UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Informática

Facultad de Administración y Dirección de Empresas

DESARROLLO DE ESTRATEGIAS DE INVERSIÓN MEDIANTE APRENDIZAJE AUTOMÁTICO: ANÁLISIS MULTIFACTORIAL Y EXPLORACIÓN DE LA EFICIENCIA DE MERCADO

Trabajo Fin de Grado Integrado

Grado en Ingeniería Informática

Grado en Administración y Dirección de Empresas

AUTOR: Ignacio Bernardo Muñoz Felder

Tutor: José Alberto Sanchis Navarro

Tutor: Ismael Moya Clemente

Cotutor/a externo: (recuperado)

Director/a Experimental: (recuperado)

CURSO ACADÉMICO: 2024/2025

Resumen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed nisi turpis, iaculis a pulvinar quis, luctus et lorem. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Nullam vitae purus eros, id auctor dolor. Sed et nisl quis nibh fermentum cursus ut at elit. Etiam condimentum porta leo quis tempor. Quisque commodo lobortis aliquet. Etiam tincidunt, libero ut vehicula euismod, justo augue lobortis sem, et facilisis velit lacus tristique dolor.

Palabras clave: integer, blandit, pharetra, urna, id.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed nisi turpis, iaculis a pulvinar quis, luctus et lore. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Nullam vitae purus eros, id auctor dolor. Sed et nisl quis nibh fermentum cursus ut at elit. Etiam condimentum porta leo quis tempor. Quisque commodo lobortis aliquet. Etiam tincidunt, libero ut vehicula euismod, justo augue lobortis sem, et facilisis velit lacus tristique dolor.

Keywords : integer, blandit, pharetra, urna, id.

Tabla de Contenidos

Glosario de Términos

- ML: Machine Learning o Aprendizaje automático, es una rama de la Inteligencia Artificial que se centra en desarrollar algoritmos que permiten a sistemas informáticos aprender de datos sin ser específicamente programados para una tarea.
- IDE: Aplicación software que provee de herramientas comprensivas para el desarrollo software
- ETF: De sus siglas en inglés (Exchange Traded Fund), son fondos de inversión que replican el comportamiento de un cierto índice, producto o temática de inversión
- Benchmark: En bolsa, un benchmark (o índice de referencia) es un punto de referencia utilizado para medir el rendimiento de una inversión, como un fondo de inversión o una cartera de acciones.
- SMA: De sus siglas en inglés (Simple Moving Average), representa la media móvil de la cotización de un activo.
- CNMV: Comisión Nacional de Mercados de Valores. Es el organismo encargado de la supervisión e inspección de los mercados de valores en España.
- DNN: De sus siglas en inglés (Deep Neural Network) es un tipo de red neuronal artificial compuesta por múltiples capas ocultas entre la capa de entrada y la de salida.

Índice de Figuras

Índice de Tablas

Estructura

Dada la dualidad de un Trabajo fin de Grado Integrado, y con el fin de facilitar la lectura a continuación se expone la división de los contenidos:

1. Introducción

1.1. Contexto

La inversión en activos es un área del comportamiento económico humano que se remonta a periodos antiguos. Ya en el siglo XVII, en imperios como el Español, Holandés o Británico era habitual la interacción entre diversos agentes, con el fin de proveer financiación para descubrir o explotar nuevas rutas marítimas de comercio. Para el dueño del capital, la inversión consiste en posponer el consumo actual de bienes por la esperanza de un mayor o igual poder adquisitivo en el futuro. Dada la naturaleza humana así como sus necesidades y aversión al riesgo, la variedad de activos disponibles donde invertir capital es amplia y cada vez mayor. Según MSCI, el total de activos invertibles en 2023 alcanzaba la cifra de $213 trillones de dólares [0]

A la hora de invertir, los activos suelen agruparse en clases que permiten analizar sus riesgos, retornos y comportamientos bajo diferentes escenarios económicos. Tradicionalmente, estas clases incluyen acciones, bonos, efectivo, bienes raíces y materias primas, aunque en los últimos años han surgido opciones alternativas como criptomonedas, derivados y fondos de cobertura.

Desde una perspectiva cuantitativa, el valor de los activos inmobiliarios domina el panorama global, seguido por el mercado de deuda y, en tercer lugar, el mercado accionario. Si bien las acciones no son la clase de activo más grande en volumen, su relevancia es indiscutible: destacan por su liquidez, transparencia, diversificación y, sobre todo, por su potencial de retorno. De hecho, históricamente el mercado accionario, y en particular el S&P 500 ha sido el vehículo de inversión que ha ofrecido los mayores rendimientos a largo plazo (6.7% desde 1928), superando ampliamente a otros instrumentos tradicionales. [3]

Además, dadas las políticas económicas fiduciarias modernas, la inflación, si bien moderada, es incentivada. Ocasionando perdidas constantes de poder adquisitivo. Las acciones, que en última esencia representan participación subyacente en empresas, cuentan, de forma generalizada, con la capacidad de soportar el efecto inflacionario. A medida que los costes y salarios de empresas se incrementan, las empresas son capaces de trasladar dichos incrementos al consumidor, mediante precios unitarios mayores o bien combatiendo la inflación, por medio de innovaciones tecnológicas continuas y disciplina de costes, al ser estos deflacionarios. Estas son cualidades que otras clases de activos como el dinero o renta fija no poseen.

.

Encontrar ineficiencias que reflejen discrepancias entre valor y precio es un proceso exhaustivo que requiere de un análisis profundo. La capacidad analítica del inversor se ve limitada y herramientas capaces de procesar grandes cantidades de datos y generar modelos de predicción se presentan como una opción interesante y con gran potencial. En este contexto surge la posibilidad de aplicar técnicas de ML en los mercados capitales. Según Tom M. Mitchell, [4] (1997, p. 2), «Se dice que un programa de ordenador aprende de la experiencia E respecto a alguna clase de tareas T y una medida de rendimiento P, si su rendimiento en tareas de T, medido por P, mejora con la experiencia E». Adicionalmente, Arthur L. Samuel (1959, p. 210) añade: «Machine learning es el campo de estudio que otorga a los ordenadores la capacidad de aprender sin ser programados explícitamente.»

Este conjunto de técnicas que conforman el término ML, ya han sido efectivamente desplegadas en la inversión activa. Hoy en día, se conoce como trading cuantitativo al conjunto de estrategias de inversión que utilizan modelos matemáticos y estadísticos para analizar datos de mercado y tomar decisiones de trading. En vez de basar los patrones de compra y venta en la opinión o intuición personal, el trading cuantitativo utiliza algoritmos y sistemas informáticos para identificar oportunidades de inversión.

Uno de los precursores del trading cuantitativo es Jim Simons con so Hedge Fund "Medallion Fund", cuyo fondo de inversión obtuvo la impresionante rentabilidad media del 37% anual, neta de comisiones, entre 1988 y 2021 [5]. Matemático de profesión, Jim Simons decidió focalizar su atención por completo en sus modelos estadísticos, a menudo ignorando juicios humanos o siquiera intentando comprender el razonamiento detrás de las decisiones propuestas por sus modelos. "Más de la mitad de las señales que usábamos eran contraintuitivas; si pasaban los filtros estadísticos, las operábamos» – Jim Simons (citado por Zuckerman, 2019) [6].

Sin embargo, existe también la otra cara de la moneda. En varias ocasiones Jim Simons ha cancelado decisiones propuestas por los modelos. Por ejemplo en 2018 en contra de las recomendaciones de los algoritmos optó por realizar ventas masivas. Es importante recordar, que los mercados no dejan de reflejar el comportamiento de agentes, que dada su naturaleza humana, muestran comportamientos dinámicos y no estacionarios. La estructura de dicho mercado (regulación, comisiones, geopolítica, aparición de participantes, etc) es variable y estrategias que funcionaban sin razonamiento lógico alguno, pueden dejar de hacerlo. Adicionalmente, es común que dichos algoritmos tiendan a realizar el denominado "overfitting", que ocurre cuando un modelo ajusta demasiado sus parámetros a los datos de entrenamiento, incluyendo ruido o anomalías como señales, y por tanto no generalizando bien a nuevos datos.

En el presente trabajo, optamos por el segundo aproximamiento. Explorando en un primer lugar la literatura respecto la generación de retornos en renta variable. Consideraremos diversos estudios sobre factores y su puesta en práctica por inversores. Además plantearemos nuestra propia visión en el desarrollo de algoritmos construyendo sobre estudios previos en la misma diciplina. Esto acelera la elección de técnicas, parámetros, ventanas temporales y datos. Por último, discutiremos y analizaremos

diferentes modelos y evaluaremos los resultados mediante matrices de confusión, rendimientos, volatilidad y otras métricas relevantes para el inversor.

1.2. Motivación

Este trabajo supone la intersección entre mi curiosidad personal y ambas disciplinas estudiadas estos últimos 5 años. Mi interés por los mercados no son una novedad. Como previo Head of Investments del UPV Investment Club he tenido la suerte de poder formar parte de una asociación de estudiantes cuya pasión por los mercados capitales es compartidas. Mi fascinación por la tecnología y la creciente proliferación de sistemas inteligentes es altamente aplicable en estos campos. Tras mi humilde y corta experiencia operando en los mercados, reconozco la dificultad psicológica de mantenerse fiel a estrategias de inversión. Los mercados son altamente emocionales y ponen a prueba la capacidad del inversor en cada ciclo. Las máquinas en contraste no padecen de esta debilidad. Con este trabajo pretendo alimentar mi conocimiento sobre la magnitud de la oportunidad que ofrecen sistemas inteligentes al proceso inversor, así como ganar las habilidades necesarias para considerarlo como una posible salida profesional.

1.3. Objetivos

De forma general, este trabajo tiene como objetivo desarrollar y validar estrategias de inversión en acciones, basadas en modelos de ML, que generen retornos ajustados al riesgo superiores a los índices S&P 500, mediante algoritmos de ML. Como se ha mencionado previamente, este trabajo supone la integración de conceptos en las disciplinas de Finanzas y Ciencia de Datos. Por ello se ha considerado relevante dividir el anterior objetivo en 2 partes.

OG-1: Recorrer los fundamentos teóricos y estrategias de inversión sobre los que extraer variables claves con capacidad predictiva para entrenar los algoritmos.

OG-2: Abordar el desarrollo técnico de algoritmos supervisados de clasificación según la predicción de rendimiento superior al mercado

Adicionalmente, el trabajo pretende abordar una serie de objetivos específicos mencionados a continuación

- 1. Proveer el contexto necesario al que se enfrenta el inversor, alternativas y ventajas de cada tipo de gestión
- 2. Resaltar la capacidad del ML en la inversión y su implementación actual en la construcción de portfolios inteligentes

- 3. Definir de forma precisa el problema a intentar optimizar mediante ML por parte del inversor
- 4. Construcción de un Dataset histórico mensual sobre empresas constituyentes del SP500
- 5. Diseñar un Pipeline en Python que descargue, normalice y almacene los datos sobre el universo
- 6. Evaluar la existencia de Alpha en un portfolio cuya distribución de pesos ha sido basada según un algoritmo supervisado de clasificación
- 7. Comparar diversos algoritmos de ML según su capacidad de predecir outperformance
- 8. Discutir la implementación y optimización de dichos algoritmos mediante la creación de portfolios optimizados
- 9. Categorizar la capacidad predictiva de diversos ratios y variables sobre el retorno de acciones según su magnitud

1.4. Impacto Esperado

A primera vista, el desarrollo de mejores y más avanzadas técnicas de inversión mediante ML tienen como impacto esperado, el de posibles ganancias económicas derivadas de su uso por parte de aquellas organizaciones o individuos que las desarrollan. Para los analistas financieros, suponen un ahorro de tiempo al acelerar el cribado de señales predictivas.

Más allá del lucro económico que estos pueden derivar, el desarrollo de modelos más eficientes permite entender mejor los mercados capitales. Al identificar posibles ineficiencias y explotarlas, dichos modelos hacen del mercado un lugar más eficiente donde el capital se reparte de correctamente entre los agentes que lo demandan para sus proyectos, contribuyendo así al desarrollo de la economía y evolución tecnológica.

En caso de que los modelos no consigan batir los resultados de una cartera indexada, el trabajo aporta información al inversor sobre la complejidad de la tarea, que puede llegar a ser solo realizable mediante una mayor infraestructura, capacidad de cómputo y datos, que los empleados en este trabajo. Esta información provee al inversor particular la tranquilidad de que indexarse pueda ser de hecho la mejor opción disponible dados sus recursos.

