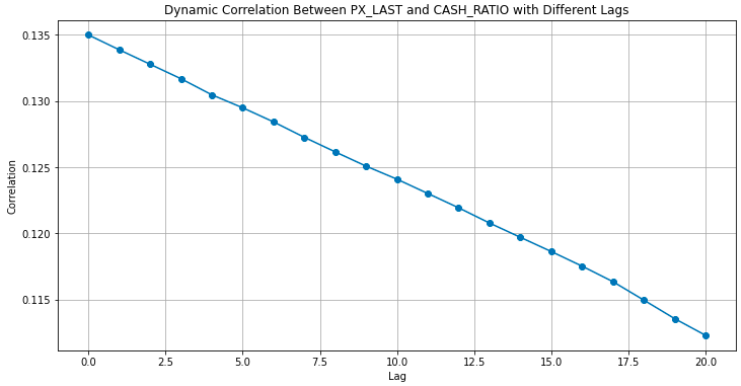
Fuente: Elaboración propia.

Ilustración 41. Correlación dinámica entre precio y PE_*current ratio*



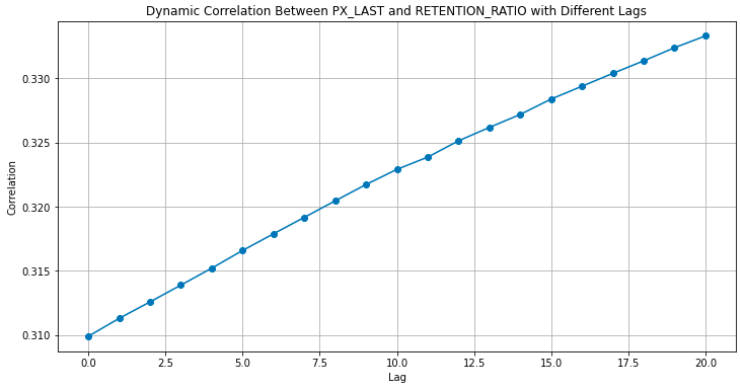
Fuente: Elaboración propia.

Ilustración 42. Correlación dinámica entre precio y CF



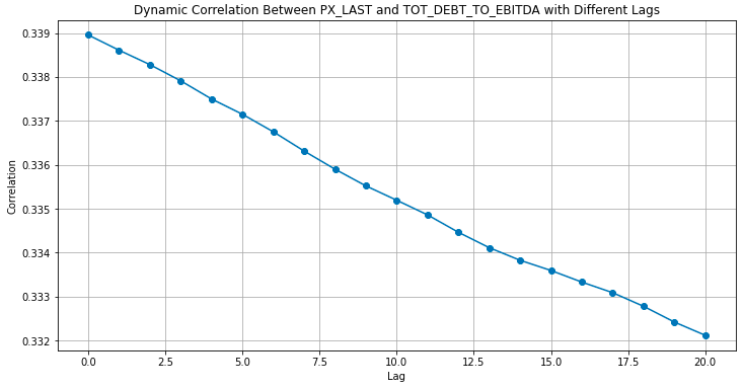
Fuente: Elaboración propia.

Ilustración 43. Correlación dinámica entre precio y *retention rate*



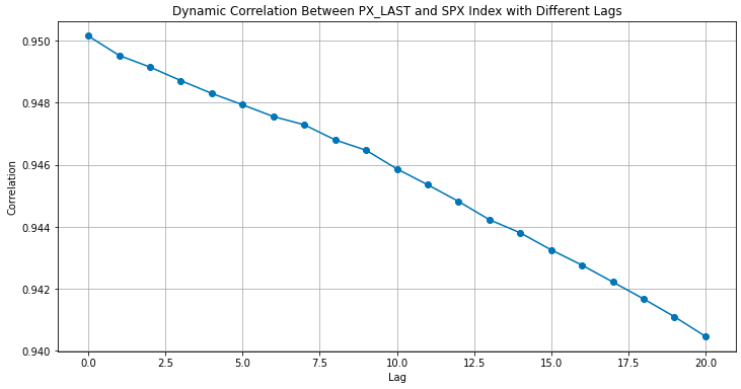
Fuente: Elaboración propia.

Ilustración 44. Correlación dinámica entre precio y *debt-to*-EBITDA



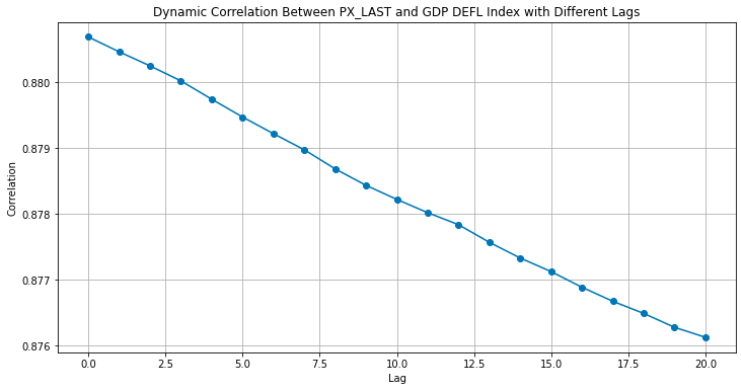
Fuente: Elaboración propia.

Ilustración 45. Correlación dinámica entre precio y S&P500



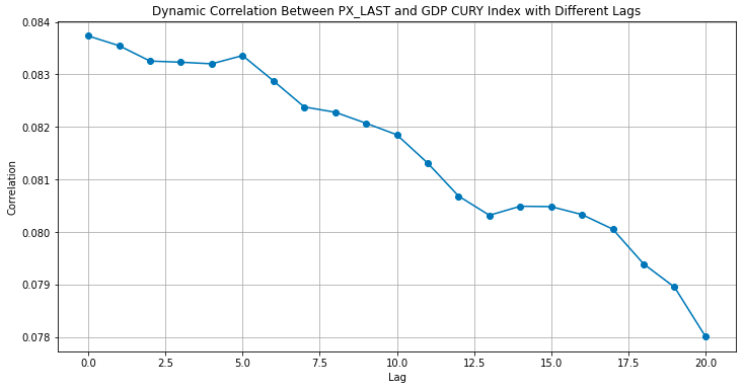
Fuente: Elaboración propia.

Ilustración 46. Correlación dinámica entre precio y deflactor del GDP



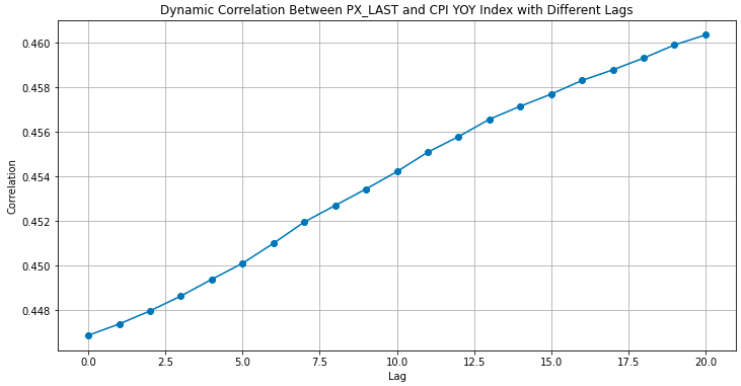
Fuente: Elaboración propia.

Ilustración 47. Correlación dinámica entre precio y GDP



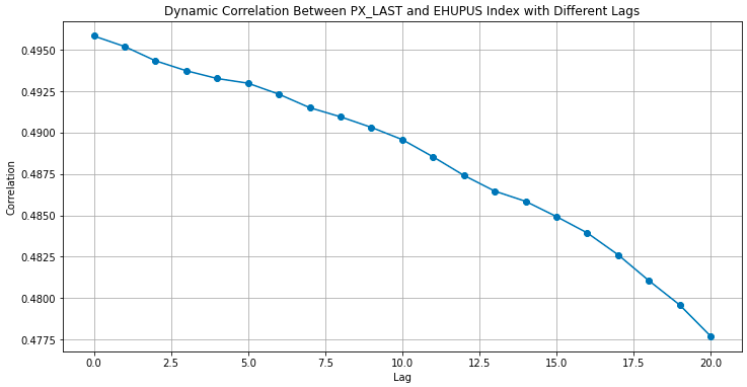
Fuente: Elaboración propia.

Ilustración 48. Correlación dinámica entre precio y CPI



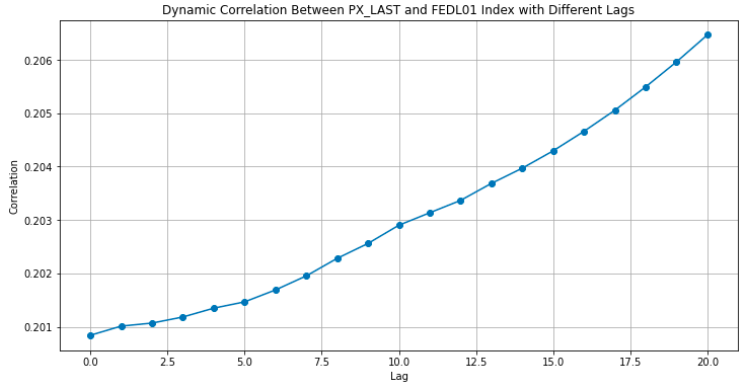
Fuente: Elaboración propia.

Ilustración 49. Correlación dinámica entre precio y PE_*ratio*



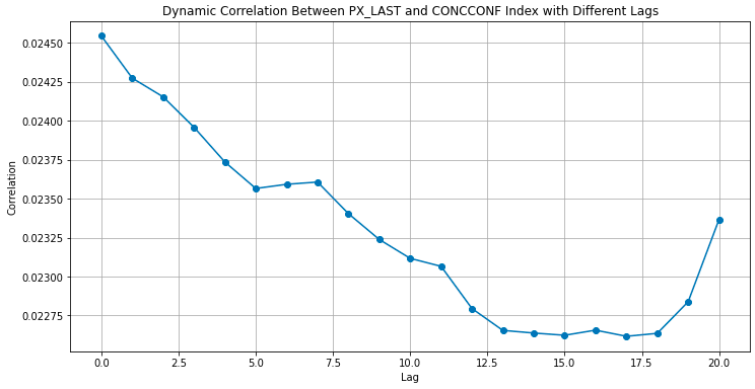
Fuente: Elaboración propia.

Ilustración 50. Correlación dinámica entre precio y FED *rate*

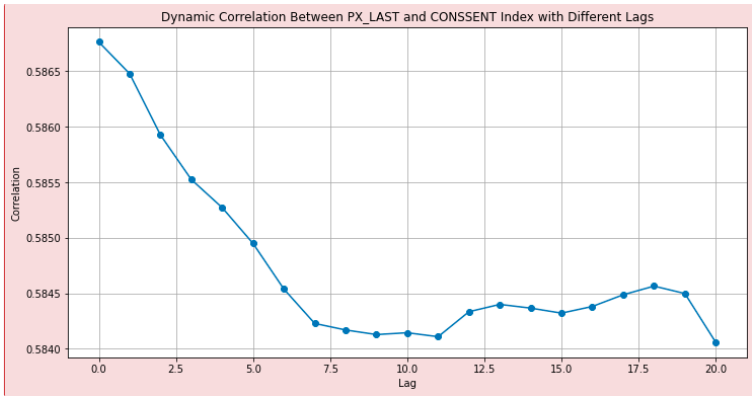


Fuente: Elaboración propia.

