**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE**

**Departamento de Matemáticas y Física**

**Sustentabilidad y Tecnología**

**PROYECTO DE APLICACIÓN PROFESIONAL (PAP)**

**Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio II**


# 4J05 Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio

# Stock Picking con Machine Learning

**PRESENTAN**

Programas educativos y Estudiantes

Ing. Financiero. Juan Carlos Gutiérrez Valdivia.

Ing. Financiero. Carlos Daniel Ponce Anguiano.

Profesor PAP: Sean Nicolás González Vázquez

Tlaquepaque, Jalisco, mayo de 2024

**ÍNDICE**

## Contenido

# REPORTE PAP

## Presentación Institucional de los Proyectos de Aplicación Profesional

*Los Proyectos de Aplicación Profesional (PAP) son una modalidad educativa del ITESO en la que el estudiante aplica sus saberes y competencias socio-profesionales para el desarrollo de un proyecto que plantea soluciones a problemas de entornos reales. Su espíritu está dirigido para que el estudiante ejerza su profesión mediante una perspectiva ética y socialmente responsable.*

*A través de las actividades realizadas en el PAP, se acreditan el servicio social y la opción terminal. Así, en este reporte se documentan las actividades que tuvieron lugar durante el desarrollo del proyecto, sus incidencias en el entorno, y las reflexiones y aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.*

## Resumen

En el proyecto trabajado en el PAP denominado " Stock Picking con Machine Learning ", nos hemos dedicado a automatizar la selección y evaluación de activos financieros para asistir a los inversionistas en el proceso de toma de decisiones de inversión. Este proyecto se ha centrado en el Value Investing para el desarrollo de herramientas analíticas.

Nos propusimos y logramos establecer los fundamentos y recolectar datos para nuestro análisis. Identificamos los activos y fundamentales significantes, diseñamos y desarrollamos un algoritmo para extraer datos de las empresas seleccionadas, y estructuramos esta información de manera óptima para su uso posterior.

En la fase de análisis fundamental y evaluación de modelos, evaluamos el rendimiento histórico de las acciones, las clasificamos y probamos diversos modelos de clasificación hasta seleccionar el más eficiente. La validación y backtesting fueron esenciales para asegurar la solidez y confiabilidad de nuestra estrategia de inversión.

A través de nuestra metodología, llegamos a resultados prometedores. El producto final, un modelo desarrollado con técnicas de Machine Learning como Gradient Boosting, ha probado ser efectivo en la predicción de rendimientos financieros positivos. Las simulaciones de backtesting han mostrado métricas favorables en términos de rendimiento, lo que subraya la capacidad del algoritmo para generar ganancias.

Nuestro enfoque ha permitido abordar de manera integral las complejidades del entorno financiero y tiene implicaciones significativas para la toma de decisiones estratégicas en el ámbito de las inversiones.

# 1. Introducción

## 1.1. Objetivos

Los objetivos del proyecto se enfocan en mejorar la toma de decisiones en la inversión de acciones mediante el uso de técnicas avanzadas de modelado matemático y aprendizaje automático. Estos incluyen:

1. Automatizar la selección y evaluación de activos financieros para asistir a los inversionistas en el proceso de toma de decisiones de inversión.
2. Establecer los fundamentos y recolectar datos para el análisis de activos financieros significativos.
3. Diseñar y desarrollar un algoritmo para extraer y estructurar datos de empresas seleccionadas.
4. Evaluar el rendimiento histórico de las acciones y seleccionar modelos de clasificación eficientes para la inversión.
5. Implementar la validación y backtesting para asegurar la solidez y confiabilidad de la estrategia de inversión.

Para lograr la implementación exitosa de este proyecto, se establecerá un flujo de trabajo riguroso y estructurado que abarcará desde la recopilación y procesamiento de datos hasta la generación y evaluación de recomendaciones de inversión. El proceso comenzará con la obtención de datos financieros relevantes, que posteriormente serán sometidos a un proceso de limpieza y transformación para asegurar su calidad y consistencia. A

continuación, se desarrollarán modelos matemáticos avanzados que analizarán de manera integral los factores clave que influyen en la valoración de los activos financieros. Estos modelos permitirán generar recomendaciones basadas en evidencia objetiva y establecerán criterios cuantitativos
para la toma de decisiones. Finalmente, se implementará una interfaz de usuario intuitiva que presentará de manera clara y concisa las recomendaciones generadas. Este enfoque de flujo de trabajo garantiza la coherencia y la precisión en todas las etapas del proceso, brindando confiabilidad y eficacia a la aplicación desarrollada.

## 1.2. Justificación

Una problemática central identificada es la prevalencia de sesgos conductuales entre los inversionistas, que incluyen desde el exceso de confianza hasta la aversión a la pérdida, afectando negativamente su juicio y conduciendo a decisiones de inversión no óptimas.

Además, se ha observado que los inversionistas minoristas a menudo se enfrentan a barreras significativas al intentar acceder a las inversiones bursátiles a través de instituciones bancarias. Las altas comisiones y los requisitos de montos mínimos de inversión excluyen a un amplio segmento de la población que, a pesar de tener la voluntad de invertir, se encuentra marginado de estos servicios financieros.

La colaboración entre el equipo PAP y PILOU ha sido fundamental para la solución de estas problemáticas. A través de un proceso de diagnóstico, hemos podido identificar las necesidades y desafíos específicos que enfrentan los inversionistas, lo que nos ha permitido diseñar soluciones dirigidas que abordan estas cuestiones de manera efectiva.

## 1.3 Antecedentes

Los antecedentes históricos nos muestran una evolución desde la especulación impulsiva hasta la inversión basada en análisis detallado. La práctica de la inversión ha sido

transformada por la introducción de teorías como la de los mercados eficientes y el modelado de riesgos, lo que ha redefinido el panorama bursátil. En la actualidad, nuestro algoritmo representa un esfuerzo continuado por mejorar las prácticas de inversión y respaldar las decisiones financieras con una metodología robusta y cuantitativa.