En relación con los ODS, el fomenta la innovación en la industria de gestión de activos y democratizando el acceso a herramientas de análisis avanzado, reduciendo por tanto la brecha entre grandes instituciones y pequeños inversores. En 2015, los líderes mundiales adoptaron un conjunto de objetivos globales para erradicar la pobreza, proteger el plantea y asegurar la prosperidad para todos como parte de una nueva agende de desarrollo sostenible. Entre estos objetivos el trabajo se relaciona con:

- **Fin de la pobreza (ODS-1)**, el ahorro e inversión en mercados de capitales libres, temática en la que se basa este proyecto, son receta que ha servido en el pasado, y servirá en el futuro para acabar con la pobreza. Por tanto para acabar con la pobreza tendremos que seguir ahorrando, invirtiendo y poder hacerlo de forma libre.

- **Educación de calidad (ODS-4)**, el presente trabajo promueve el acceso a fuentes de información de calidad, como se ha demostrado accediendo a bases de datos profesionales, También busca crear conocimiento excepcional a través del análisis concienzudo y detallado. Es por ello que considero que este proyecto cumple con el objetivo de fomentar la educación de calidad.

- **Trabajo decente y crecimiento económico (ODS-8). E**l presente trabajo promueve la profesión de analistas de mercado e inversores. Una profesión que de forma común no es entendida en amplios sectores de la sociedad, pero que realmente aporta un valor añadido a través de la alocación eficiente de recursos. Esta eficiencia es la que acaba produciendo el crecimiento económico, del que finalmente la sociedad en su conjunto se beneficia.

1.5. Metodología

El trabajo emplea dos metodologías de forma consecutiva con el fin de estructurar el desarrollo del modelo correctamente y augmentar la capacidad preductiva del resultado final. Cada una de las metodologías se aferra al cumplimiento de cada objetivo general descrito en la sección de Objetivos. Cabe mencionar que se entrará en más detalle a cerca de la selección de dichos procesos en sus respectivos apartados.

<u>Metodología 1. “Feature selection”</u>

1. Comprensión del Problema que se busca optimizar. En esta sección hacemos una introducción a la predicción de retornos mediante ML, proveemos el contexto ¿Qué supone predecir retornos? Focalizando así nuestro estudio hacia ciertas áreas específicas
2. Revisión Bibliográfica y estudio sobre estrategias tradicionales de inversión
3. Técnicas de selección de variables clave que parecen tener cierto éxito en predecir *outperformance*
4. Propuesta sobre variables elegidas
5. Análisis y selección del proveedor

Metodología 2: "ML Pipeline"

El pipeline de ML se refiere a la cadena de procesos automatizados y reproducibles que llevan a unos datos crudos hasta un modelo entrenado y validado. Aunque existen diversas variantes que enfatizan más ciertas partes del desarrollo, en este caso nos centraremos en los siguientes procesos:

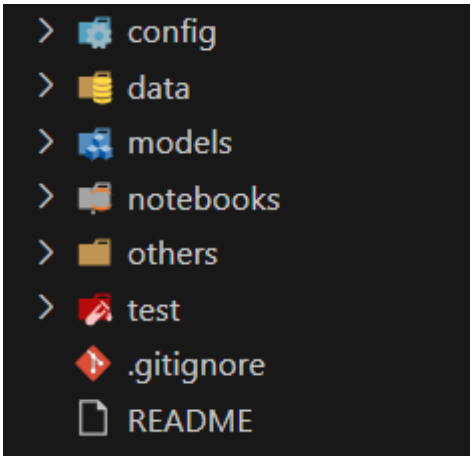
1. Extracción de Datos. También conocido como *Data Ingestion*
   - Realizar conexión a fuente de datos mediante API
   - Extracción de datos
   - Almacenamiento estructurado (tabal ancha) y persistente (.csv)
2. Limpieza y preprocesado
   - Tratar valores faltantes
   - Eliminación de columnas duplicadas y estandarización
3. Ingeniería de Características. También conocido como *Feature Engineering*
   - Creación de nuevas variables
   - Escalado, normalización y codificación de variables categóricas
4. Selección Final de Variables
   - Filtros estadísticos: Correlación y tes de hipótesis
   - Métodos wrapper y embedded
   - Reducción de dimensionalidad en caso de proceder
5. Partición del Set
   - Emplearemos Time Series Split. Se trata de una técnica útil para estimar de forma más robusta el desempeño de modelos. Similar a K-Fold Cross Validation permite reducir la varianza de la estimación al entrenar y validar el modelo en particiones pero respetando el orden temporal.
6. Entrenamiento y ajuste de modelos
7. Evaluación.
8. Presentación de Resultados

En cuanto a la implementación práctica:

Se ha elegido como **IDE,** VS Code con la extensión oficial de Jupyter notebooks. Este entorno permite lanzar, depurar y versionar notebooks de forma integrada, facilitando la iteración rápida en cada etapa

La creación de un entorno virtual aislado ha permitido el control de versiones de paquete y evitar conflictos entre estas. Adicionalmente facilita la futura reproducibilidad del pipeline por el lector. También se ha optado por una estructura de carpetas que facilita la ejecución y desarrollo de la solución

Figura 1: Estructura de Carpetas



Fuente: Elaboración Propia

ACTUALIZAR

```
THESIS-ML-NASDAQ/
│
├── config/ # Configuración de parámetros y rutas
├── data/ # Datos en formato .csv (no es necesario replicar la descarga)
├── models/ # Modelos entrenados y/o scripts de entrenamiento
├── notebooks/ # Notebooks con todo el flujo de procesamiento
│   ├── 1. data_extraction.ipynb
│   ├── 2. data_join.ipynb
│   ├── 3. return_variable.ipynb
│   ├── 4. riskadjusted.ipynb
│   ├── 5. data_cleansing.ipynb
│   ├── 6. data_description copy.ipynb
│   ├── 7. feature engineering copy.ipynb
│   ├── 8. base_model.ipynb
│   ├── 10. XGBoost.ipynb
│   ├── 11. XGBoost-Probabilistic.ipynb
│   ├── en progreso.ipynb
│   └── prices.zip
├── others/ # Otros archivos y recursos útiles
├── test/ # Tests automáticos (si aplica)
├── .gitignore
└── README.md
```

Se ha empleado tecnología git, en concreto haciéndose uso de Git-Hub, para asegurar la persistencia del trabajo y la correcta monitorización del desarrollo entre tutor y alumno. Por último, los datos usados en el entrenamiento de modelos han sido extraídos de 2 proveedores altamente reconocidos en el sector financiero: Nasdaq Data Link y LSEG Data Platform. Se hará más hincapié en su selección en futuros capítulos.

1.6. Estructura del documento

DIAGRAMA

1.7. Convenciones

Con el fin de facilitar el seguimiento por parte del lector. Se han decidido las siguientes convenciones

- Se empleará snippets con tipografía `consolas` para el código fuente
- Palabras extranjeras serán remarcadas en *cursiva*.
- Las citas textuales por otros autores serán ilustradas “entre comillas”
- La primera vez que aparezca un término disponible en el glosario de términos será remarcado con una asterisco *

2. Estado del Arte

La cuestión sobre si es posible generar rendimientos superiores a los del mercado ha sido uno de los debates centrales en las finanzas modernas. En esta sección introducimos al lector en la inversión activa, en el significado de generar *outperformance*. Además proveemos una revisión de la literatura tanto sobre el los modelos que han intentado dar una explicación a la generación de rendimientos como a su aproximamiento mediante modelos de aprendizaje. Finalmente revisamos otras estrategias discutidas con menor frecuencia en la literatura.

2.1. Inversión Activa y Pasiva

De forma general, existen dos aproximaciones principales a la inversión en acciones: la inversión activa y la inversión pasiva.

La inversión pasiva busca replicar el comportamiento de un índice de referencia, como el S&P 500, MSCI World o STOXX Europe 600. Esta estrategia parte de la hipótesis de que los mercados financieros son mayormente eficientes, lo que implica que resulta extremadamente difícil, y costoso, obtener rendimientos consistentemente superiores al promedio del mercado. Por tanto, para la mayoría de inversores, la mejor opción suele ser exponerse de forma diversificada a todo el mercado a través de vehículos como los fondos indexados o los ETFs* (Exchange Traded Funds). La popularización de estos productos en las últimas décadas ha facilitado el acceso a la inversión pasiva para todo tipo de inversores, reduciendo los costes y aumentando la transparencia y liquidez (Ferri, 2017).

En contraste, la inversión activa implica la búsqueda constante de acciones individuales cuyo futuro retorno se espera sea superior al resto del mercado. Mediante la construcción de una cartera compuesta por varias de estas oportunidades, el inversor activo trata de superar el rendimiento del índice de referencia. Esta aproximación requiere un análisis exhaustivo, selección de valores, gestión dinámica, y asume mayores costes de transacción, comisiones y, en muchos casos, una mayor volatilidad de resultados. La base teórica de la inversión activa radica en la creencia de que el mercado no es perfectamente eficiente y que existen ineficiencias temporales o estructurales (por ejemplo, errores de valoración, reacciones excesivas o lentitud en la incorporación de nueva información) que pueden ser aprovechadas para obtener *alpha* (Jensen, 1968).

Numerosos estudios académicos y reportes del sector financiero han demostrado que la mayoría de los gestores activos no logran batir de forma consistente a los índices, especialmente después de descontar comisiones y gastos de gestión. Por ejemplo, según el informe anual de SPIVA (S&P Indices Versus Active) de 2023, más del 85% de los fondos de gestión activa en EE. UU. quedaron por detrás del S&P 500 en periodos de 10 y 15 años (S&P Dow Jones Indices, 2023). Este fenómeno se ha observado en mercados desarrollados y emergentes, así como en diferentes clases de activos, reforzando la idea de que "el mercado" suele ser difícil de superar de manera persistente.

No obstante, la inversión activa sigue teniendo defensores y un papel relevante en los mercados. Las posibles recompensas para quienes logran *outperformance* son elevadas, y el atractivo del interés compuesto es poderoso.

Como ejemplo, un inversor que invierte 10,000 € durante 30 años al retorno histórico del S&P 500 (6.7% anual) obtendría 74,217 €. Si en cambio consiguiera un *outperformance* sostenido del 2% anual (8.7% anual), alcanzaría casi el doble: 134,716 €.

En resumen, la literatura y los datos empíricos respaldan la idea de que la inversión pasiva suele ser la mejor opción para la mayoría de los inversores a largo plazo. Sin embargo, la existencia de estrategias activas ganadoras, aunque poco frecuentes, motiva la investigación en nuevos enfoques, como el uso de aprendizaje automático para detectar patrones y anomalías que puedan generar rendimientos superiores a los del mercado

En los últimos años, la proliferación de algoritmos y el desarrollo de nuevas técnicas de aprendizaje automático han generado un enfoque intermedio entre la gestión activa y la pasiva. Conocidas como "smart portfolios" o carteras inteligentes, estas estrategias semi-pasivas seleccionan activos que cumplen con una serie de filtros o factores cuantitativos históricamente asociados a la obtención de mayores retornos ajustados al riesgo, como valor, calidad, momentum, o baja volatilidad (Profiles Software, 2023). Estos enfoques automatizados permiten a los gestores diseñar carteras que se benefician de reglas sistemáticas basadas en datos, reduciendo el sesgo humano y los costes de transacción asociados a la selección manual, pero manteniendo la posibilidad de obtener rendimientos superiores a una simple réplica del índice.

Así, la frontera entre gestión activa y pasiva se difumina, dando lugar a estrategias híbridas que buscan combinar lo mejor de ambos mundos: la eficiencia, simplicidad y bajo coste de la indexación con la capacidad de capturar primas de riesgo adicionales a través de criterios cuantitativos y modelos predictivos avanzados.

2.2. SP500 como referencia de Mercado

Como se ha mencionado previamente, los gestores comparan sus rendimientos frente a los de un *benchmark**. Esto facilita a los partícipes evaluar fondos de inversión.