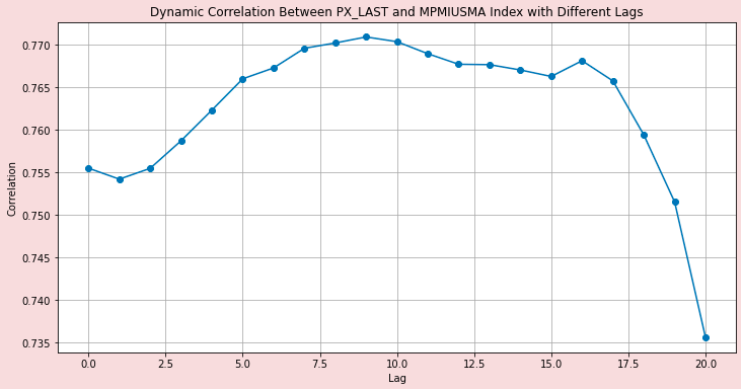
Ilustración 51. Correlación dinámica entre precio e índice de confianza



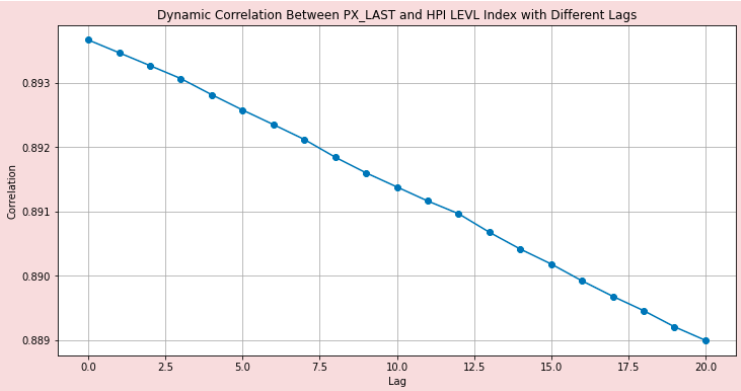
Fuente: Elaboración propia.



Fuente: Elaboración propia.



Fuente: Elaboración propia.



Comentado [U1]: Faltan los nombres de estas ilustraciones

Fuente: Elaboración propia.

Análisis de colinealidad

La colinealidad se evaluará mediante el cálculo del **factor de inflación de la varianza** (VIF). El VIF es una medida que se utiliza para evaluar la multicolinealidad en un modelo de regresión. Indica cuánto aumenta la varianza de los coeficientes de regresión debido a la multicolinealidad.

Se ha definido un umbral de aceptación igual o inferior a 10 para el análisis de colinealidad entre las variables. Durante este análisis, se identificó una alta colinealidad entre las variables, lo que podría ocasionar problemas en el modelo. Para abordar esta colinealidad, se ha decidido eliminar aquellas variables que presentan una correlación muy fuerte, según el umbral establecido. Además de esto, aplicamos los *lags* que maximizaban la correlación en el paso anterior.

En términos generales, encontramos que para la mayoría de las variables, algunos múltiplos y principalmente algunas variables macroeconómicas, como las tasas de política monetaria y los índices de sentimiento de los consumidores, las variables que hayan pasado este análisis se compararán con el análisis realizado con la correlación, y estas se continuarán utilizando para las demás validaciones.

Las variables son las siguientes:

- *Price to earnings ratio* (PER).
- *Return on assets* (ROA).
- *Dividend payout.*
- *Retention rate.*
- EBITDA.
- *Gross margin.*
- *Quick ratio.*
- *Price to book value* (P/BV).
- *Price to* CFO.
- Tasa de la FED.
- *Consumer price index* (CPI).

- Índice de confianza del consumidor.

Microsoft

Ilustración 52. Colinealidad de Microsoft y demás variables

	Variable	VIF
0	const	3864.495899
1	PE_RATIO	2.176991
2	RETURN_ON_ASSET	3.152601
3	GROSS_MARGIN	1.703648
4	FEDL01 Index	1.338989
5	CONSSENT Index	2.362718

Fuente: Elaboración propia.

Apple

Ilustración 53. Colinealidad de Apple y demás variables

	Variable	VIF
0	const	1807.285588
1	PE_RATIO	1.935476
2	PX_TO_BOOK_RATIO	1.877685
3	EBITDA	1.488363
4	FEDL01 Index	2.118805
5	CONSSENT Index	1.649807

Fuente: Elaboración propia.

Tesla

Ilustración 54. Colinealidad de Tesla y demás variables

	Variable	VIF
0	PE_RATIO	5.396623
1	PX_TO_BOOK_RATIO	3.114712
2	GROSS_MARGIN	2.819527
3	DVD_PAYOUT_RATIO	NaN
4	RETENTION_RATIO	1449.154816
5	FEDL01 Index	1.598140
6	CONSSENT Index	1.484785

Fuente: Elaboración propia.

Amazon

Ilustración 55. Colinealidad de Amazon y demás variables

	Variable	VIF
0	PE_RATIO	13.688067
1	PX_TO_CASH_FLOW	19.664404
2	GROSS_MARGIN	6.519299
3	DVD_PAYOUT_RATIO	NaN
4	QUICK_RATIO	11.372912
5	RETENTION_RATIO	6187.536033
6	FEDL01 Index	2.494027
7	CONSSENT Index	2.144997

Fuente: Elaboración propia.

Google

Ilustración 56. Colinealidad de Google y demás variables

	Variable	VIF
0	PE_RATIO	1.923247
1	PX_TO_BOOK_RATIO	1.754268
2	DVD_PAYOUT_RATIO	NaN
3	QUICK_RATIO	2.062415
4	RETENTION_RATIO	3534.937874
5	FEDL01 Index	1.892184
6	CONSSENT Index	1.607452

Fuente: Elaboración propia.

Análisis de estacionalidad

Es importante revisar la estacionalidad de una variable antes de utilizarla en un modelo de proyección por varias razones:

1. Identificar estos patrones es esencial para comprender cómo se comporta la variable en el tiempo y para predecir su comportamiento futuro.

2. Si una variable tiene componentes estacionales significativos y no se tienen en cuenta en el modelo, las proyecciones pueden ser imprecisas.

3. Dependiendo de la presencia de estacionalidad en los datos, es posible que se necesite utilizar modelos específicos que puedan capturar y modelar esta estacionalidad de manera efectiva.

4. Ignorar la estacionalidad puede llevar a sesgos en las proyecciones y los análisis.

Ilustración 57. Estacionalidad PE_ratio

Fuente: Elaboración propia.

Ilustración 58. Estacionalidad *price to book*

Fuente: Elaboración propia.

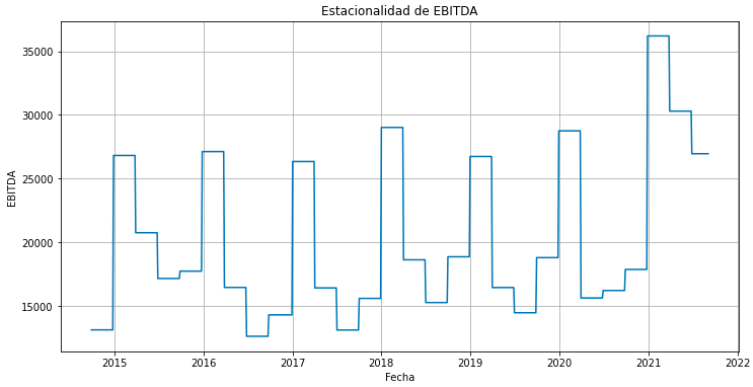
Ilustración 59. Estacionalidad *price to* CF

Fuente: Elaboración propia.

Ilustración 60. Estacionalidad ROA

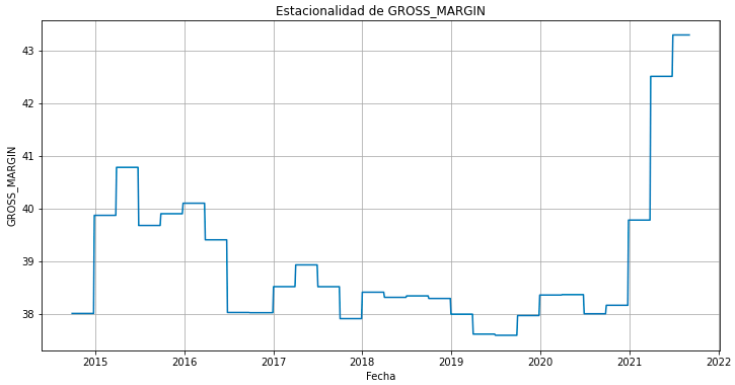
Fuente: Elaboración propia.

Ilustración 61. Estacionalidad EBITDA



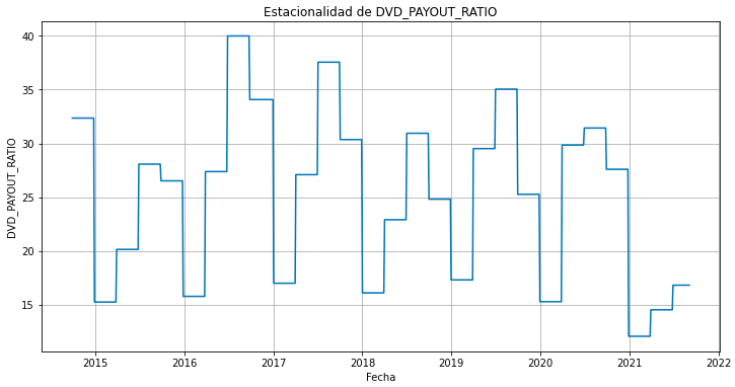
Fuente: Elaboración propia.

Ilustración 62. Estacionalidad *gross margin*



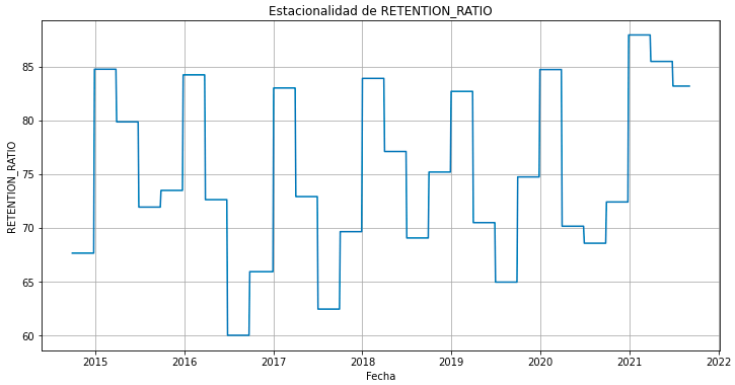
Fuente: Elaboración propia.

Ilustración 63. Estacionalidad *dividend payout*



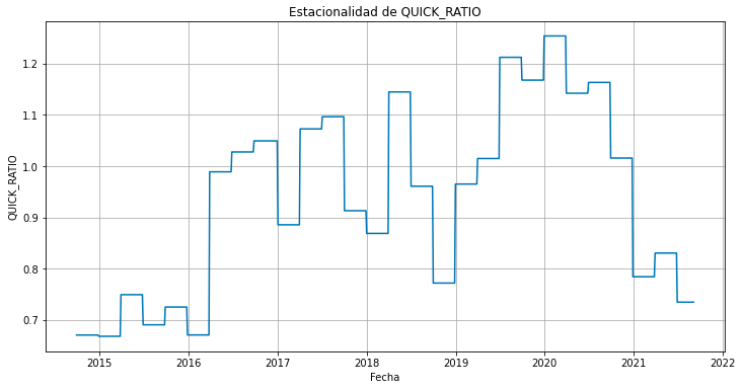
Fuente: Elaboración propia.