### 1.4. Contexto

El proceso de toma de decisiones en inversiones bursátiles está marcado por una rica historia de avances teóricos y prácticos. Históricamente, los inversionistas han buscado estrategias para optimizar la colocación de capital, pero no sin enfrentarse a sesgos cognitivos y emocionales que pueden distorsionar la racionalidad de sus elecciones. El reconocimiento de estos sesgos ha dado lugar a la búsqueda de métodos cuantitativos que apoyen una toma de decisiones más objetiva y fundamentada. En este contexto, nuestro programa se presenta como una herramienta cuantitativa y confiable, diseñada para mitigar la influencia de sesgos y mejorar la eficacia del análisis de inversión.

Desde una perspectiva teórica, este programa se alinea con la lógica de la gestión de inversiones, donde la marginación de la intuición subjetiva a favor de modelos cuantitativos y algoritmos de Value Investing se ha convertido en una tendencia creciente. Estos avances se sustentan en principios de economía conductual y análisis financiero, proporcionando a los inversionistas medios para evaluar activos con mayor precisión y realizar colocaciones de capital que apunten a una rentabilidad sostenida.

## 2. Desarrollo

### 2.1. Sustento teórico y metodológico

El Value Investing constituye la piedra angular de la metodología empleada en el proyecto. Este enfoque de inversión, propuesto por Benjamin Graham y posteriormente perfeccionado por su discípulo Warren Buffett, se basa en la identificación y adquisición de

acciones que se negocian a precios inferiores a su valor intrínseco (Graham & Dodd, 1934). La premisa es que el mercado a menudo reacciona de manera exagerada a eventos buenos y malos, lo que afecta temporalmente el valor verdadero de una empresa.

**Teoría Fundamental de Value Investing**

La teoría fundamental detrás del Value Investing es que comprar acciones de compañías robustas y bien gestionadas a precios significativamente bajos respecto a su valor intrínseco ofrece un margen de seguridad sustancial y es probable que produzca un rendimiento superior a largo plazo (Buffett, 1984). Este enfoque se concentra en el análisis fundamental, evaluando estados financieros detalladamente para discernir la verdadera salud financiera y las perspectivas de crecimiento de una empresa, así como la solidez de su liderazgo y su posición competitiva dentro de la industria.

**Integración con Tecnologías Avanzadas**

En este proyecto, el Value Investing se moderniza y enriquece mediante la integración de tecnologías de aprendizaje automático. Utilizando modelos avanzados como Gradient Boosting y Random Forest, el proyecto no solo evalúa criterios financieros fundamentales, como el ratio precio/beneficio (P/E), el ratio precio/valor contable (P/B) y otros indicadores clave de desempeño, sino que también analiza patrones en los datos históricos para predecir futuros rendimientos de las acciones (Hastie, Tibshirani, & Friedman, 2009). Esta metodología permite un escrutinio más profundo y una evaluación cuantitativa que supera los métodos tradicionales de análisis fundamental por sí solos.

**Metodología de Datos y Backtesting**

El enfoque metodológico también abarca la recolección y limpieza exhaustiva de datos, esencial para el entrenamiento efectivo de los modelos de machine learning. Las técnicas de imputación y manejo de datos faltantes, como K-Nearest Neighbors (KNN), aseguran la integridad y calidad de los datasets utilizados para el entrenamiento y la validación de modelos (James, Witten, Hastie, & Tibshirani, 2013).

El backtesting se utiliza para validar la eficacia de los modelos desarrollados, simulando su desempeño en escenarios históricos para evaluar su robustez y capacidad de generalización antes de su implementación en tiempo real. Esta práctica es ampliamente reconocida y valorada en el ámbito financiero por proporcionar una perspectiva realista sobre cómo podrían haber funcionado las estrategias de inversión en diferentes condiciones de mercado (Lo & MacKinlay, 1999).

**Conclusión**

En conclusión, el sustento tecnológico y metodológico de este proyecto representa una fusión innovadora de principios tradicionales de inversión y técnicas avanzadas de análisis y modelado. La aplicación de Value Investing, enriquecido con análisis cuantitativo y tecnología de aprendizaje automático, promete una estrategia de inversión más informada, precisa y rentable. Esta metodología no solo respeta los principios fundamentales del Value Investing, sino que también adopta modernas prácticas de análisis de datos para crear un sistema de inversión robusto y adaptativo que está bien equipado para enfrentar los desafíos del mercado financiero actual.

## 2.2. Planeación y seguimiento del proyecto

### *Descripción del Proyecto y Flujo de Trabajo*

En seguida se encuentra el desglose del flujo de trabajo, abordando brevemente las tareas que se llevan a cabo en cada fase y ahondando en la descripción del proyecto.

**Establecimiento de Fundamentos y Recolección de Datos**

Iniciaremos nuestro proyecto identificando los activos y fundamentales clave para nuestro análisis. A continuación, diseñaremos y desarrollaremos un algoritmo eficiente para extraer estos datos fundamentales de las compañías seleccionadas. Una vez recopilados, organizaremos la información de manera estructurada para facilitar su uso en la siguiente etapa de modelado.

**Análisis Fundamental y Evaluación de Modelos**

Procederemos a evaluar el rendimiento histórico de las acciones elegidas, clasificándolas en términos de rendimientos positivos o negativos (0 o 1, respectivamente). Estos datos serán divididos en conjuntos de entrenamiento y prueba para permitir una validación rigurosa. Posteriormente, implementaremos y compararemos varios modelos de clasificación para determinar cuál ofrece la mejor capacidad predictiva y seleccionaremos el más adecuado para nuestro propósito.