El SP500 (Standard & Poor's 500) es un índice bursátil frecuentemente usado como *benchmark*. Este está compuesto por las 500 empresas de mayor capitalización en Estados Unidos. Dada la gran importancia de la economía estadounidense sobre el conjunto global, el SP500 es reconocido como una medida de rendimiento de renta variable. Creado en 1957 por Standard & Poor's, incluye empresas como Apple, Microsoft o Tesla, abarcando el 80% de la capitalización de mercado total en los Estados Unidos

El S&P500 es considerado un benchmark ideal por varias razones:

- Alta diversificación sectorial
- Alta liquidez dado que sus componentes son diariamente intercambiados
- Transparencia y replicabilidad: Cuenta con un gran conjunto de productos indexados (ETF's) que replican fielmente su comportamiento, así como información disponible que facilite el análisis de sus componentes

Además, otro hecho interesante es su carácter dinámico. Dado que se compone de empresas a las cuales nos referiremos como constituyentes. Cuando una empresa deja de pertenecer al conjunto de 500 empresas más relevantes, el índice deja de replicar su comportamiento y añade un nuevo constituyente.

2.3. Predecir Retornos

La noción de Valor Actual Neto (VAN) establece que el valor de una empresa equivale al valor de los flujos de caja futuros esperados, descontados a valor presente, menos la inversión inicial requerida.

$$VAN = \sum_{t=1}^{n} \frac{F_t}{(1+k)^t} - I_0 \quad (1)$$

En este sentido, los precios de las acciones, que representan participaciones en dichas empresas, reflejan las expectativas colectivas del mercado sobre esos flujos futuros. Los retornos de una acción corresponden a los cambios en su precio a lo largo del tiempo, lo

cual implica, en última instancia, cambios en las expectativas sobre los flujos de caja futuros de la empresa.

Predecir retornos consiste, por tanto, en anticipar cómo y cuándo varían estas expectativas de mercado. Es habitual utilizar la evolución de los flujos de caja históricos para analizar la trayectoria financiera de la empresa y, junto a otras variables relevantes, construir estimaciones sobre sus resultados futuros. Estas predicciones, descontadas a valor presente, proporcionan la mejor aproximación al valor intrínseco de la empresa en cada momento.

2.4. Introducción al Outperformance

En la disciplina de gestión patrimonial, el *Outperformance* se refiere a inversiones, carteras o gestores que muestran un rendimiento superior al mercado. En este contexto, se entiende el mercado como la cesta amplia de productos directamente comparables en términos de riesgo, rentabilidad y clase de activo. (El concepto de mercado será desarrollado en detalle en el siguiente apartado.)

La Teoría de los Mercados eficientes (Fama,1970) postula que los precios de los activos reflejan toda la información disponible de forma rápida y sin sesgos sistemáticos. La derivación lógica por tanto, es que la única fuente de rentabilidad adicional es la prima recibida por asumir mayor riesgo. En concreto, el tipo de riesgo no eliminable, incluso tras una correcta diversificación de la cartera, es decir el riesgo sistemático.

Esta intuición se ve reflejada en uno de los pilares fundamentales de la valoración de activos, el CAPM (Capital Asset Pricing Model), inicialmente presentado por Sharpe (1964) y refinado por Linter (1965) y Mossin (1966). Este modelo propone un aproximamiento al cálculo del rendimiento sobre activos individuales, sugiriendo que este se encuentra linealmente relacionado con la rentabilidad media del mercado. De esta forma, el rendimiento esperado de un activo se deriva como la tasa libre de riesgo (Rf) más la prima de mercado ($E(Rm) - Rf$) ajustada por la sensibilidad (β) al mercado. Dicha sensibilidad mide la volatilidad o riesgo sistemático de una acción con relación al mercado.

$$E(Ri) = Rf + \beta i * (E(Rm) - Rf) \quad (2)$$

Posteriormente, Jensen (1968) conceptualizó el Alpha, definido como la rentabilidad adicional obtenida por un activo o cartera, que no puede ser explicada por su nivel de

exposición al riesgo sistemático del mercado. Esto es, dado el retorno del Portfolio (Rp), Alpha se calcula como:

$$Alpha = Rp - [Rf + \beta(Rm - Rf)] \quad (3)$$

Otro concepto, también ideado por W. Sharpe en 1966, fue el Sharpe-Ratio. Como es lógico, lo inversores no solo demandan rendimientos satisfactorios, sino una volatilidad adecuada a sus necesidades de liquidez y estabilidad.

Por ello de forma común, los fondos de inversión son comparados en términos de retornos ajustados por riesgo. Esto es precisamente lo que refleja el Sharpe-ratio. En esencia, permite a los inversores conocer cuanto retorno en exceso se obtiene por el riesgo que toman en una inversión

$$Sr = \frac{E[Re - Rf]}{\sigma} \quad (4)$$

2.5. Teoría de Factores

Sin embargo, la evidencia empírica muestra que es posible seleccionar activos que superan sistemáticamente al mercado, incluso después de ajustar por volatilidad o beta. Esto sugiere que existen otras fuentes de riesgo y retorno que el CAPM no captura.
Para explicar estas anomalías, en 1993 surgen los modelos multifactoriales, destacando el modelo de tres factores de Fama y French. Además del factor de mercado, Eugene Fama y Kenneth French (1992, 1993) identifican dos primas sistemáticas adicionales:

- **SMB (Small Minus Big):** recoge la tendencia de las empresas de pequeña capitalización a superar a las grandes.

- **HML (High Minus Low):** refleja la mayor rentabilidad de las empresas con alta relación valor contable/precio (value) frente a las de baja relación (growth).

Este modelo mostró que muchos *alphas* identificados con CAPM desaparecen al controlar por estos dos riesgos sistemáticos adicionales.
Poco después, Carhart (1997) introduce un cuarto factor:

- **MOM (Momentum):** representa la persistencia en el comportamiento de los activos, es decir, aquellos que han tenido buenos rendimientos recientes tienden a seguir haciéndolo bien a corto plazo.

En 2015, Fama y French amplían de nuevo su esquema, apoyados en bases de datos más completas. Mantienen el factor de mercado y SMB, pero reemplazan HML por dos nuevos factores:

- **RMW (Robust Minus Weak):** captura la rentabilidad operativa robusta frente a la débil.
- **CMA (Conservative Minus Aggressive):** diferencia entre empresas con políticas de inversión conservadoras y agresivas.

Su evidencia apunta a que la importancia del factor "value" clásico disminuye cuando se incluyen RMW y CMA, reinterpretando así la naturaleza del alfa en muchas estrategias.
En los últimos años, la literatura financiera ha identificado aún más factores, como quality, low volatility o liquidity. El modelo Q-Factor (Hou, Xue y Zhang, 2015) combina rentabilidad esperada (ROE), inversión y tamaño desde una óptica de economía real.
Este fenómeno, conocido como el "factor zoo", refleja la proliferación de nuevos factores candidatos a explicar los rendimientos. Así, el concepto de *outperformance* se vuelve cada vez más exigente: cuantos más factores explicativos se incorporan, menor es el espacio para obtener un *alpha* genuino que no pueda explicarse por riesgos sistemáticos identificados.

2.6. ML aplicado a Outperformance

El uso de técnicas de aprendizaje automático (ML) para la predicción de retornos bursátiles ha experimentado un crecimiento significativo en la última década, gracias a la disponibilidad de poder computacional y de grandes cantidades de datos sobre los que se entrenan los modelos.

Mientras que los enfoques tradicionales de “asset pricing” se basaban en modelos, la disponibilidad de grandes volúmenes de datos financieros y el avance de las capacidades computacionales han permitido a los investigadores explorar relaciones no lineales y de alta dimensión entre características de las empresas y sus rendimientos futuros.

Uno de los trabajos más influyentes en este campo es el de Gu, Kelly y Xiu (2020), quienes comparan múltiples métodos de aprendizaje automático para estimar las primas de riesgo esperadas de acciones individuales. Su estudio demuestra que modelos como redes neuronales profundas y random forests superan sistemáticamente a las regresiones lineales tradicionales en la predicción de retornos mensuales. Los autores modelan la expectativa condicional del retorno a un mes vista como una función de más

de 90 características financieras y macroeconómicas, mostrando que ML puede captar interacciones complejas entre señales predictoras.

Aunque gran parte de la literatura se centra en horizontes mensuales, algunos autores han explorado ventanas más largas. Por ejemplo, Asness, Frazzini y Pedersen (2014) estudian el factor value y el momentum en ventanas anuales, mostrando que muchas anomalías persisten a largo plazo.

Alternativamente, en un estudio reciente centrado específicamente en el S&P 500, se explora la rentabilidad de utilizar algoritmos de aprendizaje automático para seleccionar subconjuntos de acciones a partir de factores cuantitativos como características (Rodríguez et al., 2024). El trabajo utiliza algoritmos basados en árboles (Decision Tree, Random Forest y XGBoost) por sus capacidades explicativas (white-box), lo que permite analizar la importancia relativa de las variables predictoras en diferentes momentos del tiempo. A través de un sistema de backtesting con rebalanceo periódico, se demuestra que las carteras generadas superan al índice de referencia, aunque con mayor riesgo. Una contribución clave del estudio es mostrar cómo la importancia de los factores cambia con el tiempo, lo que ofrece información valiosa para diseñar estrategias dinámicas y explicables de inversión basada en ML.

2.7. Introducción de Modelos

Entre los modelos empleados en finanzas cuantitativas, destacan con frecuencia los modelos no paramétricos. Estos se caracterizan por no tener una forma fija, como un polinomio. Esto hace a los modelos no paramétricos más flexibles, además a medida que nueva información se crea con el paso del tiempo, los modelos no paramétricos tienen la capacidad de incorporarla augmentando el número de parámetros. Un aproximamiento común suele ser usar un modelelo lineal como base y comparar con otros modelos. En este caso aplicamos regresión logística como modelo base (justificar)

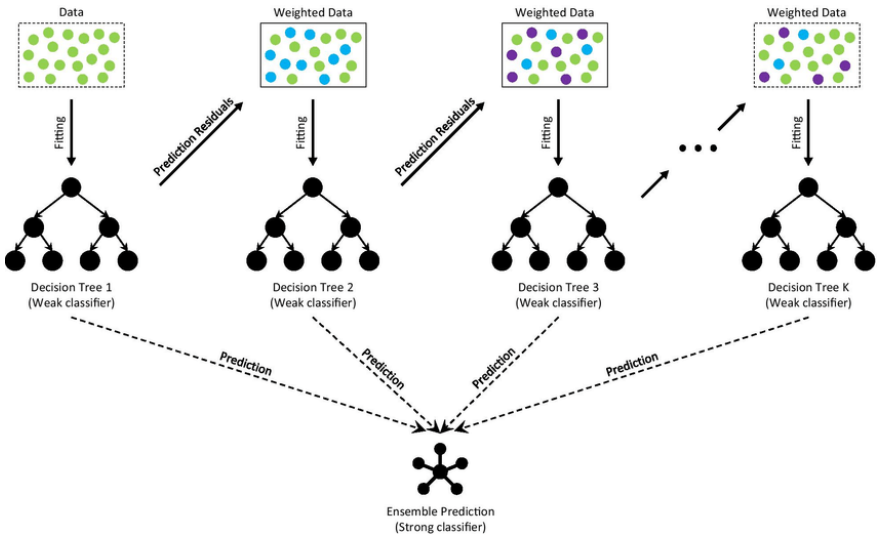
Los modelos no paramétricos como Random Forest, XGBoost y Redes Neuronales son populares porque los mercados no siguen relaciones lineales simples, y necesitas esa flexibilidad para capturar patrones más complejos y no obvios. (mercado relaciones no lineales?)

Ridge

XGBoost

XGBoost es una implementación eficiente y optimizada del algoritmo Gradient Boosting Decision Trees (GBDT). Esta es una técnica de aprendizaje automático basada en arboles de decisión que construyen un modelo fuerte a partir de la combinación secuencial de modelos débiles, habitualmente árboles poco profundos. Como se puede observar en la Figura 2, en cada iteración se entrena un nuevo árbol para corregir los errores cometidos por el conjunto de árboles anteriores, en un proceso aditivo y orientado por gradiente. Este procedimiento se repite hasta alcanzar un número de K de iteraciones o hasta minimizar la función de pérdida.

Figura 2:Estructura secuencia de XGBoost



Fuente: Song et al. (2021), "Ensemble learning for the early prediction of neonatal jaundice with genetic features."