Ilustración 64. Estacionalidad *retention rate*



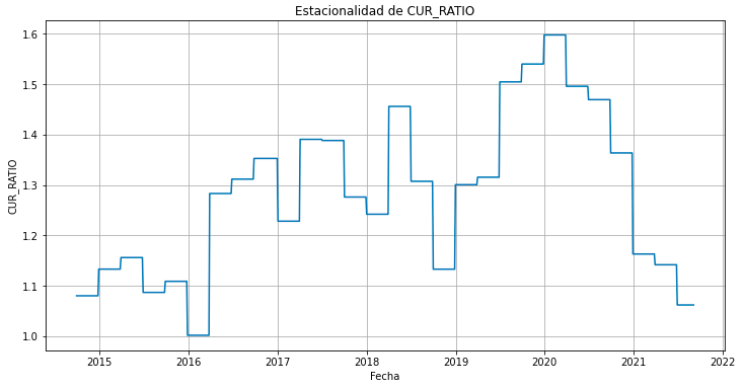
Fuente: Elaboración propia.

Ilustración 65. Estacionalidad *quick ratio*



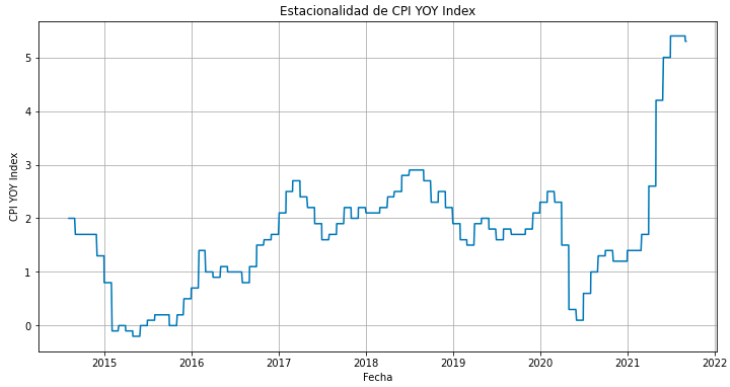
Fuente: Elaboración propia.

Ilustración 66. Estacionalidad *current ratio*



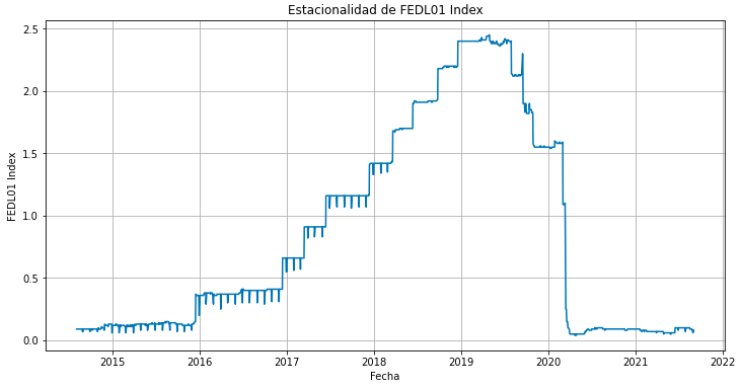
Fuente: Elaboración propia.

Ilustración 67. Estacionalidad CPI



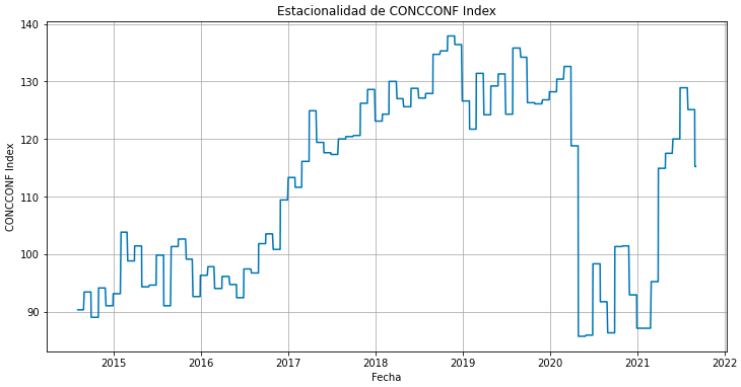
Fuente: Elaboración propia.

Ilustración 68. Estacionalidad FED *rate*



Fuente: Elaboración propia.

Ilustración 69. Estacionalidad índice de confianza



Fuente: Elaboración propia.

Mediante estas gráficas se puede realizar un análisis de descomposición de los factores de la estacionalidad. En términos generales encontramos una tendencia creciente en la mayoría de las variables, y al ser una tendencia significativa podríamos estar esperando que continúe moviéndose en esta dirección. Ahora, al revisar los patrones repetitivos y su magnitud podemos encontrar cómo la mayoría de las variables que incluyen insumos de los estados financieros o macroeconómicos tienen este factor bastante marcado.

Análisis de cointegración

El análisis de cointegración de Granger es una técnica utilizada para determinar si dos o más series de tiempo están cointegradas, lo que implica que comparten una relación de largo plazo.

Variable: **PX_TO_BOOK_RATIO.**

Estadística de cointegración: -3.6308429215059648.

Valor p: 0.02242866188725039.

Valor crítico al 5 %: -3.3614948521275867.

Variable: **EBITDA.**

Estadística de cointegración: -3.779836479363289.

Valor p: 0.014418101382643812.

Valor crítico al 5 %: -3.3614948521275867.

Variable: **DVD_PAYOUT_RATIO.**

Estadística de cointegración: -3.3585852861237036.

Valor p: 0.047163100159921034.

Valor crítico al 5 %: -3.3614948521275867.

Variable: **RETURN_ON_ASSET.**

Estadística de cointegración: -3.4277367306366036.

Valor p: 0.03936183798790379.

Valor crítico al 5 %: -3.3614948521275867.

Variable: **PE_RATIO.**

Estadística de cointegración: -3.5166406908920638.

Valor p: 0.03094846505889567.

Valor crítico al 5 %: -3.3614948521275867.

Variable: **GROSS_MARGIN.**

Estadística de cointegración: -3.3475809934454746.

Valor p: 0.048515319390231945.

Valor crítico al 5 %: -3.3614948521275867.

Variable: **QUICK_RATIO.**

Estadística de cointegración: -3.522258540597674.

Valor p: 0.03047257356608746.

Valor crítico al 5 %: -3.3614948521275867.

Variable: **FEDL01 Index.**

Estadística de cointegración: -3.971861136548644.

Valor p: 0.007873029410350816.

Valor Crítico al 5%: -3.3614948521275867.

Variable: **RETENTION_RATIO.**

Estadística de cointegración: -4.137119171291154.

Valor p: 0.0045336698585009905.

Valor crítico al 5 %: -3.3614948521275867.

Variable: **CPI YOY Index.**

Estadística de cointegración: -6.522258540597674.

Valor p: 0.02037257356608746.

Valor crítico al 5 %: -6.3614948521275867.

Variable: **CONSSENT Index.**

Estadística de cointegración: -4.079836479363289.

Valor p: 0.024418101382643812.

Valor crítico al 5 %: -7.4524948521275867.

Con base en los resultados obtenidos, concluimos que las variables con un valor p que sea menor que el valor crítico, en este caso de 0,05, están cointegradas. En los resultados mostramos las variables que pasaron esta prueba estadística.

Análisis de homocedasticidad y normalidad

Este análisis se utiliza para verificar si la varianza de los errores en un modelo de regresión es constante a lo largo de todos los niveles de las variables independientes. La homocedasticidad es un supuesto importante en muchos modelos de regresión y su violación puede llevar a resultados sesgados e incorrectos.

La verificación de la normalidad de los residuos es importante en el análisis estadístico y en la modelación de datos por varias razones:

1. Muchos modelos estadísticos, como la regresión lineal, ANOVA, pruebas t y otros, asumen que los errores (residuos) siguen una distribución normal. Si esta suposición no se cumple, los resultados de los análisis pueden ser incorrectos o sesgados.

2. La normalidad de los residuos es fundamental para realizar inferencias estadísticas válidas. Si los residuos no son normales, los estimadores de los coeficientes pueden ser sesgados y las predicciones pueden ser inexactas.

3. En modelos de regresión, la estimación de los coeficientes y la precisión de las predicciones se basan en la suposición de normalidad de los residuos.

Ambos análisis se realizarán una vez se haya generado el modelo final con las variables definidas según las validaciones realizadas y se podrán encontrar los resultados de este análisis una vez definido el modelo.

Análisis de sentimientos

En este modelo hemos utilizado la API de Twitter Developers para obtener nuestros datos. Hemos aplicado filtros para cada una de las acciones ($AAPL, $AMZN, $TSLA), y debido a las restricciones de la API, hemos reunido datos a partir del 2015 y hasta el 2020,

recopilando más de un millón de registros para las 5 empresas analizadas. La importación de los datos se llevó a cabo utilizando la biblioteca *tweepy*. Una vez nos conectamos con las claves proporcionadas por Twitter Developers, realizamos la siguiente consulta:

```
search_words = "$APPL"

date_since = "2015-01-01"

# Collect tweets

tweets = tw.Cursor(api.search,

q=search_words,

lang="en",

since=date_since).items()
```

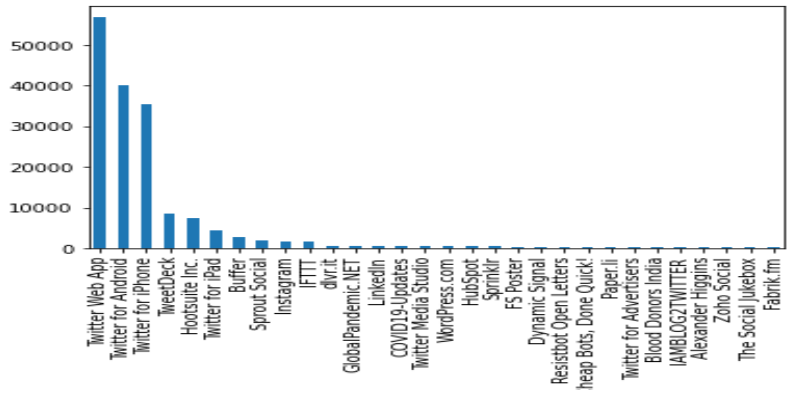
Ilustración 70. *Tweets*

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date
0	·V·i●ι¢†	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	18775	False	2020-07-25 12:27:21
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	1677	24	True	2020-07-25 12:27:17
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	9525	7254	False	2020-07-25 12:27:14
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	2019-03-07 01:45:06	197	987	1488	False	2020-07-25 12:27:10
4	DIPR-J&K	Jammu and Kashmir	🖋Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	168	101	False	2020-07-25 12:27:08

Fuente: Elaboración propia.

Una vez tenemos los datos, procedemos a seleccionar las frases únicas y contamos cuántas veces se utilizan.

Ilustración 71. Frecuencia de frases x



Fuente: Elaboración propia.

Para el desarrollo de esta parte del análisis se decidió utilizar el modelo preentrenado BERT y su tokenización. El modelo BERT *(bidirectional encoder representations from transformers)* es un modelo de lenguaje basado en la arquitectura *transformer* que ha revolucionado el campo del procesamiento del lenguaje natural (NLP). BERT se destacó por su capacidad para comprender el contexto de una palabra en un texto al procesar todo el contexto en ambas direcciones, es decir, de izquierda a derecha y de derecha a izquierda. A continuación, se detalla cómo funciona BERT y su proceso de tokenización (Chang, 2018).