**Validación y Backtesting**

La fase final incluirá un análisis de la importancia y significación de cada variable dentro del modelo elegido. Realizaremos simulaciones de backtesting en múltiples escenarios para verificar la eficacia del modelo seleccionado, asegurándonos de que nuestra estrategia de inversión sea sólida y confiable.

A continuación se muestra un diagrama que muestra de manera visual el flujo de trabajo que se seguirá para la alcanzar los objetivos propuestos:


*Figura 1. Flujo de Trabajo*

## Descarga de datos de API Alphavantage

En la primera etapa de desarrollo, se diseñó un código destinado a la descarga de datos desde la API Alphavantage. El límite de solicitudes permitido por la API KEY es 100 solicitudes por día y 5 solicitudes por minuto por usuario. Para optimizar este proceso de obtención de datos, cada miembro del equipo procedió a descargar archivos de Excel que contenían 100 conjuntos de datos relacionados con el balance general y el estado de resultados. El propósito de esta acción fue preparar los datos para su posterior consolidación en dos DataFrame, uno por cada estado financiero.

De esta manera, el código toma una lista de símbolos de acciones como entrada y utiliza la clave de API de Alphavantage para realizar solicitudes web al proveedor de información y obtener los datos financieros. Luego, organiza estos datos en un DataFrame, donde cada columna representa una partida contable y cada fila corresponde a un símbolo de acción y un período de tiempo. Finalmente, guarda estos datos en un archivo de Excel con el nombre especificado en la variable *"excel_name".* El código también incluye la opción de manejar errores y proporciona un seguimiento del progreso durante la descarga de datos.

## Unión de documentos .xlsx

Después de completar la fase inicial de descarga de datos, se procedió a consolidar toda la información descargada en múltiples archivos de Excel en uno solo. Este nuevo documento contiene la totalidad de los datos recopilados, y se considera esencial para simplificar y agilizar el proceso de análisis. Una vez lograda esta unificación, se importó este archivo bajo el esquema de DataFrame utilizando la librería pandas. Este tipo de dato consiste en una estructura tabular extremadamente versátil que permite llevar a cabo una variedad de operaciones y análisis de datos de manera eficiente. De esta manera, se logró abstraer la

información financiera relevante y concentrarla en un solo lugar bajo un formato estandarizado y fácilmente manejable.

## Cálculo de Ratios Financieros

Habiendo realizado los procesos anteriormente descritos, el siguiente paso consistió en calcular algunos fundamentales (también conocidas como razones financieras). Cabe mencionar que en este proceso se rellenaron los valores faltantes con cero; si bien esto pudo causar algunos errores en el cálculo, esto se corrigió más adelante. En seguida se encuentran descritas las métricas financieras seleccionadas:

Price Earnings Ratio (PER): Se considera uno de los más importantes, pues indica cuántas veces se está pagando el beneficio neto anual de una empresa al adquirir una accion de la misma. En otras palabras, indica qué tan cara o barata es una acción, y se recomienda nunca comprar una acción cara o sobrevaluada. De esta manera, si el PER de una empresa es mayor a 25, no se recomienda comprarla, por más buena que sea.

$$\text{PER} = \frac{\text{Precio de acción} \times \text{Número de acciones}}{\text{Ingresos netos}}$$

*Figura 1. Ecuación Price Earnings Ratio (PER)*

Price Book Value: Es un indicador que compara el precio actual de una acción con su valor contable. Así, proporciona una medida relativa de cuánto paga el mercado por cada dólar de valor en libros de una empresa y se utiliza comúnmente para evaluar si una acción se encuentra sobrevalorada o subvalorada en relación con su valor intrínseco. Un P/B ratio bajo puede indicar que una acción está infravalorada, mientras que un P/B ratio alto puede sugerir que está sobrevalorada, aunque su interpretación precisa depende del contexto de la industria y otras consideraciones.

$$\text{PBV} = \frac{\text{Número de acciones} \times \text{Precio de acción}}{\text{Activo total}}$$

*Figura 2. Ecuación Price Book Value*

Prueba del ácido: Es una métrica financiera utilizada para evaluar la capacidad de una empresa para cumplir con sus obligaciones financieras inmediatas; mide la liquidez de una empresa a corto plazo. Una prueba del ácido saludable es generalmente igual o superior a 1, lo que sugiere que la empresa tiene suficientes activos líquidos para cubrir sus obligaciones en un plazo menor a un año. Por otro lado, un valor inferior a 1 indica una posible falta de liquidez y puede ser una señal de riesgo financiero.

$$\text{Acidtest} = \frac{\text{Activo Circulante} - \text{Inventarios}}{\text{Pasivo Circulante}}$$

*Figura 3. Ecuación Prueba del ácido*

Rotación de los activos: Mide la eficiencia con la que una empresa utiliza sus activos para generar ingresos. Expresado de otra forma, otorga una visión crítica de la capacidad de una empresa para generar ganancias con los recursos que tiene a su disposición, lo que ayuda a los inversionistas y analistas a evaluar la gestión de activos y la eficiencia operativa de la empresa. Una alta rotación de activos suele ser indicativa de una gestión eficiente, mientras que una baja rotación puede señalar problemas en la utilización de activos o la necesidad de una inversión adicional para aumentar la productividad.

$$\text{ATR} = \frac{\text{Ventas etas}}{\text{Activo total}}$$

*Figura 4. Ecuación Rotación de Activos (ATR)*

Ciclo de conversión del efectivo: Es un indicador clave que mide el tiempo que transcurre desde que una empresa invierte recursos en la adquisición de materias primas y otros insumos hasta que recibe ingresos en efectivo por la venta de sus productos o servicios.