El objetivo es la minimización de una función de pérdida $\mathcal{L}$ que combina el error de predicción con un término de regularización. Este último penaliza la complejidad del modelo. Formalmente, para una predicción $\hat{y}_i$ del ejemplo i, se tiene:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Donde:

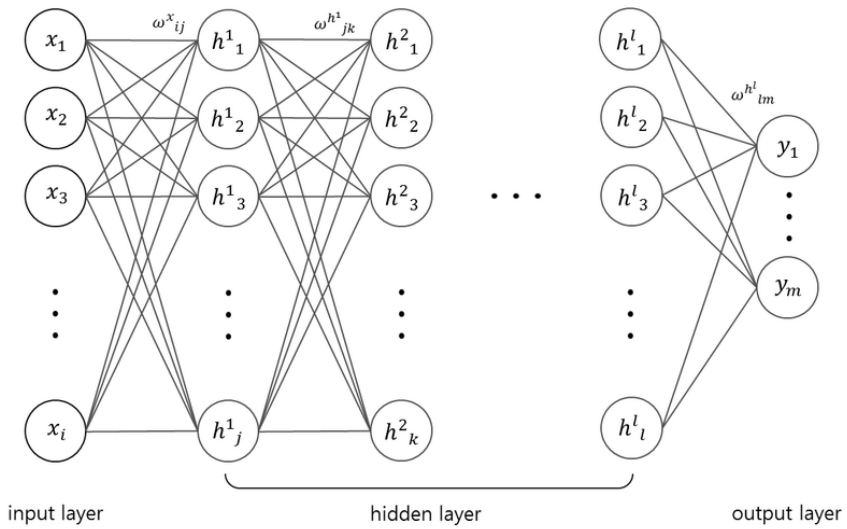
- $\mathcal{L}^{(t)}$ es una función de pérdida diferenciable (como log-loss o squared error)
- f_t es el nuevo árbol que se añade en la iteración t,
- $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ es un término de regularización que penaliza el número de hojas T y los pesos w_j de cada hoja, para evitar sobreajuste.

XGBoost destaca por su capacidad para modelar relaciones complejas no lineales (como las presentes en datos financieros) sin requerir ingeniería de variables manual intensiva, además de su robustez ante sobreajuste gracias a la regularización. Adicionalmente, mantiene ventajas comunes de los árboles de decisión como la posibilidad de interpretar la importancia de cada variable. Por último, gracias a su eficiencia y paralelización, XGBoost reacciona bien a *datasets* grandes incluso con valores faltantes y no requiere normalización. En este trabajo se hace uso de la versión de clasificación de XGBoost, mediante su librería en Python, XGBClassifier.

Redes Neuronales Profundas

Una red neuronal es un modelo computacional inspirado en el funcionamiento del cerebro humano, diseñado para resolver problemas de clasificación, predicción o reconocimiento de patrones. La red se compone de nodos (neuronas) organizadas en capas, que procesan datos de entrada, identifican patrones y generan salidas útiles, como probabilidad de pertenecer a un conjunto etiquetado. Una red neuronal profunda, DNN* es un tipo de red neuronal artificial compuesta por múltiples capas ocultas entre la capa de entrada y la de salida. Cada una de estas capas oculta aprende representaciones no lineales cada vez más abstractas, esto convierte a las DNN en potentes herramientas frente a tareas con relaciones complejas entre variables.

Figura 3: Estructura de una DNN



Fuente: Lee et al. (2017)

En una DNN, cada capa cumple un rol distinto en la construcción del conocimiento del modelo. La primera capa suele aprender combinaciones simples de las variables de entrada. Las capas intermedias combinan esos patrones iniciales en estructuras más complejas, como por ejemplo “acciones con momentum alto pero valoración baja”. Las últimas capas capturan representaciones de alto nivel directamente útiles para la tarea, como clasificar activos según su probabilidad de buen rendimiento futuro. Este proceso

permite a la red aprender representaciones progresivamente más abstractas y jerárquicas de los datos.

Matemáticamente, cada neurona aplica una función lineal

$$z = w \cdot x + b$$

Donde:

- w son los pesos del modelo: valores que determinan cuánta importancia tiene cada entrada
- x vector de entrada con datos
- b es el sesgo o *bias,* que permite desplazar la activación independientemente de las entradas.

Tras calcular z, se aplica una función no lineal $f(z)$ que introduce la capacidad de modelar relaciones complejas entre variables. Entre las funciones más populares destacan:

- ReLu: (Rectified Linear Unit) usada en capas ocultas por su simplicidad y eficiencia, activa solo los valores positivos y anula los negativos:

$$ReLu(z) = max(0, z)$$

- Softmax: Usada en la capa de salida en problemas de clasificación multiclase, transforma un vector de puntuaciones (logits) en una distribución de probabilidades que suma 1. Cada componente representa la probabilidad de pertenecer a una clase

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{para } i = 1,2,\dots,K$$

2.8. Métricas de Evaluación

En este capítulo presentamos una serie de métricas habitualmente empleados en la literatura con el objetivo de evaluar la adecuación de los modelos.

1. **Retorno Acumulado de la Cartera**

Donde rt es el retorno de la cartera en el periodo t, y T es el número total de periodos. Esta métrica refleja el crecimiento total de la inversión bajo un esquema de reinversión continua, y es clave para evaluar el rendimiento efectivo de las decisiones de asignación de activos.

$$R_{acum.} = \prod_{t=1}^{T}(1 + rt) - 1 \qquad (5)$$

2. **Volatilidad de la Cartera:**

En este caso usamos la versión simplificada para series temporales, donde $\bar{r}$ es el retorno medio de la cartera. A pesar de la definición previa de riesgo, se considera la volatilidad como una métrica relevante para el inversor. La versión simplificada evalua la variabilidad de los retornos anuales.

$$\sigma = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(\, r_t - \bar{r}\,)^2} \qquad (6)$$

3. **Matriz de confusión**

La matriz de confusión es una herramienta fundamental para evaluar el rendimiento de modelos de clasificación. Proporciona información detallada sobre el tipo de errores cometidos

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \qquad (7)$$

En nuestro caso, dada una clase *overperformance* $c \in \{c_1 c_2, \ldots c_n\}$ derivaremos nuestra cartera sobre acciones clasificadas dentro de esta clase (la clase más alta de retorno). Cada métrica representa:

- **TP** (c): instancias correctamente clasificadas como clase c
- **FP** (c): instancias de otras clases clasificadas incorrectamente como clase c
- **FN** (c): instancias de clase c clasificadas como otra clase
- **TN** (c): todas las demás instancias correctamente no clasificadas como clase c

En este caso, se proporciona una matriz de grado 5 donde cada fila representa clases reales y cada columna clases predichas. Adicionalmente consideramos métricas derivadas sobre la matriz de confusión

4. Precision

Muestra la proporción de predicciones correctas respecto al total de predicciones para una clase. Es especialmente interesante para el caso que concierne este trabajo. Mediante la elaboración de una cartera no necesitamos necesariamente encontrar todas las empresas con un retorno alto, sino asegurarnos que de las empresas que clasificamos en la clase de retornos más alta, pocas sean erróneamente clasificadas, es decir minimizar el número de falsos positivos (FP)

$$Precision = \frac{TP_c}{TP_c + FP_c}$$

5. Recall

Muestra cuantos casos de la clase Overperformance son realmente clasificados como tal. Si bien resulta menos relevante ya que no necesitamos encontrar todos los casos, sigue siendo una métrica con cierto interés.

$$Recall = \frac{TP}{TP + FN}$$

6. MDD

La siguiente versión de MDD o *drawdown* máximo mide la mayor caída año a año, o lo que es equivalente, el peor retorno anual. Resulta especialmente interesante desarrollar estrategias que no tengan años excesivamente malos ya que esto daña seriamente la confianza sobre el modelo y rompen el poder del interés compuesto.

$$\text{MDD} = \min_{t \in \{1,\dots,T\}} (R_t)$$

2.9. Crítica al estado del Arte:

1- Volatilidad y Riesgo

Es importante puntualizar que Sharpe entendió la volatilidad como una medida directa del riesgo de una acción. Esta asunción es mayoritariamente consensuada en la literatura. Si bien carteras con menores varianzas son comúnmente deseadas por los inversores, el anterior aproximamiento cuenta con un sesgo a destacar:

Como se puede observar en la tabla a continuación, El Sharpe-Ratio penaliza toda volatilidad, también al alza. Haciendo que a pesar del mayor rendimiento de B, el portfolio A obtenga un mayor Sharpe-Ratio.

Tabla 1: Comparación de Sharpe-ratios

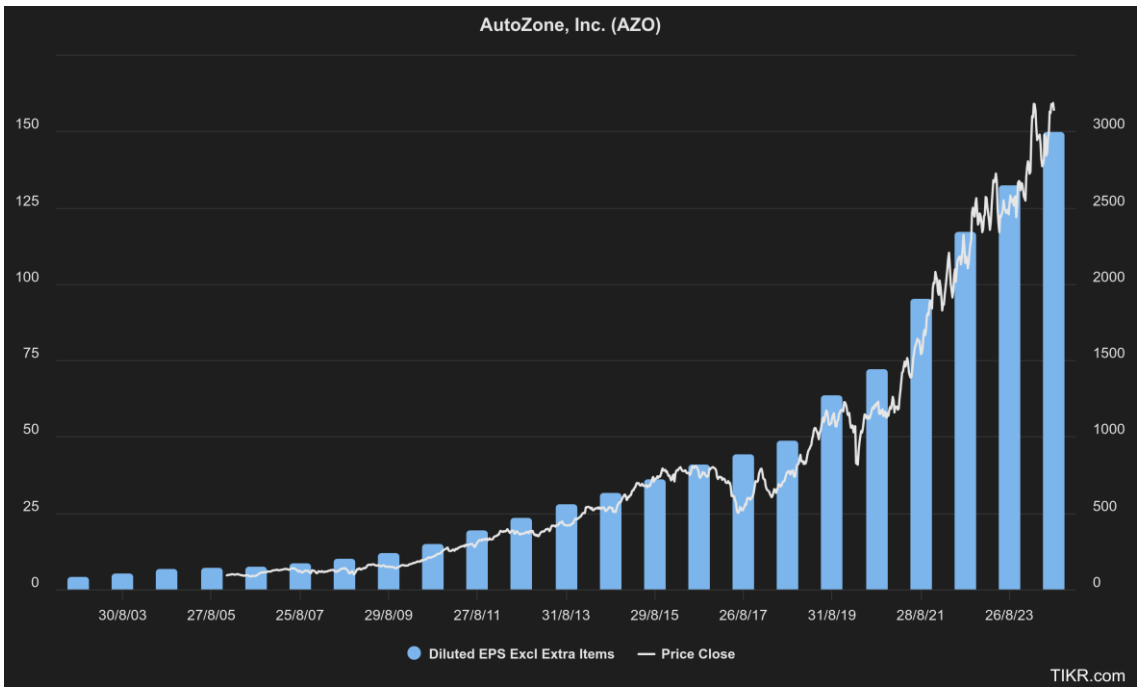
Portfolio	A	B
t=0	100	100
t=1	105	150
t=2	110	130
t=3	115	140
Rt	15%	40%
Std	6,45%	21,6%
Rf	4%	4%
Sharpe	1,704%	1,666%

Fuente: Elaboración propia

Otros autores, (Howard Marks, 2011) generalmente aquellos en campos relacionados con la inversión en valor y con metas de inversión amplias en términos temporales, sustentan la idea de que el riesgo debe ser entendido como la pérdida permanente de capital.

Además, dada la irracionalidad de los mercados puede ser común que empresas oscilen ampliamente a pesar de que su valor intrínseco no lo haga con la misma fuerza. Estos movimientos no deberían percibirse como riesgo para el inversor largoplacista. De hecho, este tipo de situaciones puede dar lugar a oportunidades. Por ejemplo, observemos el caso de Autozone, empresa líder distribuidora de piezas de automóvil en América. Podemos observar en la Figura 2, la evolución del precio y beneficio por acción. A pesar de mostrar más de 20 años de resultados sólidos, el precio ha variado considerablemente todos los años.

Figura 4: Evolución Histórica de Precio y Beneficio por acción de AutoZone Inc.



Fuente: TIKR

2- Trabajos Previos:

En su trabajo final de grado titulado “Estrategias de Creación de Carteras de Inversión Basadas en Ciencia de Datos” P. Llobregat aborda el mismo problema haciendo uso también del S&P500 como *benchmark*, también empleando métricas expuestas en este trabajo como fundamentales, Sin embargo, su dataset cuenta con histórico limitado evaluando las estrategias únicamente entre 2016 y 2024. Además, únicamente considera una selección estática de constituyentes. Para evitar el *look-ahead bias* trabaja solo con empresas activas en el índice antes del 2016. Sin embargo, esto ignora la realidad dinámica del S&P 500, donde la composición de empresas cambia anualmente, afectando significativamente a la viabilidad real de las estrategias.

Adicionalmente (Rodriguez et al. 2024), establece la variable objetivo en base a rangos fijos de estimación de retorno. Otro aproximamiento interesante podría ser establecer las clases en función de percentiles de retorno para cada ventana, de forma que estas estén balanceadas.