Funcionamiento de BERT:

BERT se entrena en una tarea de "llenado de huecos" *(cloze task)* en la que se le proporciona un fragmento de texto con una palabra oculta y se le pide que adivine cuál es esa palabra. Para lograr esto, BERT utiliza una arquitectura de codificador bidireccional que consta de múltiples capas de atención y capas *feedforward* (Chang, 2018).

1. **Tokenización de entrada:** El texto de entrada se divide en *tokens.* Los *tokens* son unidades de texto más pequeñas, que pueden ser palabras completas o subpalabras. Por ejemplo, la palabra "jugadores" podría dividirse en "jug" y "adores". Esta tokenización se realiza utilizando el tokenizador de BERT, que utiliza el modelo WordPiece.

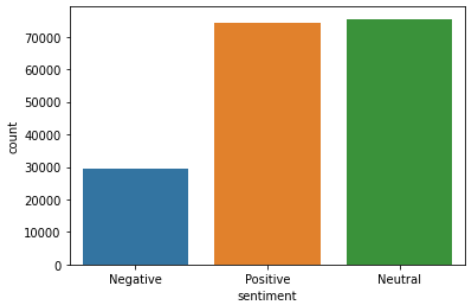
- **Tokenización inicial:** El texto se divide en palabras y se eliminan los espacios en blanco.
- **Tokenización de WordPiece:** Cada palabra se divide en subpalabras (*tokens* de WordPiece). Por ejemplo, la palabra "jugadores" se divide en "jug" y "adores". Estos *tokens* se toman del vocabulario de BERT.
- ***Tokens* especiales:** Se agregan *tokens* especiales al principio y al final del texto para indicar el inicio y el final de una oración o secuencia. También se agrega un *token* "[CLS]" al principio para representar la clase de la secuencia.
- ***Padding:*** Si es necesario, se agrega relleno (*tokens* de "[PAD]") para que todas las secuencias tengan la misma longitud.

2. ***Embedding* de *token:*** Cada *token* se asigna a un vector de alta dimensionalidad, que es parte del proceso de preentrenamiento de BERT. Estos vectores se utilizan para representar las palabras y las subpalabras en el texto de entrada.
3. **Capas de atención bidireccional:** BERT utiliza capas de atención que permiten que cada *token* mire todos los demás *tokens* en el contexto. Esto significa que BERT procesa el contexto tanto a la izquierda como a la derecha de cada palabra, capturando relaciones de dependencia complejas.
4. **Capas *feedforward:*** Después de las capas de atención, hay capas *feedforward* que realizan transformaciones lineales y aplican funciones de activación no lineales para aprender características más avanzadas.
5. **Preentrenamiento y ajuste fino *(fine tuning):*** BERT se preentrena en grandes cantidades de texto sin etiquetas para que aprenda representaciones de palabras y frases en un nivel profundo. Luego, se ajusta finamente en tareas específicas de NLP, como la clasificación de sentimientos o la extracción de entidades, utilizando un conjunto de datos etiquetado más pequeño.

Una vez hemos analizado cada uno de los *tweets,* los clasificamos de la siguiente manera:

- Positivos: 1.
- Neutral: 0.
- Negativos: -1.

Ilustración 72. Tabla de frecuencia de sentimientos



Fuente: Elaboración propia.

A continuación, se muestra un ejemplo de *tweets* clasificados como positivos por nuestro modelo.

Ilustración 73. DF *tweets* con su sentimiento

hashtags	source	is_retweet	extracted_hashtags	clean_tweet	sentiment_results	polarity	subjectivity	sentiment
NaN	Twitter for iPhone	False	[]	If I smelled the scent of hand sanitizers toda...	{'polarity': -0.25, 'subjectivity': 0.25, 'sen...	-0.25	0.250000	Negative
NaN	Twitter for Android	False	[]	Hey and wouldnt it have made more sense to ha...	{'polarity': 0.5, 'subjectivity': 0.5, 'sentim...	0.50	0.500000	Positive
['COVID19']	Twitter for Android	False	[#COVID19]	Trump never once claimed was a hoax We all cl...	{'polarity': 0.0, 'subjectivity': 0.0, 'sentim...	0.00	0.000000	Neutral
['COVID19']	Twitter for iPhone	False	[#COVID19]	The one gift has give me is an appreciation f...	{'polarity': 0.0, 'subjectivity': 0.3571428571...	0.00	0.357143	Neutral
['CoronaVirusUpdates', 'COVID19']	Twitter for Android	False	[#CoronaVirusUpdates, #COVID19]	25 July : Media Bulletin on Novel	{'polarity': 0.0, 'subjectivity': 0.0, 'sentim...	0.00	0.000000	Neutral

Fuente: Elaboración propia.

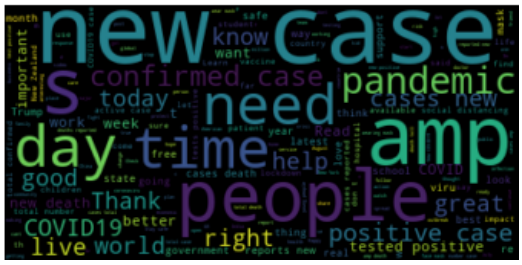
Ilustración 74. *Tweets* positivos

```
['Hey wouldnt sense players pay respects A…',
 'Change Work General (and recruiting specifically) via/',
 'Praying good health recovery',
 '👏🏻 — safe safe commit ensure…',
 'Lets protect real numbers climbing fast Continent Lets n…',
 'Second wave Flandersback',
 'COVID Update: infection rate Florida following natural curve experts predicted initial cu…',
 'Good Patriots Volunteer Election Judge Polls open without…',
 'comprehensive review amp Analysis: key ways WASH help reduce transmission…',
 'crosses 2 lakh mark 150055 ppp recovered far positive today 6988/total…',
 'Actor father MrGKReddy tested positive 15/20 days s…',
```

Fuente: Elaboración propia.

Finalmente, utilizando la función de tokenización de BERT, dividimos los *tweets* en palabras. Luego, creamos nubes de palabras para identificar las palabras más utilizadas según el sentimiento expresado.

Ilustración 75. Nube de *tweets* negativos



Fuente: Elaboración propia.

Ilustración 76. Nube de *tweets* positivos

Fuente: Elaboración propia.

Para agregar esta variable como una característica a nuestro modelo, se realizó una agrupación para cada uno de los días y promediamos los valores de los *tweets* correspondientes a cada una de las compañías analizadas. En los días en los que no se registraron *tweets* asignamos un valor neutral a esta variable.

Al realizar el análisis de esta variable contra el precio de cada una de las compañías encontramos en promedio una correlación fuerte superior al 65 %, tanto para el dato temporal como para el dato atemporal con un *lag,* y teniendo en cuenta esto decidimos realizar los análisis y el modelo con ambas variables en la misma temporalidad. Para esta variable en específico se decidió realizar una prueba de causalidad utilizando la metodología de Granger, ya que al no ser una variable muy utilizada en la literatura para este tipo de análisis queríamos validar que sí existe una relación causal entre las dos variables, y pudimos concluir estadísticamente que hay evidencia de causalidad de Granger entre el precio de la acción y el sentimiento que muestra el mercado por medio de las publicaciones en la red social X (antes Twitter).

Estadística de Causalidad: -3.646820638.

Valor p: 0.034855311945.

Valor crítico al 5 %: -3.3614948521275867.

Predicción de precios

Una vez se ha realizado la evaluación de todas las variables y se ha definido cuáles son las que harán parte de cada uno de los modelos. Después de haber realizado un análisis del estado del arte, el estudio de la literatura que trata este tipo de problemas, y de haber recibido la asesoría de expertos en el campo, se decidió tomar la arquitectura de red neuronal que combine una red neuronal convolucional (CNN) y capas de memoria a corto plazo (LSTM), las cuales son adecuadas para tareas de pronóstico de series temporales como la predicción del precio de las acciones, y para eso nos basamos en puntos como los siguientes:

- Las CNN son muy efectivas para extraer patrones espaciales y temporales de los datos. Las CNN se destacan por identificar patrones, como tendencias, ciclos o irregularidades, dentro de los datos. Al utilizar capas convolucionales, la red puede aprender y capturar automáticamente características relevantes en la serie temporal, lo que la hace adecuada para tareas en las que los patrones locales importan.

- Los LSTM son un tipo de red neuronal recurrente (RNN) que está diseñada para capturar dependencias de largo alcance en datos secuenciales. En el pronóstico de series de tiempo es crucial considerar no solo el estado actual, sino también cómo las observaciones pasadas contribuyen a las predicciones futuras. Los LSTM son capaces de aprender y recordar dichas dependencias temporales, lo que puede ser importante en la predicción del precio de las acciones, en la que los precios históricos pueden tener un impacto significativo en los precios futuros.

- Al combinar las capas CNN y LSTM se aprovechan las fortalezas de ambas arquitecturas. Las CNN son excelentes para la extracción de características y pueden ayudar a preprocesar los datos para capturar patrones relevantes. La salida de las capas CNN sirve como entrada para las capas LSTM, que se centran en modelar dependencias temporales. Este enfoque jerárquico permite que el modelo aprenda simultáneamente patrones de corto y largo plazo en los datos de series de tiempo. Se ha demostrado que la combinación de capas CNN y LSTM es efectiva en varias tareas de pronóstico de series temporales. Puede mejorar el poder predictivo del modelo en comparación con el uso de cualquiera de las arquitecturas por separado. Las CNN

pueden identificar patrones locales importantes, mientras que los LSTM pueden capturar cómo esos patrones evolucionan con el tiempo, lo que genera pronósticos más precisos.

Se utilizaron diferentes librerías, como TensorFlow, Keras, Numpy y Scikit-learn.

El uso de Keras Tuner permite una búsqueda automatizada de los mejores hiperparámetros para optimizar el rendimiento del modelo. Una breve descripción de cómo funciona este algoritmo de optimización y por qué se decidió trabajar con él es la siguiente:

- Se debe definir el espacio de posibles hiperparámetros y la arquitectura de su modelo. Esto incluye especificar hiperparámetros como tasas de aprendizaje, tasas de abandono, número de capas, unidades por capa, etc. El espacio de búsqueda se puede definir utilizando la API de Keras Tuner.
    - **Filters:** El número de filtros en la capa convolucional.
    - **Kernel_size:** El tamaño del núcleo convolucional.
    - **Lstm_units:** El número de unidades LSTM.
    - **Learning_rate:** La tasa de aprendizaje del optimizador.

Ilustración 77. Optimización de hiperparámetros

```
# Hyperparameters to search
filters = hp.Int('filters', min_value=32, max_value=256, step=32)
kernel_size = hp.Choice('kernel_size', values=[3, 5])
lstm_units = hp.Int('lstm_units', min_value=32, max_value=128, step=32)
learning_rate = hp.Float('learning_rate', min_value=1e-5, max_value=1e-2, sampling='log')
```

Fuente: Elaboración propia.

- Es necesario definir una función objetivo, que es una función que evalúa el rendimiento de un modelo para un conjunto determinado de hiperparámetros. En la mayoría de los casos esta función calcula una pérdida (p. ej., error cuadrático medio para tareas de regresión) o una métrica (p. ej., precisión para tareas de clasificación). El objetivo es minimizar o maximizar esta función objetivo. La arquitectura del

modelo es definida incluyendo los siguientes parámetros:

- Una capa convolucional 1D seguida de MaxPooling.
- Dos capas de LSTM.
- Una capa completamente conectada (densa).
- El modelo se compila con la tasa de aprendizaje seleccionada utilizando el optimizador Adam y el error cuadrático medio (MSE) como función de pérdida.