Este ciclo comprende tres etapas fundamentales: la gestión de inventarios, el período de cuentas por cobrar y el plazo de pago a proveedores. Un ciclo de conversión del efectivo eficiente implica que la empresa puede recuperar rápidamente su inversión, optimizando sus flujos de efectivo y mejorando su liquidez, lo que es esencial para mantener una operación financiera saludable y sostenible en el largo plazo.

$$\text{CCC} = \frac{365 \times \text{ Inventarios}}{\text{Costo de venta}} + \frac{365 \times \text{ Cuentas por cobrar}}{\text{Ventas}} - \frac{365 \times \text{ Cuentas por pagar}}{\text{Costo de venta}}$$

*Figura 5. Ecuación Ciclo de Conversión de Efectivo (CCC)*

Return on Assets: Cuantifica la eficiencia en el uso de los activos de una empresa mediante la expresión, en porcentaje, de ingresos netos obtenidos con relación a los activos totales durante el ejercicio. Básicamente busca encontrar una relación entre el beneficio que obtiene la empresa a raíz de los recursos que tiene a su disposición. Entre más alto sea el indicador, mayor la eficiencia de la empresa en utilizar sus medios para generar ganancias.

$$\text{ROA} = \frac{\text{Ingresos Netos}}{\text{Activo Total}}$$

*Figura 6. Ecuación Return on Assets (ROA)*

Razón Deuda a Capital: Determina la proporción de financiamiento de una empresa mediante deuda en comparación con su financiamiento a través de capital propio. Esta métrica se toma como referencia para evaluar la solidez financiera y la capacidad de endeudamiento de una empresa. Mientras que un cociente alto puede indicar altos niveles de endeudamiento y riesgo, un cociente bajo puede apuntar a una estabilidad financiera, pero también a la falta de apalancamiento para el crecimiento de esta.

$$\text{DER} = \frac{\text{Pasivo Total}}{\text{Capital Total}}$$

*Figura 7. Ecuación Razon Deuda a Capital (DER)*

Margen de Utilidad Neto: Revela la rentabilidad y eficiencia operativa de una empresa al expresar el porcentaje de ganancias netas que obtiene en relación con sus ingresos totales. Un margen más alto indica que la empresa retiene una mayor parte de sus ingresos después de cubrir sus costos, gastos y el pago de impuestos. Un margen de utilidad neto más bajo, por otro lado, sugiere una menor rentabilidad en relación con sus ingresos.

$$\text{NPM} = \frac{\text{Utilidad Neta}}{\text{Ventas}}$$

*Figura 8. Ecuación Margen de Utilidad Neto (NPM)*

Multiplicador de capital: Se refiere a la relación entre el capital propio de una empresa y su capacidad para generar ganancias o rendimientos a través de inversiones o activos. Un multiplicador de capital más alto indica que la empresa puede obtener mayores rendimientos con su capital propio, lo que generalmente se considera una señal positiva de su rentabilidad y eficacia en la gestión financiera. Por otro lado, un multiplicador de capital bajo puede indicar que la empresa depende en gran medida de financiamiento externo, lo que puede aumentar su riesgo financiero.

$$\text{EM} = \frac{\text{Activo Total}}{\text{Capital Total}}$$

*Figura 9. Ecuación Multiplicador de Capital (EM)*

## Limpieza de DataFrame

La limpieza de datos es un paso fundamental en el análisis de datos financieros, ya que la calidad de los datos influye significativamente en la validez de los resultados y conclusiones que se puedan extraer. En el presente estudio, se llevó a cabo un proceso de limpieza de datos del DataFrame que contiene diversos ratios financieras, como Price-to-Earnings Ratio (PER), Price-to-Book Value Ratio (PBV), Return on Assets (ROA), entre otros, correspondientes a distintas empresas. Es así que este proceso se inició con una exploración preliminar de la base de datos mediante la aplicación de un método similar a

"*pd.describe()".* Esta técnica permitió obtener una visión general de las características de las columnas que componen el DataFrame, incluyendo información relevante sobre el tipo de objeto contenido en cada columna, la presencia de valores faltantes y el rango numérico de los datos.

Uno de los hallazgos significativos derivados de esta exploración inicial fue la identificación de valores representados como *"np.inf"* y *"-np.inf"* en varias columnas. Estos surgen como producto de las divisiones entre cero que se realizaron durante el cálculo de algunos ratios financieros (en el momento en que se sustituyeran valores faltantes por cero). En consecuencia, se determinó que era necesario abordar esta problemática realizando una sustitución por valores más apropiados evitando la distorsión que la eliminación de estos registros pudiera causar en cualquier análisis subsiguiente; en este caso, se optó por reemplazar los infinitos por "*np.nan"* (Not-a-Number), lo que permitiría tratarlos como valores faltantes en el proceso posterior de tratamiento de valores faltantes.

Una vez realizado lo anterior, se procedió a realizar una imputación utilizando el algoritmo K-Nearest Neighbors (KNN) con un valor de n igual a 5. El algoritmo KNN es una técnica de imputación que se basa en la idea de que los valores faltantes en un conjunto de datos pueden ser estimados de manera precisa tomando en cuenta los valores de los vecinos más cercanos en función de una métrica de distancia. En este caso, los vecinos más cercanos se determinaron en función de las características financieras similares de las empresas.
La utilidad y los beneficios de utilizar el algoritmo KNN en la imputación de datos radican en varios aspectos:

- Preservación de la estructura de los datos: KNN imputa valores faltantes basándose en las similitudes entre observaciones, lo que ayuda a preservar la estructura de los datos y a no introducir sesgos significativos.

- Flexibilidad: KNN es un método no paramétrico, lo que significa que no hace suposiciones específicas sobre la distribución de los datos, lo que lo hace adecuado para una amplia gama de situaciones.
- Robustez: KNN es robusto ante la presencia de valores atípicos y datos ruidosos, ya que se basa en la proximidad de los vecinos más cercanos en lugar de depender de estimaciones de parámetros.
- Adaptabilidad: El valor de n en KNN se puede ajustar según las necesidades del análisis. Un valor más alto de n suavizará la imputación, mientras que un valor más bajo podría ser más sensible a las diferencias sutiles entre observaciones.