2.10. Propuesta

En el presente trabajo se opta por la reciente definición de riesgo al trabajar con un *dataset* histórico de inversión superior a 20 años. Por ello el objetivo del algoritmo es predecir retornos superiores, incluso si acabamos con carteras con mayor volatilidad. Sin embargo, a modo de análisis incluimos la volatilidad de las carteras creadas frente a la volatilidad del índice de referencia como una métrica de evaluación

Además a diferencia de previos estudios altamente centrados en predecir movimientos mensuales o incluso semanales, nuestro aproximamiento opta por ventanas anuales, donde los modelos pueden ser más fácilmente interpretables. En concreto, el objeto de clasificación será el rendimiento desde el punto de vista de mayores retornos de forma consistente al mercado.

Los datos son descargados a través de proveedores altamente reconocidos en la industria financiera como lo son Nasdaq Data Link y LSEG . Esto resulta en una calidad de datos superior que minimiza la cantidad de valores faltantes o datos anómalos, eliminando la necesidad de incurrir en otras prácticas popularmente empleadas por inversores *retail* como *scrapping*.

3. Selección de Variables

En la práctica, los inversores han empleado numerosas estrategias para identificar retornos superiores. En este capítulo presentamos algunas de las estrategias al alcance del inversor particular y que además cuentan con casos de éxito con retornos auditados. En segundo lugar, identificamos métricas clave sobre las que construir nuestras variables. Algunas de las estrategias coinciden con la revisión de factores previa, mientras que otras son añadidas para intentar mejorar el modelo.

3.1. Value

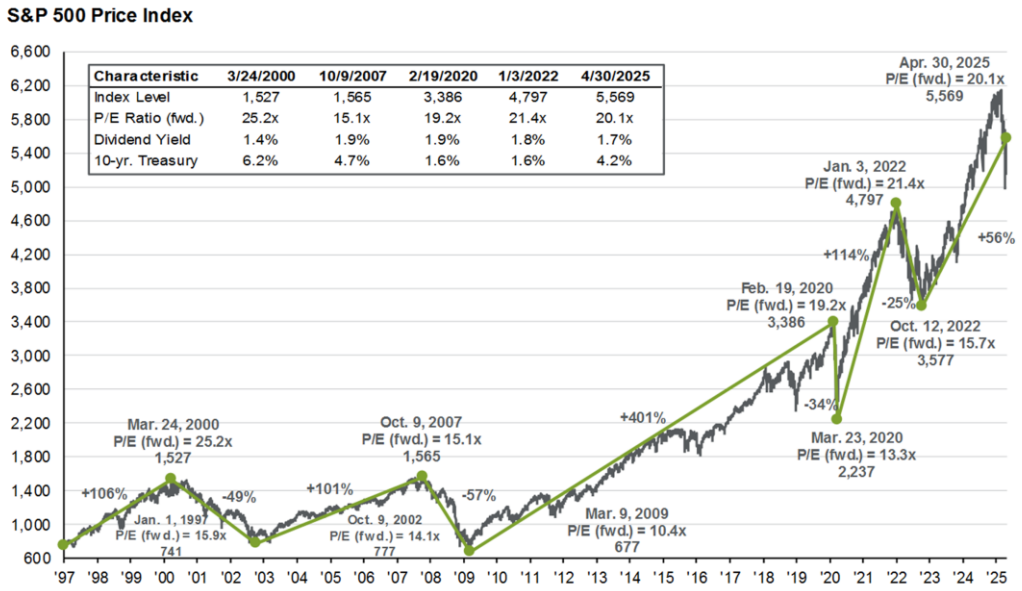
La inversión en valor o *value investing* parte de la premisa de que el mercado, por razones psicológicas o estructurales, ocasionalmente infravalora empresas en exceso. Benjamin Graham, Warren Buffett o Joel Greenblatt han construido imperios siguiendo estos principios. La evidencia empírica, especialmente la aportada por Fama y French con su factor HML (High Minus Low), demuestra que las acciones con múltiplos bajos (P/E, P/B) tienden a obtener primas de rentabilidad a largo plazo. Estas métricas buscan capturar compañías "baratas" en relación a sus fundamentales, apostando a que el mercado reconocerá ese valor a futuro.

Las principales métricas de valoración se denominan múltiplos o yields:

- Price to Book (P/B): $\frac{Precio\ por\ acción}{Valor\ en\ Libros}$
- Price to Earnings (P/E): Relación entre Precio y beneficio: $\frac{Precio\ por\ acción}{Beneficios\ por\ acción}$
- Dividend Yield: dividendo: $\frac{Dividendo\ por\ acción}{Precio\ por\ acción}$
- FCF Yield: $\frac{FCF\ por\ acción}{Precio\ por\ acción}$
- EV/EBITDA: $\frac{Capitalización\ de\ Mercado - Deuda\ neta}{EBITDA}$

Por ejemplo en la Figura 2 podemos observar la evolución del índice S&P500 y su valoración en términos de P/E. En los últimos 30 años, el P/E medio del índice ha sido de 16.93 veces. Como podemos observar los mínimos de cotización suelen coincidir con puntos bajos de valoración, mientras que en los picos, los partícipes están dispuestos a pagar hasta 25€ por cada euro de beneficio anual.

Figura 5: Evolución del S&P500 y su P/E



Fuente: JP Morgan "Guide to the Markets"

Sin embargo, los múltiplos de valoración cuentan con limitaciones a destacar. Por ejemplo, empresas poco intensivas en capital como consultoría, pueden presentar una alta relación entre su precio y sus activos, sin necesariamente indicar sobrevaloración.

De igual modo métricas como el P/E deben contrastarse con empresas comparables en crecimiento o midiendo su variación frente a periodos anteriores, reflejando cambios de atractividad de valoración.

3.2. Quality

La estrategia *quality* selecciona empresas con ventajas competitivas, balances sólidos y alta rentabilidad sobre el capital. Terry Smith o Thomas Rowe Price han popularizado este enfoque, argumentando que la calidad "compra tiempo" y protege en ciclos adversos. La literatura reciente (Fama & French 2015, Hou-Xue-Zhang, etc.) confirma que la rentabilidad operativa y el bajo apalancamiento son premiados por el mercado. Esto se debe a que empresas de calidad son capaces de subir precios sin perder clientes. Además, su solvencia les habilita adquirir competidores en momentos de dificultades financieras.

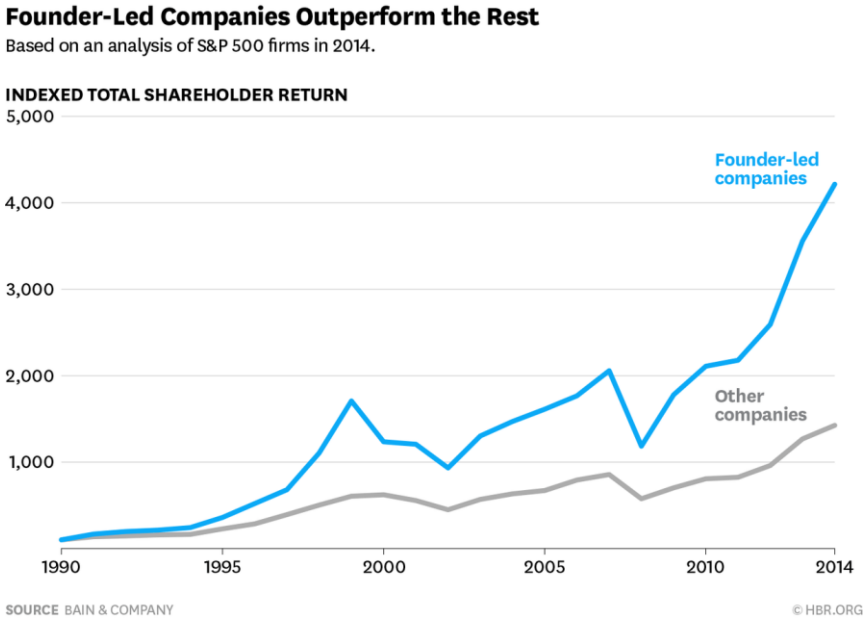
A largo plazo, estas características permiten compuestos de rentabilidad atractivas, incluso si no parecen baratas en el corto plazo.

La teoría financiera tradicional (Modigliani y Miller, 1958) sostiene que los beneficios de las empresas tienden a estabilizarse y converger hacia la media en el largo plazo. (habitualmente representado por el PIB). Sin embargo, en la práctica existen empresas excepcionales que logran sostener un crecimiento y rentabilidad superior gracias a ventajas competitivas duraderas. Pensemos en CocaCola o Apple. Dichas marcas están

tan integradas en los hábitos del consumidor, que competir contra ellas se hace extremadamente complejo.

Si bien es más abstracto encontrar datos al respecto. Otro hecho indicativo de calidad es la habilidad de la directiva y su capital humano. En concreto según un estudio de Bain & Co, las empresas lideradas por sus fundadores tienden a comportarse mejor que aquellas lideradas por un director externo. Esto puede estar relacionado con la visión y motivación que un fundador aporta en sus decisiones. Otras métricas indicativas incluyen nivel de compensación o el porcentaje de compensación variable (opciones o acciones) sobre el total.

Figura 6: Outperformance de empresas lideradas por fundadores



Fuente: Bain & Company

- ROE: $\frac{Beneficios}{Patrimonio\ Neto}$
- Margen EBIT/operativo: $\frac{EBIT}{Ventas}$
- Presencia del fundador
- Participación de la directiva $\frac{Acciones_{\ D}}{Total\ de\ Acciones}$
- Deuda/Equity: $\frac{Deuda}{Patrimonio\ Neto}$

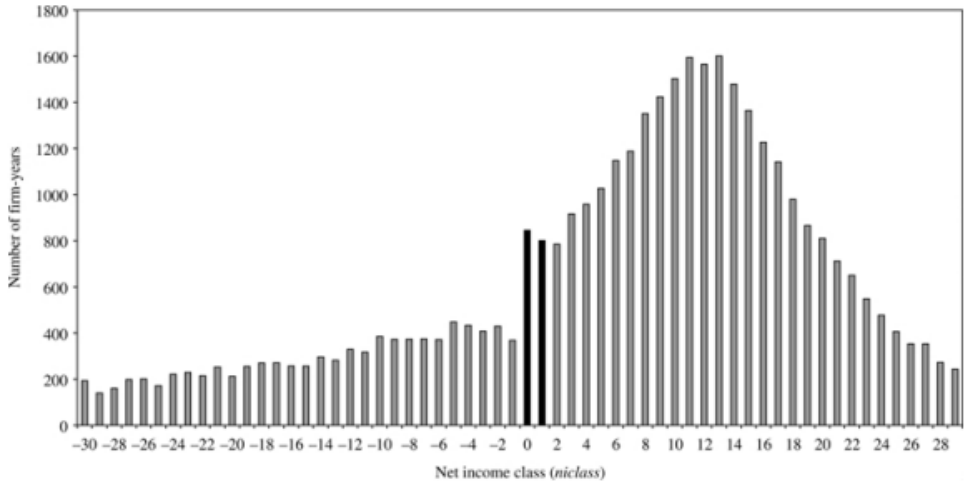
3.3. Growth

Empresas cuyas ventas, beneficios y otras métricas importantes creen rápidamente, suelen comportarse mejor en bolsa en periodos largos de tiempo. Cabe matizar que dichas empresas se suelen encontrar a una valoración superior y su compra no siempre está justificada. Sin embargo, popularizada por Warren Buffet y otros gestores, la combinación del factor *growth* con los dos factores previos, empresas de calidad y a un precio razonable es una de las temáticas más influyentes en fondos de inversión. En esencia las empresas que crecen suelen experimentar mayor volatilidad dado que su valor está determinado en mayor medida por los flujos más futuros, haciendo que cambios en las tasas de descuento o estimaciones de crecimiento tengan un fuerte impacto en los modelos desarrollados por los analistas.

Variabilidad de beneficios: $\frac{Beneficios_{t+1}}{Beneficios_{t}} - 1$

Si bien el crecimiento de beneficios es importante, las prácticas contables empleadas por cada constituyente del S&P500 son diversas y atienden a incentivos personales. La Figura 4 muestra la distribución del beneficio neto escalado por capitalización de mercado para empresas con más de 20 años listadas Se observa claramente una concentración anómala justo por encima de cero, es decir, muchas firmas reportan pequeñas ganancias positivas, y relativamente pocas reportan pequeñas pérdidas, lo que genera el famoso "*Kink*" en la distribución de beneficios. Este patrón ha sido documentado en múltiples estudios (por ejemplo, Burgstahler y Dichev, 1997) y ha motivado una reflexión crítica sobre la fiabilidad del beneficio contable como métrica de desempeño. Esto conduce a la necesidad de analizar flujos de caja junto con beneficios.