- Keras Tuner ofrece diferentes estrategias de ajuste, como búsqueda aleatoria, hiperbanda y optimización bayesiana. En este caso, el uso de la estrategia de búsqueda aleatoria que es el método RandomSearch de Keras Tuner especifica la función build_model como el objetivo a optimizar ('val_loss'), el número de pruebas a ejecutar (max_trials) y el directorio para almacenar los resultados.

- El sintonizador explora sistemáticamente el espacio de hiperparámetros. En la búsqueda aleatoria selecciona combinaciones aleatorias de hiperparámetros para entrenar y evaluar el modelo. No depende de ningún orden o patrón específico, lo que la hace más eficiente que una búsqueda en cuadrícula, especialmente cuando el espacio de búsqueda es vasto.

Ilustración 78. Optimización del modelo

```
# Create a Keras Tuner RandomSearch tuner
tuner = RandomSearch(
    build_model,
    objective='val_loss',
    max_trials=10,  # Number of hyperparameter combinations to try
    directory='my_tuner_directory',  # Directory to store results
    project_name='cnn_lstm_hyperparameters'  # Name of the project
)
```

Fuente: Elaboración propia.

- Para cada conjunto de hiperparámetros, el sintonizador entrena el modelo correspondiente utilizando un subconjunto de sus datos de entrenamiento (conjunto de validación) y evalúa su rendimiento utilizando la función objetivo definida. Luego, el sintonizador registra los resultados. Se define una *stopping callback* anticipada para detener el entrenamiento si no hay mejora en la pérdida de validación. Esto ayuda a prevenir el sobreajuste.

- Después de explorar una cantidad predefinida de combinaciones de hiperparámetros o alcanzar cierto presupuesto de tiempo, el sintonizador selecciona la configuración que produjo el mejor rendimiento (pérdida más baja o puntuación métrica más alta).

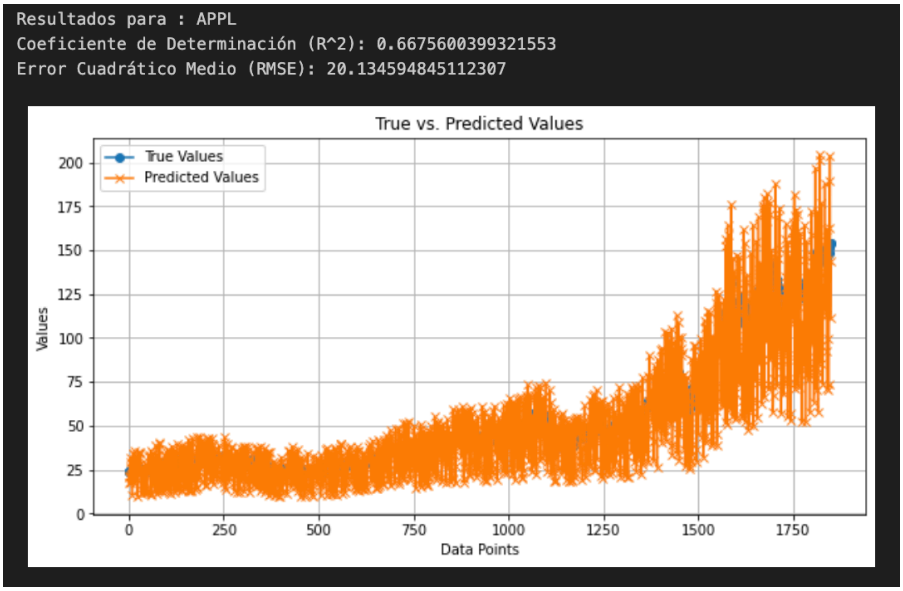
- Una vez que se determina la mejor configuración, se puede crear un modelo final utilizando esos hiperparámetros. Luego, se entrena este modelo en todo el conjunto de datos de entrenamiento durante una mayor cantidad de épocas para obtener el mejor modelo posible.

- Finalmente, se evalúa el desempeño del mejor modelo en un conjunto de datos de prueba separado para evaluar su desempeño de generalización. R cuadrado y RMSE se utilizaron como métricas de evaluación para cuantificar la precisión predictiva del modelo, mediante metodologías de *cross validation.*

Los resultados son bastante ajustados a los datos reales. Para las 5 compañías se tuvo un R^2 superior al 65 % en promedio, lo cual nos brinda una aceptable proyección de los datos según el análisis y la comparación con diferentes trabajos y métricas de la industria. Cabe destacar el buen comportamiento que se obtuvo por parte del modelo para Tesla con una métrica del 85 %, mientras que para Google se obtiene solo un 42 %. En las siguientes ilustraciones se puede observar gráficamente la comparación entre los datos reales y los proyectados mediante el modelo utilizado.

Apple

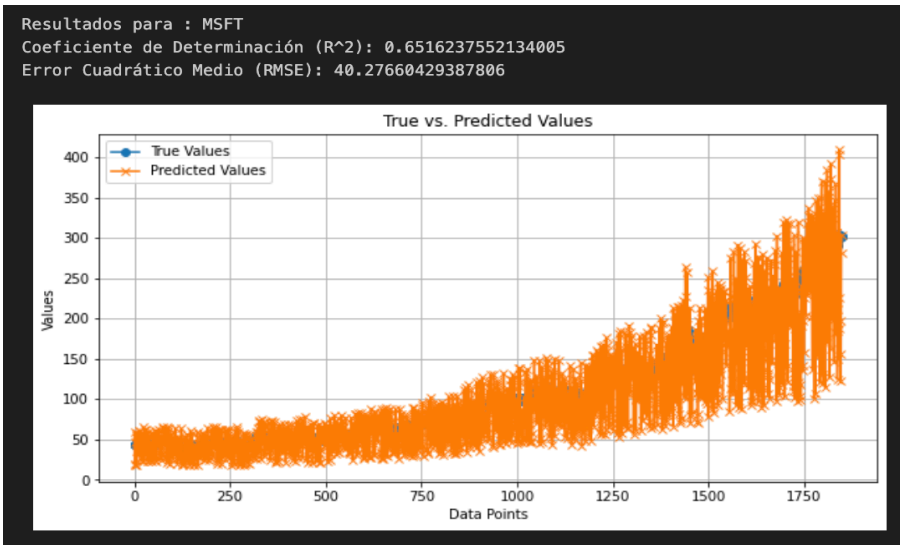
Ilustración 79. Predicción de precios para Apple



Fuente: Elaboración propia.

Microsoft

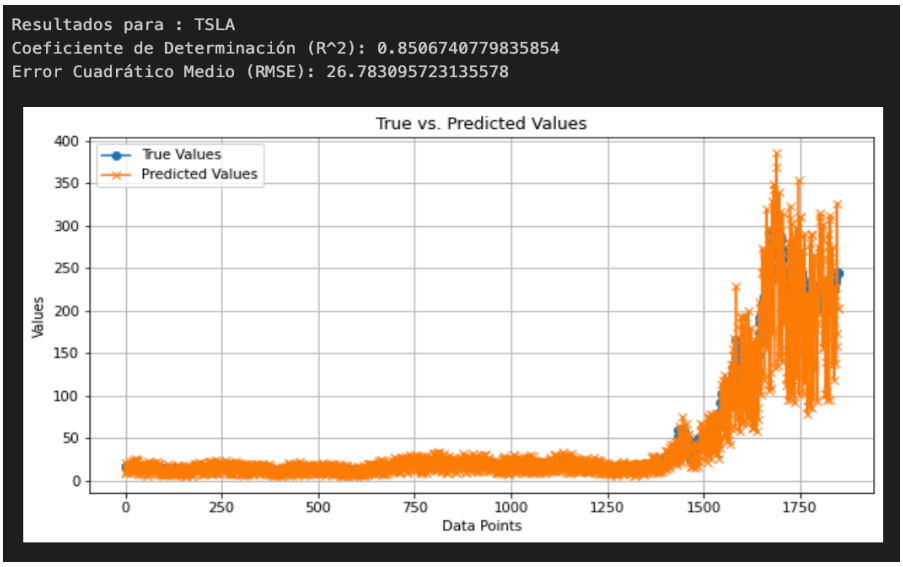
Ilustración 80. Predicción de precios para Microsoft



Fuente: Elaboración propia.

Tesla

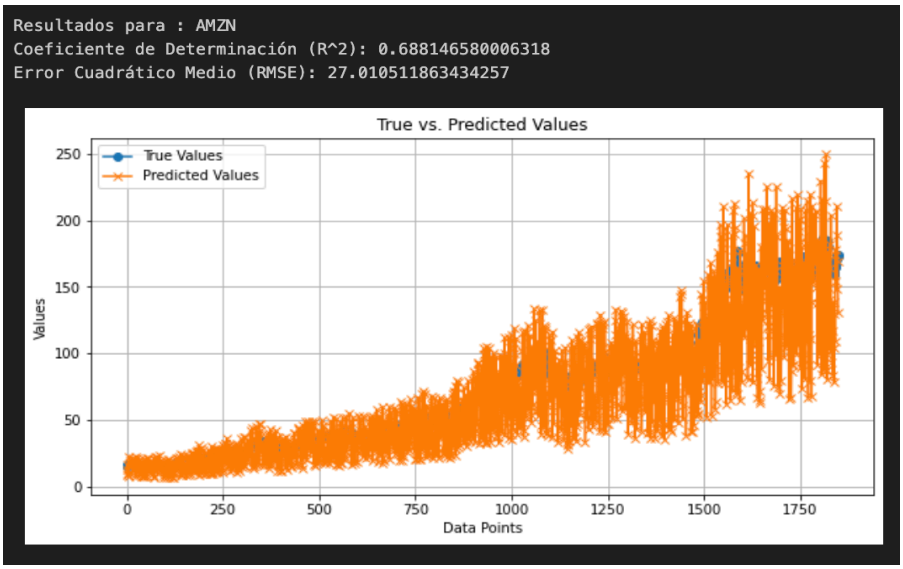
Ilustración 81. Predicción de precios para Tesla



Fuente: Elaboración propia.

Amazon

Ilustración 82. Predicción de precios para Amazon

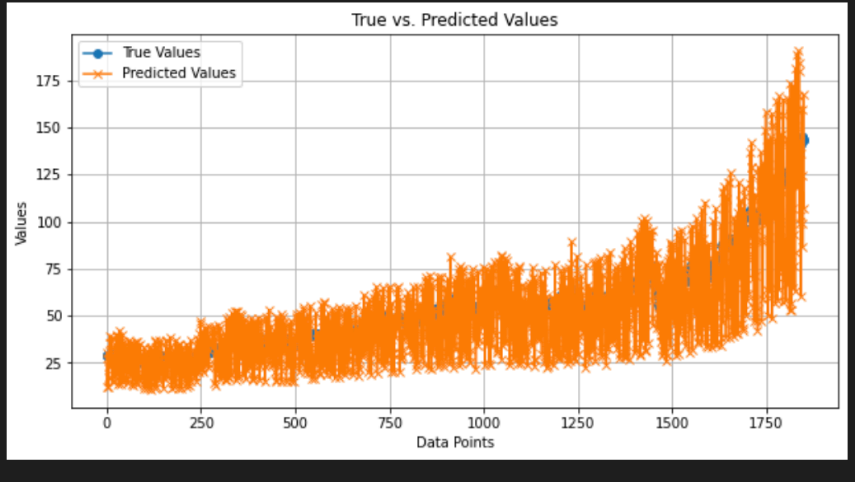


Fuente: Elaboración propia.