Es mediante este proceso de imputación de datos faltantes que se termina con la fase de limpieza de datos.

## EDA (Exploratory Data Analysis)

Posterior a la limpieza de datos, se considera pertinente realizar un análisis exploratorio con el objetivo de obtener información adicional sobre la distribución de estos. Es entonces que a partir de un análisis descriptivo realizado sobre la base de datos utilizando una herramienta de generación de informes automatizada, se proporcionó una visión general detallada de las características y relaciones presentes en la misma. Aunque más adelante se encuentran algunos de los resultados más relevantes del informe, la totalidad de este documento se muestra como anexo al proyecto. Además, se realizaron análisis adicionales con el objetivo de recabar la mayor cantidad de información posible. Dentro de este marco, se destaca la capacidad de las herramientas para identificar y resaltar aspectos de interés que requieren una atención especial. A continuación, se muestran distintos insights obtenidos como producto de esta fase.

## *Características de la base de datos*

Dataset statistics

| | |
|---|---|
| Number of variables | 11 |
| Number of observations | 7760 |
| Missing cells | 0 |
| Missing cells (%) | 0.0% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 727.5 KiB |
| Average record size in memory | 96.0 B |

Variable types

| | |
|---|---|
| Text | 1 |
| DateTime | 1 |
| Numeric | 9 |

*Figura 10. Características de la base de datos*

Dentro de la base de datos en consideración, se albergan un conjunto compuesto por un total de once variables de interés. Estas variables encapsulan información relevante relacionada con los activos financieros, reconocidos en la industria con el término técnico "ticker". Además, se incluye la fecha de cierre del período fiscal en análisis y una serie de indicadores financieros cruciales: Price-to-Earnings Ratio (PER), Price-to-Book Value Ratio (PBV), Acid Test, Asset Turnover Ratio (ATR), Cash Conversion Cycle (CCC), Return on Assets (ROA), Debt-to-Equity Ratio (DER), Net Profit Margin (NPM) y el Earnings Multiple (EM).

Es importante destacar que esta base de datos engloba un total de 7,760 registros. Estos registros se distribuyen a lo largo de 20 instancias temporales, representando cinco años de observaciones divididos en trimestres. Asimismo, la base de datos abarca información relativa a 388 acciones pertenecientes al índice bursátil S&P500. La complejidad y amplitud de esta base de datos proporciona un rico conjunto de datos que facilita un análisis exhaustivo y profundo de las tendencias financieras y de desempeño de un amplio espectro de acciones dentro del mercado financiero.

## Visualización de base de datos

| | Stock | fiscalDateEnding | PER | PBV | Acid_test | ATR | CCC | ROA | DER | NPM | EM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | A | 2023-09-30 | 295.454866 | 3.072177 | 53.781818 | 0.154379 | 519.738042 | 0.010398 | 0.920655 | 0.067354 | 1.920655 |
| 1 | A | 2023-06-30 | 117.462748 | 3.287041 | 37.066474 | 0.156505 | 589.975104 | 0.027984 | 0.866805 | 0.178804 | 1.866805 |
| 2 | A | 2023-03-31 | 115.920786 | 3.736983 | 12.466387 | 0.159538 | 570.189255 | 0.032237 | 0.946693 | 0.202067 | 1.946693 |
| 3 | A | 2022-12-31 | 119.540607 | 4.176884 | 76.111111 | 0.174041 | 496.822863 | 0.034941 | 0.985297 | 0.200764 | 1.985297 |
| 4 | A | 2022-09-30 | 108.621824 | 3.408678 | 14.888889 | 0.163201 | 498.707191 | 0.031381 | 1.059320 | 0.192285 | 2.059320 |
| 5 | A | 2022-06-30 | 128.506518 | 3.367842 | 15.462857 | 0.151698 | 497.027433 | 0.026208 | 1.041195 | 0.172762 | 2.041195 |
| 6 | A | 2022-03-31 | 138.870374 | 3.805589 | 5.919964 | 0.160066 | 459.087301 | 0.027404 | 1.003686 | 0.171204 | 2.003686 |
| 7 | A | 2021-12-31 | 107.844614 | 4.452809 | 9.967901 | 0.154507 | 457.619360 | 0.041289 | 0.986454 | 0.267231 | 1.986454 |
| 8 | A | 2021-09-30 | 178.528000 | 4.492555 | 21.646154 | 0.149175 | 461.585143 | 0.025164 | 1.121108 | 0.168690 | 2.121108 |
| 9 | A | 2021-06-30 | 204.468603 | 4.247472 | 13.282927 | 0.144739 | 450.431800 | 0.020773 | 1.161746 | 0.143522 | 2.161746 |

*Figura 11. Visualización de base de datos*

En el DataFrame que se exhibe en la figura superior, se presentan las diez primeras filas de la base de datos. En esta representación tabular, la primera columna corresponde a los identificadores ticker asociados a los instrumentos financieros en consideración. En la segunda columna, se dispone la fecha de cierre del período correspondiente. A continuación, se encuentran dispuestos los nueve indicadores financieros, también conocidos como ratios financieros, que han sido calculados a partir de la información disponible para cada uno de los trimestres y para cada instrumento financiero en estudio. Estos ratios proporcionan una visión cuantitativa y relacional de diversos aspectos clave de la salud financiera de las empresas, permitiendo un análisis detallado de su desempeño y solidez en el ámbito económico y financiero.