Figura 7: Distribución de Beneficios



Fuente: Burgstahler, D., & Dichev, I. (1997)

Contrastamos pues el crecimiento en beneficios con el crecimiento en flujo de caja.

$$\text{Variabilidad de flujos de caja:} \quad \frac{FCF_{t+1}}{FCF_t} - 1$$

Además, interesa que ambas métricas sean representadas por acción, para reflejar los efectos de dilución que el accionista sufre.

3.4. Momentum

Otra estrategia, defendida por inversores como Richard Driehaus, consiste en comprar lo que ha subido recientemente y vender lo que ha caído, bajo la premisa de que las tendencias suelen persistir. El factor momentum, formalizado por Carhart (1997), es una de las anomalías más robustas internacionalmente y se usa ampliamente en hedge funds cuantitativos.

Uno de los inversores más aclamados, Stanley Druckenmiller, sostiene que las grandes tendencias en el mercado suelen durar entre 12 y 24 meses, por lo que identificar movimientos fuertes en una etapa temprana y mantener la posición mientras persista el impulso puede generar retornos excepcionales. Su enfoque se basa en detectar cambios macroeconómicos, tecnológicos o de política monetaria que desencadenan ciclos

sostenidos, como ocurrió con las acciones tecnológicas entre 2020 y 2021 impulsadas por la digitalización y la expansión monetaria.

Este marco respalda el uso de estrategias de momentum basadas en ventanas de retorno de 6 a 12 meses, ya que capturan tendencias que se pueden prolongar durante el año siguiente, justo en el tramo temporal de mayor euforia previo a que la tendencia se revierta o se lateralice.

$$\text{Retornos últimos 6 meses} \quad \frac{Precio_{\,t-1/2}}{Precio_{\,t}}$$

$$\text{Retornos últimos 12 meses} \quad \frac{Precio_{\,t-1}}{Precio_{\,t}}$$

3.5. Estrategias técnicas

Otros inversores han empleado reglas técnicas simples como señales de compra/venta, como el cruce de medias móviles o la superación de máximos previos ("breakout"). Estas estrategias buscan aprovechar tendencias o revertir movimientos extremos, asumiendo que los precios reflejan información y psicología colectiva.

$$\text{Cruce de SMA*:} \quad \text{SMA } 1y \; - \; \text{SMA } 2y$$

Utiliza dos medias móviles (SMA) de distinta duración para generar señales de compra y venta en función de su interacción. Por ejemplo SMA(y=1) y SMA(y=2). Donde 'y' indica la duración de la media en años. Como se observa en la Figura 5, cuando la SMA corta cruza hacia arriba la SMA larga, se genera una señal alcista (*long*) Cuando la SMA corta cruza hacia abajo la SMA larga, se genera una señal bajista (*short*)

Figura 8: Crossover de SMA

Si bien se trata de un indicador simple y objetivo, cuenta con ciertas limitaciones. Como retrasos en las señales. Al ser un indicador *lagging*, puede reaccionar tarde ante cambios bruscos de tendencia. Además, puede ocasionar falsas señales en rangos laterales: en mercados sin tendencia clara, puede generar múltiples entradas y salidas sin éxito conocidos como *whipsaws*. Para mitigarlo proponemos usar medias móviles amplias como son 1 y 2 años en ventanas anuales, dado que al ser más largos, estos SMAs filtran mejor el ruido del corto plazo.

- Distancia de máximos/mínimos

  Usado en trading cuantitativo, la diferencia entre el precio actual y el máximo o el mínimo del año, muestra que tanta distancia hay entre la cotización actual y la mínima o máxima del año anterior. En estudios como Jegadeesh & Titman (1993) y más recientemente en el *52-week high effect* (George & Hwang, 2004), se observa que las acciones que están cerca de su máximo de 52 semanas tienden a seguir subiendo a corto plazo

$$\text{Distancia a máximos: } \frac{Precio\ MAX_{t-1}}{Precio_t}$$

$$\text{Distancia a mínimos: } \frac{Precio\ MIN_{t-1}}{Precio_t}$$

3.6. Esfuerzos inversores

Una métrica común para medir el la agresividad en inversión por parte de las empresas es el Capex* .En concreto su magnitud frente a ventas. Esta métrica refleja el factor CMA

mencionado anteriormente, donde las empresas conservadoras son menos glamorosas, pero en el tiempo tienden a superar a las agresivas en rendimiento, ajustando por riesgo.

$$\frac{Capex}{Ventas}$$

3.7. Tamaño

Otro factor mencionado previamente era **SMB**. Este argumenta que las empresas de baja capitalización tienen mayor potencial de crecimiento que las de gran capitalización, ofreciendo mejores resultados. En nuestro caso, todas las empresas son realmente de gran capitalización al pertenecer al S&P500. Sin embargo, se considera relevante extraer información de la capitalización de mercado para observar si existe diferencia alguna en rendimiento. De hecho, en la última década la concentración del índice ha augmentado rápidamente, haciendo que un selecto grupo de empresas tecnológicas sean las protagonistas. La Figura 6 muestra la evolución en la contribución del top 5 empresas a la capitalización total del índice, resaltando tendencias dinámicas de preferencia de tamaño

Exhibit 4: The concentration of market cap in the largest stocks has soared
as of April 23, 2020

Source: Compustat, Goldman Sachs Global Investment Research

3.8. Otros Indicadores

El entorno macro y la liquidez afectan el atractivo relativo de la renta variable. Estrategias exitosas suelen incorporar variables como la tasa libre de riesgo, spreads de crédito o medidas de volatilidad.

- Tasa libre de riesgo (*risk-free rate*)

  Se ha tomado como tasa libre de riesgo, el rendimiento a vencimiento (YTM) del bono del Tesoro de Estados Unidos a 10 años, considerado el principal referente mundial por su bajo riesgo de impago. La tasa libre de riesgo es fundamental para valorar inversiones, ya que representa la alternativa de inversión sin riesgo y sirve como mínimo exigible de rentabilidad para los inversores. Cuando esta tasa aumenta, el coste de oportunidad de invertir en acciones también sube, haciendo más difícil justificar inversiones en renta variable y favoreciendo la salida de capital hacia activos seguros. Además, una tasa libre de riesgo elevada suele traducirse en mayores costes de financiación para las empresas, lo que puede afectar negativamente su capacidad de crecimiento e inversión, y en consecuencia influir en la valoración de sus acciones. Volatilidad pasada

- Volumen_{t-1}

  El volumen negociado durante el año pasado se ha incorporado como métrica relevante porque representa el nivel de liquidez y el interés de los inversores en cada valor. Un mayor volumen suele indicar una mayor facilidad para comprar o vender acciones sin provocar grandes variaciones en el precio, lo que reduce el riesgo de iliquidez para los inversores.

- Volatilidad_{t-1}

  La volatilidad del último año representada como la desviación estándar (σ_{t-1}) al desearse generalmente empresas que experimenten menos variaciones de precio.

4. Selección del Proveedor

Como hemos podido observar en el capítulo previo, una gran cantidad de variables proviene de información referente a las actividades operativas, de inversión y financiación de la empresa. Esta información se hace pública mediante la publicación de los estados financieros de forma trimestral y anual. En Estados Unidos el organismo encargado de regular la correcta y efectiva publicación de resultados de empresas cotizadas es la SEC (Securities and Exchange Comission). Esta información es accesible mediante su buscador EDGAR. Sin embargo, el entrenamiento de modelos requiere de datos estructurados comúnmente en formato de tabla ancha. Una posible solución sería realizar *scrapping*, si bien esto aumenta el nivel de detalle y personalización sobre la información escogida, complicaría exponencialmente nuestro objetivo final y materia de este trabajo, la elaboración de modelos.

Otra posible alternativa es el uso de proveedores que ofrecen directamente datos sobre estados financieros. Un proveedor común, es Nasdaq-Data-Link (NDL) conocido previamente como Quandl. A través de su plataforma NDL comercializa bases de datos de terceros como Sharadar, una firma independiente fundada en 2013 especializada en extracción, estandarización y agrupación de información financiera en base a presentaciones de empresa. Una de las publicaciones mencionados previamente (Rodríguez et al., 2024) emplea precisamente este proveedor.

En concreto cuentan con un producto denominado “Sharadar Core US Equities Bundle” con un historial desde 1998 hasta 2025, cubriendo más de 16.000 empresas en Estados Unidos. El producto tiene un coste de $69/mes y consiste a su vez de la agrupación de 5 sub-productos:

- Core US Fundamentals data
- Core US Insiders data (solo disponible desde 2008)
- Core US Institutional Investors
- Sharadar Equity Prices.
- Sharadar Fund Prices

El presente trabajo se centra en Core US Fundamentals Data para la obtención de datos fundamentales y Equity Prices para el cálculo de indicadores técnicos, rendimiento, volumen y momentum. Desgraciadamente Core US Insiders cuenta solo con datos desde 2008. En este trabajo se ha preferido aprovechar un amplio dataset (1998-2025) a incluir el factor Insider, por lo que este se deja para futuras ampliaciones o mejoras. El resto de productos no son relevantes para el trabajo.

Core US Fundamentals Data:

Se compone de 7 tablas que agrupan información en función de su carácter. La más importante es SF1 que cuenta con 150 indicadores fundamentales. Además otra tabla denominada SP500, provee información sobre las empresas activas, y deslistadas en el índice, permitiendo así la reconstrucción de un *dataset* con empresas solo activas en el índice en cada momento.

Equity Prices

Cuenta con una tabla SEP con datos diarios de precio EOD*, esto resulta interesante ya que si bien trabajaremos con ventanas anuales, la granularidad diaria permite capturar patrones temporales más finos, abriendo la puerta a métricas más sofisticadas para alimentar el modelo de predicción, como la volatilidad anualizada, momentum o el *drawdown* máximo.

Además, Equity prices ofrece precios ajustados por *splits*, *spin-offs* y dividendos. Esto es fundamental para el análisis de desempeño relativo entre acciones, ya que elimina distorsiones causadas por eventos corporativos. Por ejemplo, una acción que experimenta un *split* puede parecer haber perdido valor si no se ajusta correctamente el precio. De igual forma, los dividendos reducen el precio de mercado, pero no representan una pérdida real para el inversor

LSEG Data Platform

Adicionalmente, durante la elaboración de este trabajo, el alumno realizó una estancia en un programa de intercambio, y se interesó por una base de datos denominada LSEG Data Analystics, ofrecida a través de la universidad de destino. LSEG es junto a Bloomberg una de las plataformas usadas por profesionales de la inversión cuantitativa. Sin embargo, el acceso ofrecido limitaba la posibilidad de extracción a consultas de tipo REST ligeras, con una servicio a parte para descargas *bulk*, por lo que su aplicación a la descarga de grandes cantidades de datos se veía limitada. Es por ello que se decidió hacer un uso marginal de esta, completando los indicadores de NDL con datos macroeconómicos comunes a todas las empresas para cada año como es la tasa libre de riesgo.

Figura 9: Proveedores empleados

Fuente: nasdaq.com, lseg.com

5. Análisis del Problema

5.1. Definición formal del Problema

El problema a consiste en la creación de algoritmos, capaces de clasificar correctamente acciones según su rendimiento a un año. Esto permitirá la construcción de una cartera con rendimientos superiores a los del mercado. Dicha cartera es construida mediante la selección de ciertos constituyentes del S&500, a principio de cada ventana anual con una alta probabilidad de obtener rendimientos superiores al mercado.

Para ello primero debemos definir que rendimientos son superiores a nuestro benchmark, el S&P500. En la sección de Contexto se comentó que el rendimiento del SP&500 desde 1928 a 2024 fue del 6,7%. Al añadir la rentabilidad por dividendos, el rendimiento se incrementa a 9,6%. Rendimientos por encima de esa cifra son considerados superiores al mercado y serán de forma general el objeto de identificación. Adicionalmente es importante establecer las clases de forma balanceada.

Si bien gran cantidad de estudios en ML usan ventanas de baja duración. Dada la naturaleza de este trabajo y su importante relación con el análisis financiero de las empresas, se considera que en ventanas de mayor duración, métricas derivadas del análisis financiero pueden tener más éxito. En esencia buscamos predecir retornos de acciones a 1 año. Para ello se parte de un universo de empresas que históricamente han formado parte del índice bursátil S&P500.