Google

Ilustración 83. Predicción de precios para Google

```
Resultados para : GOOGL
Coeficiente de Determinación (R^2): 0.4186124759069366
Error Cuadrático Medio (RMSE): 18.79027158644975
```



Fuente: Elaboración propia.

Creación del portafolio

Una vez hemos realizado la predicción de los precios utilizando el modelo y las variables seleccionadas anteriormente, procedemos a realizar el desarrollo del portafolio, su composición y su política de rebalanceo más óptimo.

Lo primero a tener en cuenta es que para este desarrollo se tomarán algunos supuestos establecidos en la teoría moderna de portafolios de Markowitz..

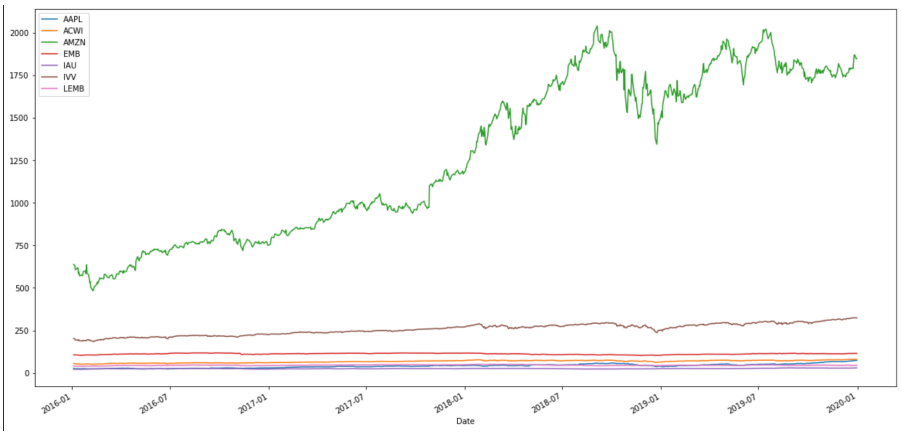
- La teoría se centra en la inversión en activos financieros, como acciones y bonos. No se consideran activos no financieros, como bienes raíces o inversiones directas en negocios. En este caso, el análisis se centrará en inversiones en el mercado de renta variable de los Estados Unidos.

- Se asume que los inversionistas son racionales y buscan maximizar su utilidad

esperada.

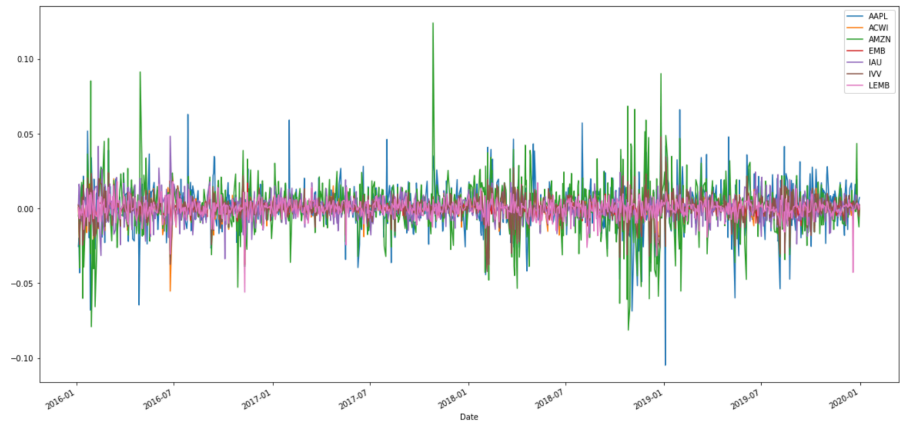
- La teoría se desarrolla para inversionistas con un horizonte de inversión a largo plazo.
- Uno de los supuestos más importantes es que los inversionistas pueden diversificar su portafolio al agregar diferentes activos financieros.
- La teoría presupone que los rendimientos de los activos financieros siguen una distribución normal.
- La teoría asume que los mercados son eficientes, lo que significa que los precios de los activos reflejan toda la información disponible en el mercado.
- Se supone que no hay costos de transacción al comprar o vender activos, y no se consideran implicaciones fiscales en las decisiones de inversión.
- Los supuestos se basan en expectativas estables de los inversionistas sobre los rendimientos futuros y la volatilidad de los activos.

Ilustración 84. Precios principales de Acciones S&P500



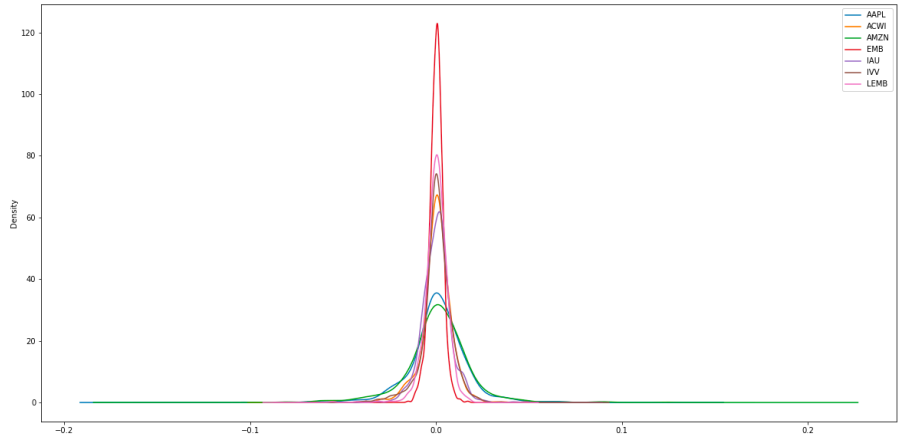
Fuente: Elaboración propia.

Ilustración 85. Retornos principales de Acciones S&P500



Fuente: Elaboración propia.

Ilustración 86. Distribución de retornos principales de Acciones S&P500



Fuente: Elaboración propia.

Para realizar la validación de la normalidad de los retornos de las diferentes empresas, como se mencionaba en las pruebas estadísticas a realizar del modelo, se decidió utilizar la metodología de Anderson-Darling, la cual es una prueba estadística utilizada para evaluar si un conjunto de datos proviene de una población con una distribución específica, como una distribución normal (gaussiana). La prueba fue desarrollada por Theodore W. Anderson y

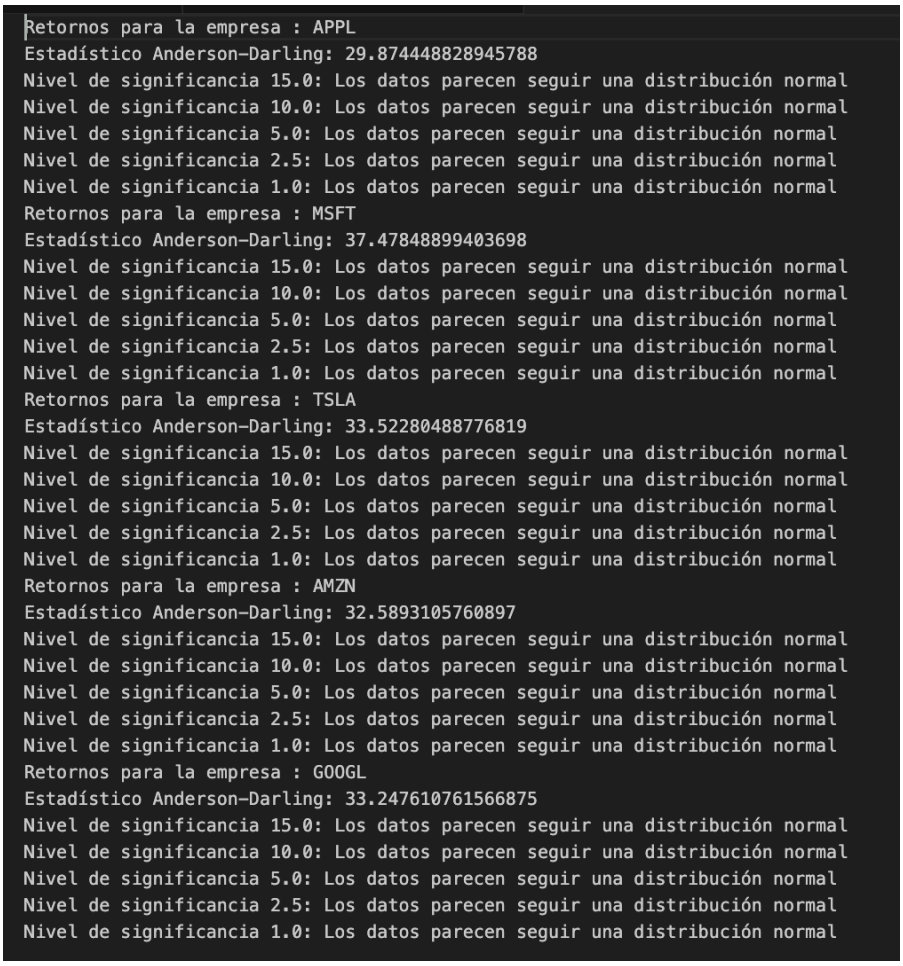
Donald A. Darling en 1952. Se utiliza para verificar si los datos muestrales se ajustan a una distribución teórica o si muestran desviaciones significativas de esa distribución.

La metodología de la prueba de Anderson-Darling implica los siguientes pasos:

1. La hipótesis nula (H0) establece que los datos provienen de la distribución normal. La hipótesis alternativa (H1) sugiere que los datos no provienen de esa distribución.
2. Los datos se ordenan de manera ascendente.
3. Se calcula un estadístico de Anderson-Darling a partir de los datos ordenados.
4. Se determinan los valores críticos a partir de tablas estadísticas para diferentes niveles de significancia (alfa). Estos valores críticos son umbrales que se utilizan para comparar con el estadístico de Anderson-Darling calculado.
5. El estadístico de Anderson-Darling calculado se compara con los valores críticos correspondientes al nivel de significancia deseado.
6. Si el estadístico de Anderson-Darling es menor que el valor crítico, no se rechaza la hipótesis nula, lo que sugiere que los datos se ajustan bien a la distribución teórica. Si es mayor, se rechaza la hipótesis nula, indicando que los datos no siguen la distribución teórica.

Los resultados obtenidos se muestran en la siguiente gráfica, que indica que siguiendo la metodología utilizada de Anderson Darling podemos concluir estadísticamente que los residuos para cada una de las series analizadas siguen una distribución normal en diferentes niveles de confianza.

Ilustración 87. Revisión de la normalidad de retornos



Fuente: Elaboración propia.

Para el desarrollo del portafolio hemos optado por la optimización del *ratio* de Sharpe sobre la optimización por mínima varianza, aunque ambos son dos enfoques clave en la teoría moderna de portafolios. Esto con base en la idea de que un inversionista conservador puede optar por un portafolio de mínima varianza, mientras que un inversionista dispuesto a asumir un riesgo moderado puede preferir un portafolio optimizado por el *ratio* de Sharpe, y para el ejercicio se asume una disposición a tomar riesgo moderada/alta por parte del inversionista.

Portafolio de mínima varianza:

- **Definición:** Un portafolio de mínima varianza es aquel que está diseñado para minimizar la volatilidad o el riesgo total del portafolio.
- **Enfoque principal:** El principal objetivo de un portafolio de mínima varianza es la reducción de la variabilidad de los rendimientos.
- **Adecuado para:** Los inversionistas que tienen una aversión al riesgo significativa y desean proteger su capital contra movimientos bruscos del mercado suelen optar por portafolios de mínima varianza.