## Análisis de correlación Pearson

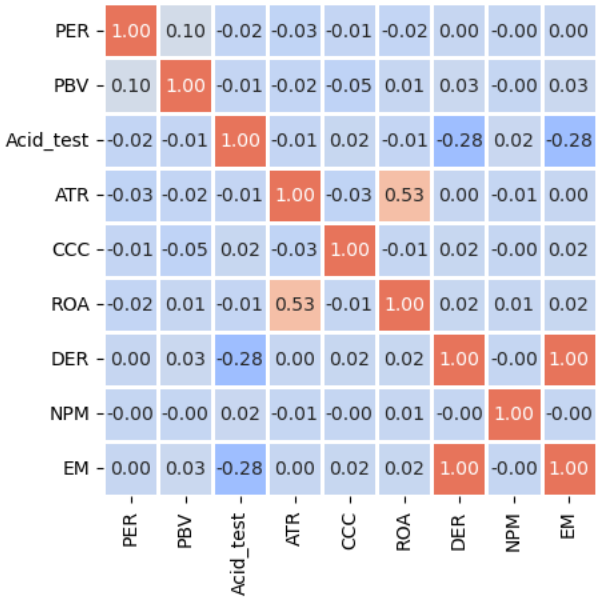


*Figura 12. Análisis de correlación Pearson*

Se realizó una matriz de correlación por el método de Pearson debido a que las variables involucradas son cuantitativas continuas. El análisis entre los ratios financieros revela una rica variedad de información. En su mayoría, los ratios financieros muestran correlaciones insignificantes, tanto positivas como negativas, cercanas a cero con respecto a otros ratios. Esto indica que estas variables son relativamente independientes entre sí y no presentan una homogeneidad significativa. En otras palabras, no comparten similitudes que podrían repetirse en el modelo. Cada ratio puede estar influenciado de manera diferente por factores internos y externos, lo que sugiere que se está teniendo en cuenta la mayor cantidad de información posible para obtener resultados más precisos.

Sin embargo, se observa una excepción en los ratios EM (Earnings Multiple) y DER (Debt-to-Equity Ratio), los cuales muestran una correlación perfecta entre ellos. Esta alta correlación indica que ambas variables podrían estar repitiendo información en el modelo, lo que podría redundar en una pérdida de eficiencia en el análisis. En consecuencia, se valora la posibilidad de eliminar uno de estos ratios para evitar la duplicación de información y simplificar el modelo; sin embargo, este paso se considerará en una etapa futura.

## Análisis de distribuciones

En los gráficos que se presentan a continuación, se muestra la distribución de cada ratio financiero, considerando una porción relevante de las observaciones correspondientes a todas las empresas incluidas en el análisis. No obstante, se destaca que no se representa la totalidad de los datos debido a la presencia de valores atípicos, también conocidos como outliers, los cuales pueden distorsionar la información visual proporcionada por las gráficas. Esta representación gráfica permite una visión integral y conjunta de la variabilidad y estructura de cada uno de los indicadores financieros, abarcando los datos recopilados de todos los activos financieros involucrados en el estudio. La visualización de estas distribuciones resulta fundamental para identificar tendencias y patrones en el conjunto de datos. A continuación, se presenta la tabla que detalla el porcentaje de datos reflejados en cada gráfico:

| Ratio | % de datos contenidos |
|---|---|
| PER | 95.93% |
| PBV | 98.89% |
| Acid test | 100.00 |
| ATR | 97.35% |
| CCC | 94.48% |
| ROA | 97.54% |
| DER | 100.00% |
| NPM | 98.89% |
| EM | 100.00% |

*Figura 13. Análisis de distribuciones*

En seguida, se muestran las distribuciones de los ratios financieros:

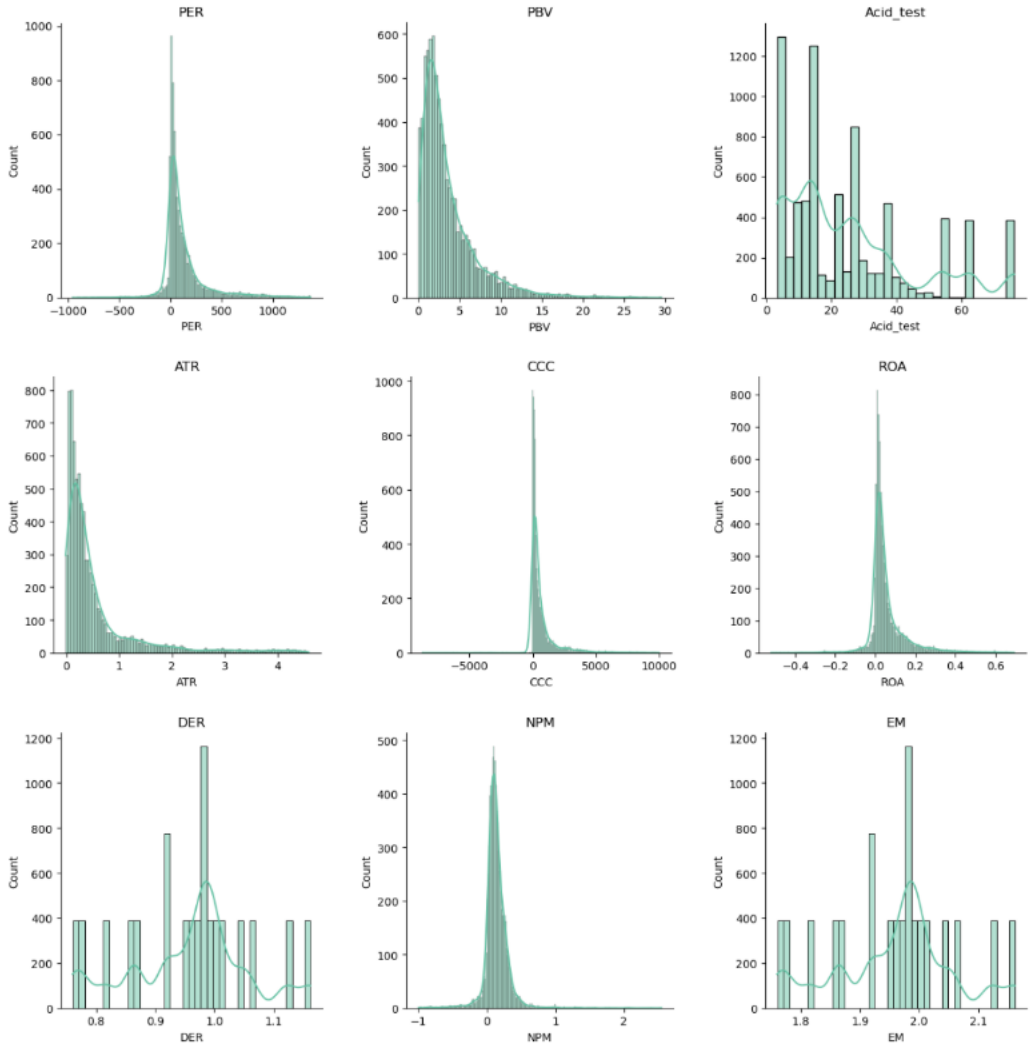


*Figura 14. Distribuciones de ratios financieros*

Tomando en cuenta todos los datos, incluyendo outliers, ajustamos los datos de la mejor manera a las siguientes distribuciones: poisson, lognormal, normal, chi2, uniforme y binomial.

A continuación, se presentan tablas con los ratios financieros y la distribución que mejor se ajustó, incluyendo sus respectivos parámetros.

| Ratio | Tipo de distribución | Media | Varianza |
|---|---|---|---|
| DER | Normal | 0.96 | 0.10 |
| NPM | Normal | 0.78 | 51.06 |
| EM | Normal | 1.96 | 0.10 |

| Ratio | Tipo de distribución | Media | Varianza |
|---|---|---|---|
| PER | Lognormal | 0.01 | 305360.35 |
| PBV | Lognormal | 0.89 | 0.17 |
| Acid test | Lognormal | 5.00 | 3.22 |
| ATR | Lognormal | 1.22 | 0.01 |
| ROA | Lognormal | 0.02 | 16.85 |

| Ratio | Tipo de distribución | k |
|---|---|---|
| CCC | chi2 | 0.2 |

*Figura 15. Mejor distribución para ratios financieros*

Comprender cómo están distribuidos los datos puede ayudar a seleccionar el modelo de clasificación más apropiado, ajustar sus parámetros de manera efectiva y evaluar su desempeño de manera precisa. Además, permite identificar posibles desafíos, como desequilibrios de clases o valores atípicos, que pueden afectar la calidad de las predicciones. Además, se pueden observar sesgos a través de ver que tan desviado de 0 está la media en las dsitribuciones normal y lognormal. Esta alerta de sesgo es de suma importancia, ya que sugiere que estas variables pueden estar sujetas a influencias no deseadas o a patrones que podrían introducir ruido en los modelos. La identificación temprana de este sesgo brinda la oportunidad de abordar estos problemas y ajustar adecuadamente el enfoque analítico.

## Variable dependiente: Rendimiento

La variable objetivo a predecir se refiere al rendimiento financiero, el cual se mide de manera trimestral para cada acción. Este rendimiento se obtiene a través del cálculo de un cambio porcentual simple entre dos periodos consecutivos. Concretamente, el rendimiento en el tiempo "n" se formula de la siguiente manera:

$$Rendimiento_n = \frac{Precio_{n+1}}{Precio_n} - 1$$

*Figura 16. Fórmula de rendimiento en el tiempo "n"*

En función de si el rendimiento en el próximo trimestre resulta ser negativo o positivo, se ha establecido una clasificación binaria. En esta clasificación, la "Clase 1" representa la decisión de SÍ invertir dado un rendimiento efectivo positivo en el trimestre, mientras que la "Clase 0" denota la decisión de NO invertir dado un rendimiento efectivo negativo en el trimestre. Este enfoque de clasificación se basa en las fluctuaciones del rendimiento lo que permite determinar si es aconsejable abstenerse de realizar una inversión, considerando el comportamiento de los activos financieros, cabe destacar que se toma una periodicidad trimestral debido a que los reportes financieros de las empresas son publicados cada 3 meses.

## Análisis de correlación Spearman

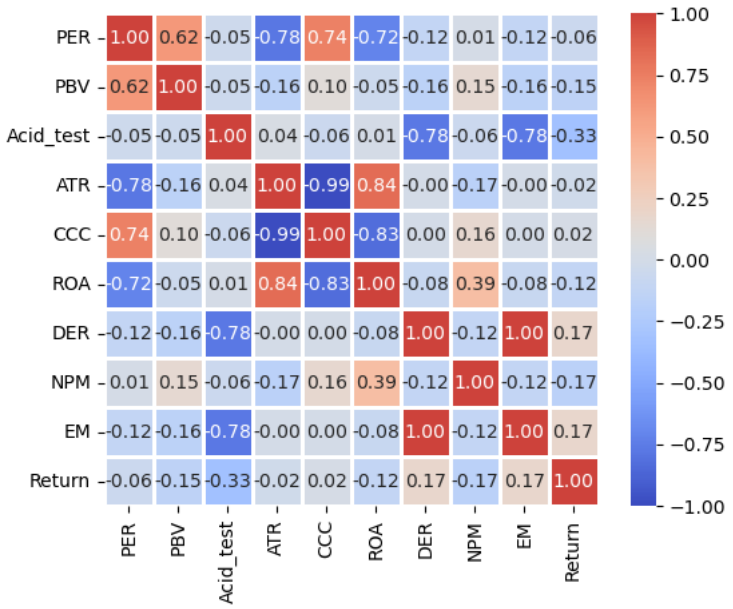


*Figura 17. Análisis de correlación Spearman*

Posteriormente se realizó una matriz de correlación por el método de Spearman. Este análisis de correlación entre los ratios financieros es más estricto, pero aun así nos muestra

la variedad de información que podemos obtener para fortalecer nuestro modelo. Se observan ratios financieros con correlaciones muy cercanas a cero tanto positivas como negativas, lo cual nos indica la independencia entre ellas y la variedad de información que obtiene el modelo para mejorar su eficiencia. En otras palabras, no comparten similitudes que podrían repetirse en el modelo. Cada ratio puede estar influenciado de manera diferente por factores internos y externos, lo que sugiere que se está teniendo en cuenta la mayor cantidad de información posible para obtener resultados más precisos.