Se trabaja por tanto con múltiples empresas (500) en diversos años, lo que hace de este problema una tarea con datos panel o longitudinales. Así pues combinan variación temporal y variación entre muestras.

5.2. Requisitos Funcionales y no Funcionales

Requisitos Funcionales

El sistema propuesto debe ser capaz de realizar las siguientes funciones para permitir el desarrollo y evaluación de estrategias de inversión basadas en aprendizaje automático:

- **Ingestión y preprocesamiento de datos**: Importar datos históricos de precios y fundamentales de activos financieros, gestionar valores faltantes, y asegurar la calidad de los datos.
- **Ingeniería de características**: Calcular métricas relevantes como medias móviles, volatilidades históricas, ratios financieros y otros indicadores técnicos y fundamentales.
- **Entrenamiento de modelos predictivos**: Implementar y entrenar modelos de regresión y clasificación, tales como Ridge Regression, Random Forest y XGBoost, para predecir retornos o categorías de rentabilidad futura.
- **Validación temporal**: Emplear técnicas de validación específicas para series temporales, como el esquema walk-forward o TimeSeriesSplit, con el fin de evitar el uso indebido de datos futuros (*look-ahead bias*).
- **Evaluación de modelos**: Generar métricas de desempeño como el error cuadrático medio (RMSE), Sharpe Ratio, Sortino Ratio, matriz de confusión y clasificación multiclase.
- **Selección de activos**: Construir carteras de inversión seleccionando los activos con mejores predicciones de rentabilidad, según un umbral definido (por ejemplo, top 40% o 50%).

- **Visualización de resultados**: Producir gráficos de rendimiento acumulado, evolución de las métricas a lo largo del tiempo, y visualizaciones de clasificación como matrices de confusión y gráficos de dispersión Predicted vs Actual.

Requisitos No Funcionales

Además de las funcionalidades descritas, el sistema debe cumplir con los siguientes requisitos de calidad:

- **Eficiencia computacional**: Los tiempos de entrenamiento y evaluación deben ser razonables para *datasets* medianos, permitiendo iteraciones rápidas y exploración de diferentes configuraciones.
- **Integridad de los datos**: Debe garantizarse la consistencia y corrección de los datos durante todo el proceso, detectando y tratando adecuadamente valores extremos, inconsistencias y datos faltantes.
- **Modularidad**: El sistema debe estar estructurado de forma modular para facilitar la reutilización y la extensión de componentes individuales, como los módulos de ingestión de datos, preprocesamiento, modelado y evaluación.
- **Escalabilidad**: El sistema debe ser capaz de adaptarse a un incremento en la cantidad de activos analizados, en el número de años de datos históricos o en el número de características utilizadas.
- **Reproducibilidad**: Se debe garantizar que, bajo condiciones idénticas de datos y parámetros, los resultados obtenidos sean reproducibles mediante la fijación de semillas aleatorias y el control del entorno de ejecución.
- **Robustez frente a errores**: El sistema debe ser capaz de manejar errores inesperados, como fallos de lectura de datos o cálculos numéricos, proporcionando mensajes claros y mecanismos de recuperación cuando sea posible.

5.3. Análisis del Marco Legal y Ético

El desarrollo de estrategias de inversión basadas en técnicas de aprendizaje automático involucra consideraciones legales y éticas que deben ser tenidas en cuenta para garantizar el cumplimiento normativo y la responsabilidad profesional.

Marco Legal

- **Protección de Datos**: Los datos utilizados en este proyecto son de naturaleza financiera y no contienen información personal identificable (PII). En algoritmos que operen con otro tipo de datos, es esencial asegurar que el tratamiento de los datos cumpla con los principios de protección de datos y privacidad establecidos, por ejemplo, por el Reglamento General de Protección de Datos (GDPR) en Europa. Además, los servicios contratados mediante NDL y LSEG permiten la construcción y desarrollo de herramientas en base a los datos que ofrecen, pero no la reventa de datos a terceros.

- **Regulación Financiera**: En el ámbito financiero, los algoritmos que impactan decisiones de inversión pueden estar sujetos a regulación bajo directivas como 2014/65/EU - MiFID II (Markets in Financial Instruments Directive) en la Unión Europea. Únicamente entidades autorizadas por la CNMV* en España pueden ofrecer recomendaciones de inversión. Aunque este proyecto se realiza con fines académicos y no implica recomendaciones reales de inversión, es importante considerar que, en un entorno profesional, las estrategias basadas en aprendizaje automático deben ser auditables, explicables y cumplir con los requisitos regulatorios de transparencia y gestión de riesgos.

Marco Ético

Además de los aspectos legales, el uso de machine learning en estrategias de inversión plantea importantes cuestiones éticas:

- **Transparencia y Explicabilidad**: Es fundamental que los modelos empleados sean interpretables hasta cierto nivel, para evitar decisiones de tipo caja negra o *black-box*. La trazabilidad de las decisiones de inversión es especialmente relevante en el cumplimiento de MiFID II, que exige que los clientes puedan comprender los riesgos de las recomendaciones recibidas.

- **Evitar Overfitting y Optimización Excesiva**: Éticamente, se debe evitar la construcción de modelos sobreajustados a datos históricos, conocido como *data snooping bias* que puedan presentar expectativas de rentabilidad no realistas.

- **Responsabilidad sobre Decisiones Automatizadas**: En caso de implementación real, sería necesario establecer claramente la responsabilidad en las decisiones tomadas por sistemas automáticos, conforme a las normativas de protección al inversor.

5.4. Análisis de Riesgos

Los principales riesgos en la construcción de modelos:

Máximo Drawdown

En el presente trabajo se proponen estrategias de inversión basadas en acciones sin apalancamiento, es decir, sin recurrir a financiación externa o endeudamiento. Este enfoque implica que el inversor no puede perder más que el capital inicialmente aportado. En consecuencia, la pérdida máxima posible en el peor de los escenarios sería la pérdida total del capital invertido.

Look-ahead bias

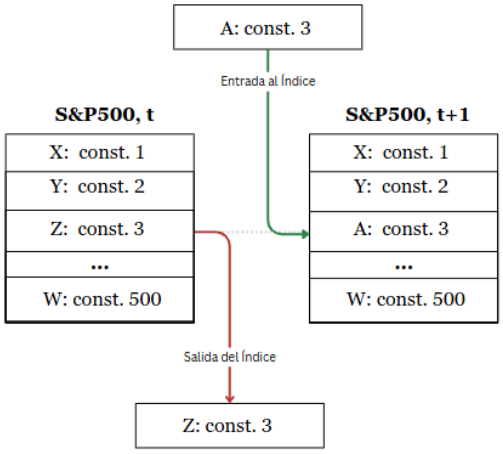
El *look-ahead bias* ocurre cuando el modelo utiliza información que no estaría disponible en el momento real de la predicción. Esto genera una visión artificialmente optimista del desempeño futuro, ya que el modelo, inadvertidamente, aprende de datos futuros.

Para mitigar este riesgo es crucial respetar el orden temporal en el entrenamiento y validación, así como emplear técnicas de validación específicas para series temporales, como walk-forward validation o TimeSeriesSplit.

(INSERTAR IMAGEN)

Se utilizan constituyentes dinámicos de los índices de mercado, es decir, listas de activos que reflejan la composición vigente en cada momento histórico, evitando usar información de constituyentes futuros no conocidos en ese instante.

Figura 10: Constituyentes dinámicos



Fuente: Elaboración propia

- Se debe asegurar ventanas de entrenamiento que solo contienen información pasada hasta el momento de cada predicción. Por ejemplo, con los datos fundamentales, se asegura que únicamente se utilicen estados financieros publicados antes de la fecha de referencia. No se utilizan cifras revisadas *o re-statements* posteriores que serían desconocidos en tiempo real.

Overfitting

El overfitting ocurre cuando el modelo se ajusta excesivamente a los datos de entrenamiento, capturando ruido en lugar de patrones generalizables. Esto es especialmente crítico en mercados financieros, donde la relación señal-ruido es baja y los patrones históricos pueden no repetirse en el futuro. Mediante un *backtest* pretendemos validar las estrategias sobre datos no observados por el modelo.

Medidas adoptadas para mitigar el riesgo de overfitting:

- Regularización: Se emplean técnicas como la penalización L2 en Ridge Regression o la regularización incorporada en XGBoost (L1/L2).

- Separación de conjuntos de datos: El conjunto de datos se divide en tres bloques:
  - Train set: Para entrenar el modelo.
  - Validation set: Para ajustar los hiperparámetros mediante técnicas de búsqueda como grid search o random search.

    - Test set: Reservado exclusivamente para la evaluación final, asegurando que los resultados son verdaderamente fuera de muestra.

- Control de la complejidad del modelo: Se limitan parámetros como la profundidad máxima de los árboles o el número de estimadores en métodos ensemble para evitar modelos demasiado flexibles.

- Evaluación de desempeño en datos no vistos: Se reportan métricas de desempeño (Sharpe Ratio, F1-Score, matriz de confusión) utilizando exclusivamente el Test set para garantizar la validez de los resultados.

5.5. Identificación de Posibles Soluciones

El desarrollo de carteras basadas modelos de ML presenta múltiples grados de libertad que deben ser definidos cuidadosamente. Estas decisiones abarcan desde la elección del modelo y la forma de validar, hasta cómo se preprocesan los datos o se definen las clases objetivo. A continuación, se describen los principales aspectos analizados:

1. Modelos de clasificación

   Existen numerosos enfoques para abordar un problema de clasificación multiclase en datos financieros en panel. Entre los modelos más frecuentes en la literatura se encuentran:

   - Modelos lineales como regresión logística o LDA.
   - Árboles de decisión y métodos ensemble como Random Forest y XGBoost.
   - Modelos más recientes como redes neuronales profundas (DNNs) y otros basados en aprendizaje profundo.
   - LightGBM, SVM

Cada uno ofrece distintos compromisos entre interpretabilidad, capacidad de modelado no lineal y rendimiento computacional.

2. Estrategia de validación

Al trabajar con datos temporales, se debe respetar la estructura cronológica para evitar *data leakage* (introducir datos futuros al modelo). Algunas estrategias son

- División aleatoria: desestimada por romper la estructura temporal.
- Validación tipo *k-fold* adaptada temporalmente: útil para algunos estudios, pero compleja de implementar de forma estable.
- Validación por ventana deslizante (walk-forward): seleccionada por su coherencia con el uso real del modelo. Se entrena con los últimos *n* años y se evalúa sobre el siguiente.

3. Definición de las clases objetivo

Una decisión fundamental es cómo transformar un retorno continuo en una clase discreta. Dos aproximaciones evaluadas son:

- Clasificación por quintiles o percentiles, como se hace habitualmente en finanzas cuantitativas.
- Clasificación por rangos de retorno absolutos, como se ve en estudios centrados en interpretabilidad financiera (ej. retorno ≥ 15% = clase 4).

4. Horizonte temporal de entrenamiento y datos

Se consideraron diferentes longitudes para la ventana de entrenamiento, validación y test. Una ventana más larga puede incluir más datos pero perder relevancia, mientras que una más corta puede capturar mejor el comportamiento reciente. Además la selección de diferentes periodos históricos puede resultar en modelos diversos.

5. Tratamiento de valores faltantes

Se consideran varias alternativas de tratamiendo de registros con valores nulos: -

- Imputación con medias.
- Eliminación de columnas o filas
- Forward fill, que se utilizó por su coherencia con datos financieros y por no introducir información futura.

6. Restricciones de cartera

El tamaño de la cartera seleccionada afecta directamente a la diversificación. Además existen múltiples estrategias según la teoría de portfolios como liquidez, apalancamiento, distribución de pesos, carteras *long-only* o *long-short* que incluyen posiciones bajistas.

7. Criterio de normalización de datos

5.6. Solución Propuesta

La solución propuesta consiste en construir un marco comparativo de modelos de clasificación supervisada para predecir el rendimiento futuro relativo de activos financieros y generar carteras a partir de las predicciones.

Modelos seleccionados para comparación:

1. **Modelo base**: regresión logística multiclase, útil como línea base interpretable.
2. **XGBoost**: modelo potente y bien establecido para datos estructurados, capaz de modelar relaciones complejas sin requerir normalización.
3. **Red neuronal profunda** (DNN): modelo con alta capacidad expresiva, adecuado para problemas no lineales y tareas de ranking.