Portafolio optimizado por el *ratio* de Sharpe:

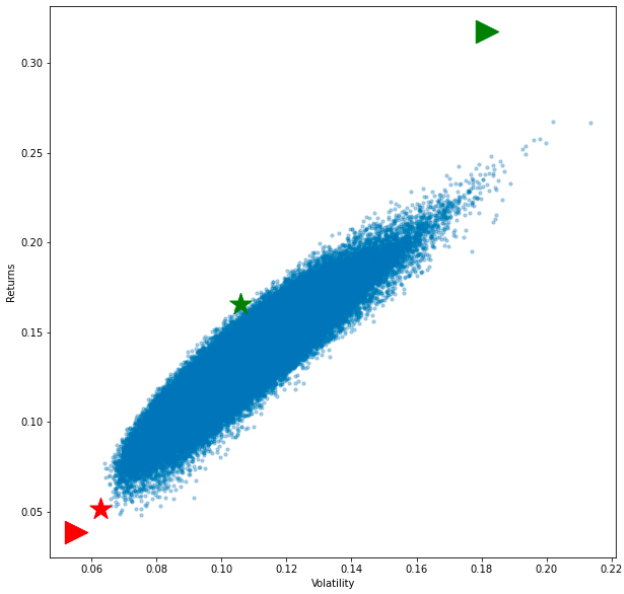
- **Definición:** Un portafolio optimizado por el *ratio* de Sharpe busca maximizar la relación riesgo-retorno.
- **Enfoque principal:** El enfoque principal de este tipo de portafolio es equilibrar el riesgo y el rendimiento.
- **Adecuado para:** Los inversionistas que buscan un equilibrio entre riesgo y rendimiento, es decir, un mayor rendimiento potencial sin asumir un riesgo excesivo, suelen optar por portafolios optimizados por el *ratio* de Sharpe.

Comparación:

- Un portafolio de mínima varianza se centra exclusivamente en la reducción del riesgo y la volatilidad. Su objetivo principal es proteger el capital del inversionista en situaciones de mercado difíciles.
- Un portafolio optimizado por el *ratio* de Sharpe busca el equilibrio entre riesgo y rendimiento, tratando de maximizar el rendimiento ajustado al riesgo. Es adecuado para inversionistas que buscan un mayor rendimiento sin estar dispuestos a asumir un riesgo significativamente más alto.

En la siguiente gráfica se muestra el punto óptimo de una solución analítica y discreta con más de 5.000 simulaciones para ambas metodologías de optimización, con lo cual seguimos reafirmando nuestra posición de realizar el modelo con la metodología de optimización por *Sharpe ratio*:

Ilustración 88. Frontera eficiente



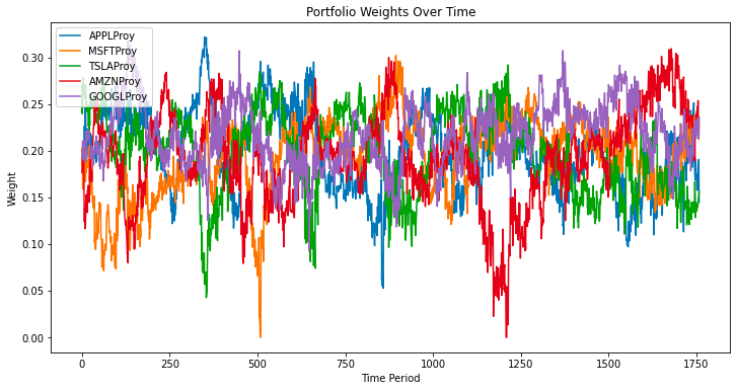
Fuente: Elaboración propia.

Para el periodo comprendido entre el 2015 y el 2020, se decidió realizar rebalanceos trimestrales, teniendo en cuenta que se espera que la metodología presentada pueda ser implementada en la práctica y en esta ya sí se deben tener en cuenta costos operativos, comisiones, impuestos y demás, por lo que un rebalanceo más frecuente no sería óptimo.

A la hora de realizar el análisis de la estrategia y el modelo seleccionados, se decidió tomar diferentes medidas de rentabilidad sobre el *benchmark,* que en este caso sería el S&P500, y para eso se llevó a cabo un análisis de *Sharpe Ratio*, Alfa sobre *benchmark* y Alfa de Jensen

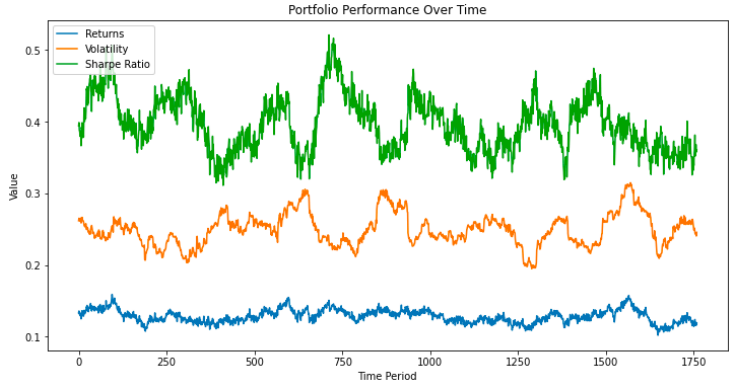
para cada una de las ventanas en las que se realizaron los respectivos rebalanceos. Los resultados son los siguientes:

Ilustración 89. Pesos de los diferentes activos



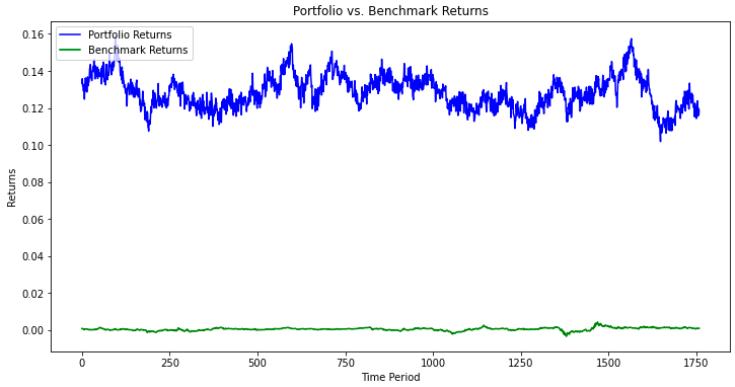
Fuente: Elaboración propia.

Ilustración 90. *Performance*



Fuente: Elaboración propia.

Ilustración 91. Retornos portafolio Vs S&P 500



Fuente: Elaboración propia.

Conclusiones

En este proyecto aplicamos técnicas de aprendizaje supervisado para predecir la tendencia del precio para varias acciones y su posterior aplicación para optimizar un portafolio utilizando la teoría de portafolio moderna. Nuestros hallazgos se pueden resumir en los siguientes aspectos:

Los rendimientos conseguidos mediante el modelo definido presentan consistentemente un rendimiento mayor al del S&P 500, lo cual tendría un impacto significativo en la gestión de inversiones y podría generar un valor sustancial para los inversionistas. Aunque la aplicación de este tipo de modelos en la toma de decisiones financieras tiene limitaciones, como la falta de transparencia y la posibilidad de que existan sesgos, el enfoque aplicado en el análisis tiene el potencial de mejorar la eficiencia y la efectividad de la gestión de activos, y la toma de decisiones de inversión.

La principal ventaja de incorporar el análisis de sentimientos como una variable de entrada en el modelo de predicción es la mitigación del riesgo asociado a noticias o eventos inesperados que podrían impactar el comportamiento de los precios. Sin embargo, la desventaja de este enfoque radica en la obtención de los datos necesarios para construir la variable de "Sentimiento". Esto implica depender de la API de X y la creación de una cuenta que permita realizar múltiples solicitudes, lo que conlleva costos para el desarrollador.

Los resultados obtenidos en esta investigación demuestran que la aplicación de modelos de inteligencia artificial y aprendizaje automático en la industria financiera puede mejorar significativamente la precisión de las predicciones de precios, para posteriormente ser utilizada en la optimización de portafolios. Aplicar la combinación de modelos de inteligencia artificial y aprendizaje automático con la teoría moderna de portafolios puede conducir a mejoras significativas en la toma de decisiones de inversión. La inclusión de datos no estructurados, como noticias y redes sociales, en el análisis financiero, puede mejorar aún más la precisión de las predicciones y proporcionar una visión más completa del mercado.

Comentado [JB2]: No vi en ninguna parte del documento dónde se demostraba esto. Es decir, dónde está la tabla mostrando los retornos esperados vs los reales, tanto por un portafolio de media-varianza tradicional como por un portafolio usando AI y ML.???

Para futuras investigaciones relativas a este campo se recomienda explorar aún más la aplicación de modelos de inteligencia artificial y aprendizaje automático en la industria financiera, especialmente en áreas como la gestión de riesgos. También se sugiere investigar

la aplicación de modelos de inteligencia artificial y aprendizaje automático en otros mercados financieros, como el mercado de divisas y el mercado de materias primas, para los cuales actualmente existe gran cantidad de estudios que referencian este tipo de análisis. Además, que se realicen más investigaciones para evaluar la efectividad de los modelos de inteligencia artificial y aprendizaje automático en diferentes condiciones del mercado y en diferentes períodos.

Estos resultados tienen implicaciones prácticas en la gestión de activos y la toma de decisiones de inversión en la industria financiera, y pueden beneficiar a los profesionales del mercado al proporcionar estrategias de inversión más avanzadas y efectivas. Para implementar este enfoque se necesitaría una infraestructura tecnológica sólida y un equipo de expertos en finanzas y tecnología. En primer lugar, se requeriría una plataforma de datos que pueda recopilar y procesar grandes cantidades de datos financieros estructurados y no estructurados. Estos datos se necesitan para que puedan desarrollar y ajustar los modelos de inteligencia artificial y aprendizaje automático para la previsión de precios y la optimización de portafolios. Estos modelos deberían ser capaces de identificar patrones y tendencias en los datos financieros, y proporcionar recomendaciones precisas y oportunas para la toma de decisiones de inversión. Una vez se tenga el modelo, se requeriría un proceso de validación riguroso para evaluar la efectividad de los modelos de inteligencia artificial y aprendizaje automático en diferentes condiciones del mercado y en diferentes períodos.

Referencias

Álvares, R. (2012). *Un paseo por los derivados financieros.* Trabajo de grado para máster, Universidad de León, Facultad de Ciencias Económicas y Empresariales.

AMV (2006). *Guía estudio renta variable.* https://www.amvcolombia.org.co/wp-content/uploads/2021/01/Guia-Renta-Variable-Operador-enero-2021.pdf.

Antut, G., *et al.* (s. f.). La minería de datos. *Monografías.* https://www.monografias.com/docs112/mineriadedatos/mineriadedatos.shtml.

Attigeri, G. V., *et al.* (2015). Stock market prediction: A big data approach. *TENCON 2015-2015 IEEE Region 10 Conference,* 1-5.

Bandgar, B. M., y Sheeja, S. (2016). Analysis of real time social tweets for opinion mining. *International Journal of Applied Engineering Research, 11*(2), 1404-1407.

Baker, M., y Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives, 21*(2), 129-152.

Banco de la República (2010a). Estadísticas: Banco de la República. *Banco de la República.* https://www.banrep.gov.co/es/estadisticas/trm.

Banco de la República (2010b). Política monetaria: Banco de la República. *Banco de la República.* https://www.banrep.gov.co/es/politica-monetaria/como-se-implementa-politica-monetaria.

Banco de la República (2018). *Banco de la República.* https://www.banrep.gov.co/es/intermediario-del-mercado-cambiario-imc.

BBVA (25 de octubre del 2018). BBVA Trader: Análisis fundamental vs. análisis técnico. *BBVA.* https://www.bbva.com/es/invertir-bolsa-desde-cero-analisis-fundamental-analisis-tecnico/.

Bengio, Y. C., Courville, A., y Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence, 35*(8), 1798-1828.

Bollen, J. M. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8.

Brealey, R. A. (2017). *Principios de finanzas corporativas.* McGraw-Hill.

Campbell, J. Y. (1997). *The econometrics of financial markets.* Princeton University Press.

Cardozo, N., Rassa, J. S., y Rojas, J. S. (2014). Caracterización del mercado de derivados cambiarios en Colombia. Borradores de economía. *Banco de la República,* (860), 2-45.

Caridad y Ocerin, J. M. (1998). *Econometría: Modelos econométricos y series temporales* (Vol. II). Reverté.

Chandni, M., Chandra, N., Gupta, S., y Pahade, R. (2015). Sentiment analysis and its challenges. *International Journal of Engineering Research & Technology, 4*(3), 968-970.

Chang, M.-W., y Devlin, J. (2018). Open sourcing BERT: State-of-the-art pre-training for natural language processing. *Research Scientists, Google AI Language.* https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html.

Chen, A. S., Leung, M. T., y Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers & Operations Research, 30*(6), 901-923.

Community, A. O. (2011). *Open NLP.* https://opennlp.apache.org/.

Correa, J. A., Ramírez, L. J., y Cataño, C. E. (2010). La importancia de la planeación financiera en la elaboración de los planes de negocio y su impacto en el desarrollo empresarial. *Revista Facultad de Ciencias Económicas: Investigación y Reflexión, XVIII*(1), 179-194.

DANE (2022). Balanza comercial. *Departamento Administrativo Nacional de Estadística - DANE.* https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-internacional/balanza-

comercial#:~:text=Informaci%C3%B3n%20agosto%202022&text=La%20balanza%20comercial%20es%20la,e%20importaciones%20de%20un%20pa%C3%ADs.

Dattu, B. S., y Gore D. (2015). A survey on sentiment analysis on Twitter data using different techniques. *International Journal of Computer Science and Technologies.*

Elton, E. J., *et al.* (1965). *Modern portfolio theory and investment analysis.* Wiley.

Fama, E. F., y French, K. R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics, 25*(1), 23-49.

Fama, E. F., y French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance, LXV*(5), 1915-1946.

Frederick, D. (1984). Interpretation of dummy variables in semilogarithmic equations: small sample implications. *Southern Economic Journal, 50*(4), 1185-1188.

Golmohammadi, D. K. (2020). Stock price prediction using hybrid models: A systematic literature review. *Expert Systems with Applications, 142,* 113-169.

Gómez, A. (2020). *Derivados financieros en Colombia, y la nueva regulación de colaterales.* Monografía, Universidad EAFIT, Escuela de Derecho.

Graham, B. (1949). *The intelligent investor.* Harper Business.

Gunasekarage, A., y Power, M. (2001). The profitability of moving average trading rules in South Asian stock markets. *Emerging Markets Review, 2*(1), 17-33.

Gupta, A., Pruthi, J., y Sahu, N. (2017). Sentiment analysis of tweets using machine learning approach. *International Journal of Computer Science and Mobile Computing, 6*(4), 444-458.

Hall, M. (2021). *Investopedia.* https://www.investopedia.com/ask/answers/050115/what-does-dow-jones-industrial-average-measure.asp.

Hernández, J., y Zúñiga, J. (2013). *Modelos econométricos para el análisis económico.* ESIC.

Hull, J. C. (2002). Introducción a los mercados de futuros y opciones. Pearson Education.

Hutto, C., y Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216-225.

InfoWorld (2017). Python vs. R: The battle for data scientist mind share. *InfoWorld.* http://www.infoworld.com/article/3187550/data-science/python-vs-r-the-battle-for-data-scientist-.

Jabreel, M., Moreno, A., y Huertas, A. (2016). Semantic comparison of the emotional values communicated by destinations and tourists on social media. *Journal of Destination Marketing & Management, 6*(3).

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2014). *An introduction to statistical learning.* Springer.

Kenney, J. F., y Keeping, E. S. (1962). Root mean square. En *Mathematics of statistics* (pp. 59-60). Princeton.

Khan, G. F. (2015). *Seven layers of social media analytics: Mining business insights from social media text, actions, networks, hyperlinks, apps, search engine, and location data.* Gohar.

Khurshed, A. G. (2007). *Macroeconomic factors and the performance of value and glamour stocks in emerging markets.*Kimoto, K. A. (1990). Stock market prediction system with modular neural networks. *IJCNN International Joint Conference on Neural Networks, 1,* 1-6.

Kulesa, A., Krzywinski, M., Blainey, P., y Altman, N. (2015). Sampling distributions and the bootstrap. *Nat Methods, 12*(6), 477-478.

Kumar, P. N. (2021). A review on financial forecasting models: A hybrid perspective. *Journal of Ambient Intelligence and Humanized Computing.*

Liu, B. (2015). *Sentiment analysis, mining opinions, sentiments, and emotions.* Cambridge.

Lo, A. W. (2010). *Hedge funds: An analytic perspective.* Princeton University Press.

Loria, S. (2020). *TextBlob.* https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis.

Malkiel, B. G. (2003). *A random walk down wall street.* W. W. Norton & Company.

Markowitz, H. (1952). *The journal of finance.* Wiley.

Mehta, D. J. (2016). Sentiment mining and related classifiers: A review. *IOSR Journal of Computer Engineering, 18*(1), 50-54.

Ministerio de Economía y Finanzas del Perú (2022). *Política económica y social: Ministerio de Economía y Finanzas.* https://www.mef.gob.pe/es/?option=com_content&language=es-ES&Itemid=100694&view=article&catid=23&id=60&lang=es-ES.

Mishkin, F. S. (2007). *The economics of money, banking, and financial markets.* Pearson Education.

NLTK (2017). *Natural Language Toolkit.* http://www.nltk.org.

O'Leary, D. (2015). Twitter mining for discovery, prediction and causality: Applications and methodologies. *International Journal of Intelligent Systems in Accounting and Finance Management, 22*(3), 227-247.

Oviedo, A. F., y Sierra, L. P. (2019). Importancia de los términos de intercambio en la economía colombiana. *Revista CEPAL,* (128), 126-128. https://repositorio.cepal.org/bitstream/handle/11362/44740/RVE128_Oviedo.pdf?sequence=1.

Padró, L., y Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. *LREC2012.*

Pawar, A. B., Jawale, M. A., y Kyatanavar, D. N. (2016). Fundamentals of sentiment analysis: Concepts and methodology. *Sentiment Analysis and Ontology Engineering* (pp. 25-48). Springer.

Penman, S. H. (2010). *Financial statement analysis and security valuation.* McGraw-Hill Education.

Porlles, J., Carlos, Q. A., y Salas, G. (2013). Pronóstico financiero: métodos rápidos de estimación del fondo de maniobra o capital de trabajo estructural. *Revista de la Facultad de Ingeniería Industrial UNMSM, 16*(1), 29-36.

Sabarmati, D. C. (2017). Reliable data mining tasks and techniques for industrial applications. *IAETSD Journal for Advanced Research in Applied Sciences, 4*(7), 138-142.

Sutar, K., Kasab, S., Kindare, S., y Dhule, P. (2016). Sentiment analysis: Opinion mining of positive, negative or neutral Twitter data using hadoop. *International Journal of Computer Science and Network, 5*(1), 177-180.

Rachev, S. T., Stoyanov, F., y Fabozzi, F. (2005). *Advanced stochastic models, risk assessment, and portfolio optimization: the ideal risk, uncertainty, and performance measures.* John Wiley & Sons.

Rahmath, H. (2014). Opinion mining and sentiment analysis - Challenges and applications. *International Journal of Application or Innovation in Engineering & Management, 3*(5), 401-403.

Rapach, M. E. (2005). *The predictive power of macroeconomic variables for stock returns in robust tests.*

Rashid, A., y Muhammad, M. (1972). Stock prices and exchange rates: Are they related? Evidence from South Asian countries. *The Pakistan Development Review, 41*(4), 535-550.

Red Cultural del Banco de la República (2000). *Enciclopedia: Banrecultural.* https://enciclopedia.banrepcultural.org/index.php/Sector_real.

Rodrigo, J. A. (2016a). *Ciencia de datos.* https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple.

Rodrigo, J. A. (2016b). *Rpubs.* https://rpubs.com/Joaquin_AR/233932.

Rodrigo, J. A. (2017). *Ciencia de datos.* https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines.

Rodrigo, J. A. (2020). *Validación de modelos predictivos: Cross-validation.* https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap.

Roiger, R. J. (2017). *Data mining, a tutorial-based primer.* Chapman & Hall/CRC.

Roldán, P. N. (2020). *Economipedia.* https://economipedia.com/definiciones/bolsa-de-valores.html.

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 44*(1.2), 206-226.

Sharpe, W. (1970). *Portfolio theory and capital markets.* McGraw-Hill.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives, 17*(1), 83-104.

Smeureanu, I., y Cristian, B. (2012). Applying supervised opinion mining techniques on online user reviews. *Informática Económica, 16*(2).

Stock, J. H. (2019). *Introduction to econometrics.* Pearson.

Stock, J. H., y Watson, M. M. (2012). *Introducción a la econometría.* Pearson.

Taruel, S. (2021). En qué consisten los pronósticos financieros. *Emburse Captio.* https://www.captio.net/blog/en-que-consisten-los-pronosticos-financieros.

Treynor, J. L. (2014). *Market value, time, and risk.* https://ssrn.com/abstract=2600356.

Trisedya, Y. E. (2015). Stock price prediction using linear regression based on sentiment analysis. *International Conference on Advanced Computer Science and Information Systems (ICACSIS).*

Tsai, C. F. (2020). Sentiment analysis of Twitter data for predicting stock market movements. *Proceedings of the 2020 5th International Conference on Computer and Communication Systems (ICCCS).*

Twitter (s. f.). *Twitter Company Facts.* https://about.twitter.com/company.

Twitter (s. f.). *Twitter Developer Documentation.* https://dev.twitter.com/docs.

Umar, K. I., y Chiroma, F. (2016). Data mining for social media analysis: Using Twitter to predict the 2016 US presidential election. *International Journal of Operational Research in Management, Social Sciences & Education, 1*(1), 118-128.

Valora Analitik (2019). Ecopetrol y Bancolombia, las acciones colombianas con mejor desempeño en EE. UU. *Valora Analitik.* https://www.valoraanalitik.com/2019/04/01/ecopetrol-y-bancolombia-las-acciones-colombianas-con-mejor-desempeno-en-ee-uu/.

Rajput, V., y Bobde, S. (2016). Stock market prediction using hybrid approach. *International Conference on Computing, Communication and Automation.* ACM New York.

Witten, I. H., y Frank, E. (2017). *Data mining. Practical machine learning tools and techniques.* Elsevier.

Wooldridge, J. M. (2008). *Introducción a la econometría. Un enfoque moderno.* Paraninfo.

Zhao, B., He, Y., Yuan, C., y Huang, Y. (2016). Stock market prediction exploiting microblog sentiment analysis. *2016 International Joint Conference on Neural Networks. IEEE Xplore.*

Zhuge, Q., Xu, L., y Zhang, G. (2009). LSTM neural network with emotional analysis for prediction of stock price. *Engineering Letters, 25*(2).