Estos tres modelos representan diferentes niveles de complejidad y aproximaciones metodológicas. Todos se entrenan con la misma estructura de validación temporal y conjunto de datos, lo que permite una comparación justa y controlada.

Aunque el desarrollo podría realizarse en R, MATLAB u otros entornos, se optó por Python, por ser el lenguaje más extendido en ciencia de datos y finanzas cuantitativas, y por su ecosistema rico en librerías (pandas, scikit-learn, keras). Se usaron Jupyter Notebooks por su facilidad para prototipar y documentar.

Algunos modelos requieren de normalización, Dado que los modelos de redes neuronales son sensibles a la escala de los datos, se aplicó una normalización estándar (StandardScaler) entrenada solo con el conjunto de entrenamiento cada año para evitar fugas de información.

Este enfoque permite **analizar empíricamente cuál de los modelos produce mejores resultados financieros**, evaluando no solo la capacidad de clasificación, sino también su utilidad práctica al construir carteras. La solución se enmarca así no solo como un desarrollo técnico, sino como una contribución metodológica al uso de machine learning en asset allocation

Dado el planteamiento del problema en la sección “Definición del problema”, se evaluaron distintas **alternativas de modelado y estructura de validación** temporal. A continuación se presentan las opciones consideradas, junto con un análisis de sus ventajas, desventajas y criterios de selección.

Modelos: Una gran variedad de problemas similares hacen uso de arboles de decisión y más crecientemente redes neuronales, otros modelos?

Tamño de Ventana: depende el horizonte de inversión, los modelos pueden trabajar con datos de diversa granularidad

Lenguaje de Programación: R, Python, net

Split de Validación: Han de respetar la naturaleza de series temporales. Walk forward, ventana deslizante? (el que he implementado)

Entorno: Jupyter notbooks, Google collab, server especializado

Tratamiendo de datos: como se imputan los nan, ffill? Media? Eliminación?

Tamaño y restricciones de Cartera: numero de acciones que escogemos.

Selección de Clases: percentiles vs rangos fijos

Solución Propuesta

5.7. Plan de Trabajo

El desarrollo del presente TFG se ha estructurado en las siguientes fases principales, inspiradas en la metodología de proyectos software y adaptadas a la naturaleza multidisciplinar de este trabajo

Estimación de Esfuerzos:

Fase	**Descripción breve**	**Horas estimadas**	**Horas reales**
1. Revisión bibliográfica y definición	Investigación de estrategias, revisión de literatura y planteamiento del problema.	40 h	50 h
2. Obtención y procesamiento de datos	Búsqueda, contratación y extracción de fuentes de datos; limpieza y preprocesado.	60 h	100 h
3. Ingeniería de características	Selección, transformación y generación de variables predictoras.	50 h	60 h
4. Diseño y entrenamiento de modelos ML	Implementación de modelos, ajuste de hiperparámetros y validación.	80 h	60 h
5. Backtesting y análisis de resultados	Simulación de carteras, análisis comparativo y visualización.	50 h	- h
6. Documentación, memoria y conclusiones	Redacción de memoria, elaboración de anexos y materiales gráficos.	60 h	- h
7. Reuniones de seguimiento y tutorías	Coordinación con tutores, revisión de avances y correcciones.	10 h	- h
Total		**320 h**	**¿? h**

Nota: Las horas reales se han ido ajustando conforme avanzaba el desarrollo, redistribuyendo el esfuerzo en función de los retos encontrados en cada fase.

Presupuesto

En base a nuestras estimaciones de esfuerzos iniciales, el desarrollo completo del software habría requerido un total aproximado de 350 horas realizadas por un estudiante de doble grado en Ingeniería Informática y ADE aún no titulado.
El salario bruto de un recién titulado en este perfil suele situarse en torno a 21.000 €/año, lo que implica un coste empresa de 27.468 € anuales (según estimaciones estándar), es decir, 13,20 €/hora (calculado sobre 2.080 h/año).

Considerando un descuento del 30% por tratarse de un alumno no titulado:

- Coste hora estimado alumno: 13,20 € × 0,7 = 9,24 €/h
- Coste total de desarrollo: 350 h × 9,24 € = 3.234 €

A ello hay que sumar los siguientes costes asociados:

- Acceso a base de datos US Core Bundle: 60 €/mes × 3 meses = 180 €
- Acceso a base de datos LSEG Data Platform: 2000€/mes x 2 meses = 4000€
- Asesoramiento senior/tutorías: 2 tutores × 10 h × 38,65 €/h = 773 €

Total presupuesto estimado: **8.127 €**

5.8. ¿?

6. Diseño de la Solución

6.1. Arquitectura

Como se establece en el capítulo de Metodología, en el presente trabajo se opta por un *pipeline* clásico de ML. Mediante este el objetivo es transformar una serie de datos brutos (incluyen cierto preprocesado ya) en conocimientos accionables, en concreto en la forma de modelos de decisión. La naturaleza modular del pipeline permite el desarrollo por etapas, las cuales pueden ser iterativas sin necesidad de rehacer todo el progreso. Gracias a esta arquitectura cualquier resultado es reproducible y su división facilita la trazabilidad. Dado que se busca obtener modelos lo más robustos posibles, se ha incurrido en numerosas iteraciones de las etapas presentadas a continuación. algunos de estos ajustes menos relevantes se han dejado fuera de la extensión de la memoria con el propósito de ofrecer un seguimiento comprensivo al lector.

(INTRODUCIR DIAGRAMA)

Ingesta de Datos:

El primer paso consiste en la extracción de información relevante para la creación de variables escogidas en el capítulo Selección de Variables. Se ha trabajado con

Genaración de API keys		
Descarga diaria de datos técnicos		
Descarga anual de datos fundamentales		
Descarga datos macro		

6.2. Variables

En el apartado de selección de variables se ha entrado en detalle sobre las métricas seleccionadas. A continuación se proporciona una breve recopilación de estas:

1. Identificatorios
   ticker, date

2. Precio
   closeadj, max_1y, min_1y, sma_1y, sma_2y, sma_diff, dist_max_1y, dist_min_1y

3. Value:

   pb, pe, evebitda, ps, pe_yoy, pb_yoy

4. Rentabilidad
   fcf_yield, roe, ebit_margin, net_margin

5. Size
   marketcap, log_marketcap, variables dummy de marketcap

6. Esfuerzos de Inversión
   Capex / ventas

7. Técnico
   sma_diff, dist_max_1y, dist_min_1y, vol_over_sma, vol_1y, volatility_1y,

8. Fundamental
   revenue, ebitda, netinc, eps, ebit, capex, ncfo, equity, revenue_yoy, ebitda_yoy, eps_yoy, equity_yoy, capex_yoy, fcfps, fcfps_yoy, ncfo_yoy, log_revenue, log_ebitda, log_ebit, log_equity

9. Solvencia
   de

10. Momentum
    ret_6m, ret_12m

11. Macroeconómico
    risk_free_rate

12. Target

target_12m_final, sharpe_1y

6.3. Patrón de validación walk-forward / series temporales

6.4. Tecnologías y herramientas

Entorno Visual

Lenguaje

Principales librerías

La base de Datos escogida es Nasdaq Data Link. Se trata de un proveedor reconocido con más de 800.000 usuarios activos. Entre otras las cualidades que la hacen destacar son:

- **Amplia gama de datos:** Nasdaq Data Link ofrece datos de mercado, datos de empresas (informes financieros, datos de accionistas), datos económicos y más.
- **Datos de alta calidad:** La información proporcionada por Nasdaq Data Link es generalmente precisa y confiable.
- **Herramientas de acceso y análisis:** Nasdaq Data Link proporciona herramientas y interfaces para facilitar la consulta, el análisis y la visualización de los datos.
- **Datos Históricos y en Tiempo real:**
- **Amplio alcance:** Nasdaq Data Link cubre el mercado de valores de EE. UU. y también ofrece datos de otros mercados financieros internacionales

6.5. ¿?

6.6. ¿?

7. Desarrollo de la Solución Propuesta

7.1. Pipeline de ingestión (LSEG dl, nan)

7.2. Ingeniería de características y lag de 3 meses

Normalización

En modelos basados en árboles, como XGBoost, la normalización o estandarización de las variables no es necesaria y, de hecho, puede perjudicar el rendimiento. A diferencia de los modelos lineales o basados en distancias, donde la escala de las variables afecta directamente el ajuste o la medida de similitud, los árboles toman decisiones basadas en umbrales de corte que dependen únicamente del orden relativo de los datos. Normalizar los datos altera la distribución natural de las variables, puede amplificar outliers y dificultar la identificación de puntos de corte óptimos, afectando negativamente la capacidad predictiva del modelo. Por ello, en regresión con XGBoost, es recomendable

trabajar con las variables en su escala original, dejando que el modelo gestione las diferentes magnitudes internamente sin necesidad de transformación previa.

7.3. Modelos implementados

7.4. Metodología global

7.5. Código clave y estructura de repositorio(resto en anexos)

8. Implementación

8.1. Deploy Local / cloud

8.2. Scripts de Automatización y reproducibilidad

9. Pruebas

9.1. Métricas predictivas

9.2. Backtesting: retorno, drawdown, etc

9.3. Análsis de robustez

9.4. Interpretabilidad

10. Conclusiones

10.1. Síntesis de Resultados Técnicos y Financieros

10.2. Implicaciones para la eficiencia de Mercado

10.3. Limitaciones del Estudio

[0] https://www.msci.com/www/blog-posts/sizing-up-the-global-market/05073690405

[1] https://www.investopedia.com/terms/a/assetclasses.asp

[2] https://finance.yahoo.com/news/global-real-estate-valued-379-180013773.html

[3] https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html

[4] MITCHELL, Tom M. (1997). *Machine Learning*. New York: McGraw-Hill, 414 p. ISBN 0-07-042807-7.

[5] https://www.morningstar.com.au/personal-finance/the-greatest-investor-youve-never-heard-of

[6] Zuckerman, G. (2019). *The Man Who Solved the Market*.

Gu, S., Kelly, B., & Xiu, D. (2020). *Empirical Asset Pricing via Machine Learning*. Review of Financial Studies, 33(5), 2223–2273.

Fama, E. F., & French, K. R. (1993). *Common risk factors in the returns on stocks and bonds*. Journal of Financial Economics, 33(1), 3-56.

Jegadeesh, N., & Titman, S. (1993). *Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency*. The Journal of Finance, 48(1), 65-91.

Carhart, M. M. (1997). *On persistence in mutual fund performance*. The Journal of Finance, 52(1), 57-82.

Fama, E. F., & French, K. R. (2015). *A five-factor asset pricing model*. Journal of Financial Economics, 116(1), 1-22.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). *...and the cross-section of expected returns*. Review of Financial Studies, 29(1), 5-68.

Burgstahler, D., & Dichev, I. (1997). Earnings management to avoid earnings decreases and losses. *Journal of Accounting and Economics*, 24(1), 99–126. https://link.springer.com/article/10.1023/A:1024481916719

https://www.profilesw.com/insights/ai-machine-learning-in-portfolio-management/

Imágenes:

https://markets.businessinsider.com/news/stocks/sp500-concentration-large-cap-bad-sign-future-returns-effect-market-2020-4-1029133505

https://www.researchgate.net/publication/356698772_Ensemble_learning_for_the_early_prediction_of_neonatal_jaundice_with_genetic_features

https://www.researchgate.net/publication/327651247_Prediction_of_Arctic_Sea_Ice_Concentration_of_Kara-Barents_Seas_Using_RCM_Data_with_Machine_Learning

Anexo

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza.**	**x**			
ODS 2. **Hambre cero.**			**x**	
ODS 3. **Salud y bienestar.**			**x**	
ODS 4. **Educación de calidad.**	**x**			
ODS 5. **Igualdad de género.**				**x**
ODS 6. **Agua limpia y saneamiento.**				**x**
ODS 7. **Energía asequible y no contaminante.**				**x**
ODS 8. **Trabajo decente y crecimiento económico.**	**x**			
ODS 9. **Industria, innovación e infraestructuras.**		**x**		
ODS 10. **Reducción de las desigualdades**.		**x**		
ODS 11. **Ciudades y comunidades sostenibles**.				**x**
ODS 12. **Producción y consumo responsables.**				**x**
ODS 13. **Acción por el clima.**				**x**

ODS 14. **Vida submarina.**				**x**
ODS 15. **Vida de ecosistemas terrestres.**				**x**
ODS 16. **Paz, justicia e instituciones sólidas.**			**x**	
ODS 17. **Alianzas para lograr objetivos.**				**x**