Por otro lado, por ser este método más estricto nos muestra que algunos ratios están aún más correlacionados como, por ejemplo, CCC, PER, ATR y ROA por lo cual nos indica que se podría estar repitiendo información sobre el modelo y podría ocasionar una pérdida de eficiencia en nuestro análisis. En consecuencia, se valora la posibilidad de eliminar uno o más de estos ratios para evitar la duplicación de información y simplificar el modelo.

Por último, se observa la correlación de la variable dependiente "Return" con las variables independientes. Acid test es la variable que mayor correlación tiene, seguido por DER y EM. Esta información podría ser útil en caso de tomar variables significativas para el modelaje.

## Conclusiones EDA

En conclusión, este análisis exploratorio de datos ha proporcionado una comprensión sólida de la base de datos financiera. Se han identificado relaciones interesantes y se han destacado aspectos críticos para futuros análisis. Además, se ha planteado eliminar CCC y EM por su alta correlación con otros ratios financieros. Este proceso es fundamental para garantizar que el análisis subsiguiente se realice de manera precisa y efectiva, permitiendo tomar decisiones informadas bajo el contexto en el que se está trabajando.

## Modelaje

Tras llevar a cabo un análisis exploratorio de los datos, procedimos a la partición de estos en dos conjuntos fundamentales: un conjunto de entrenamiento, que abarca el 80% de los datos, y un conjunto de prueba, que comprende el restante 20%. Este proceso de partición se realizó con el propósito de poder evaluar el desempeño y capacidad predictiva de diferentes modelos. Para la elección del modelo óptimo hemos considerado y evaluado diversas opciones. Una vez que se determine el modelo más adecuado, se procederá a su validación, ajuste y optimización, evaluando la precisión de sus resultados. Este proceso nos permitirá alcanzar un modelo efectivo que se adapte de manera óptima a los datos y cumpla con los objetivos planteados en nuestra investigación.

### *Regresión logística*

El modelo de regresión logística, reconocido por su carácter simplista y funcionalidad eficaz, sirve como una excelente línea de base en el campo del aprendizaje automático. Su simplicidad radica en la facilidad con que puede ser implementado y su eficiencia computacional, lo que permite una rápida iteración y evaluación de modelos. No obstante, aunque es un modelo fácil de entender y de aplicar, la simplicidad también implica ciertas limitaciones en su capacidad de manejar la complejidad y las interacciones no lineales entre variables.

| Metric / Class | Accuracy | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|---|
| Class 0 | 0.61 | 0.42 | 0.03 | 0.06 | 510 |
| Class 1 | | 0.62 | 0.97 | 0.76 | 827 |

*Figura 18. Métricas Regresión Logística*

Los resultados al aplicar el modelo con nuestros datos de entrenamiento y prueba dan un accuracy de 61.40%, que es la proporción de predicciones correctas. Sin embargo, al medir la precisión en las predicciones en 0 y 1, podemos observar que en clase 1, que indica la

instrucción de "invertir", da un 62%, pero en la clase 0, que indica "no invertir", tenemos una precisión de 42%. Este indicador nos da indicios de que no es el modelo que buscamos, pues la precisión en la clase 1 todavía puede considerarse deficiente para el contexto de generación de riqueza sustentable que se busca, y, por lo tanto, no usaremos este modelo y se buscará otras formas de mejorar las métricas de ajuste. Nuestra principal intención es la de mejorar el rendimiento del modelo, por lo cual consideramos necesario explorar y aplicar modelos más complejos y avanzados en el proceso de optimización.

## Identificar Variables Significativas

La selección de variables significativas es un paso crucial en la construcción de modelos predictivos robustos y eficientes. No sólo mejora la interpretabilidad del modelo al reducir la complejidad, sino que también potencia la capacidad predictiva al enfocarse en las variables con mayor impacto. Durante este proceso, se implementó la técnica de Eliminación hacia Atrás (Backward Elimination), un enfoque sistemático y riguroso que empieza con todas las variables candidatas y elimina la menos significativa en cada iteración, basándose en un criterio de significancia estadística establecido previamente.

Este criterio se centra en la significancia de los coeficientes de las variables, generalmente evaluados mediante pruebas de hipótesis. El p-value asociado con cada coeficiente es examinado contra un umbral de significancia, como el estándar 0.05, para determinar si hay evidencia suficiente para rechazar la hipótesis nula de que el coeficiente es igual a cero (lo que implicaría que la variable no tiene efecto). Variables con p-values altos significan poca influencia dentro del modelo, ya que su contribución a la capacidad predictiva del modelo no es estadísticamente significativa.

Mediante este método de selección hacia atrás, se identificaron y retuvieron solo aquellas variables con p-values bajos, sugiriendo que los coeficientes asociados eran diferentes de

cero y, por ende, tenían un efecto significativo en la respuesta del modelo. La resultante fue un conjunto concentrado de tres variables de alta relevancia, cuya significancia fue respaldada por pruebas de hipótesis rigurosas. Estas variables conforman la esencia del modelo, aportando a la precisión de las predicciones y a la robustez general del mismo, asegurando que cada predictor incluido aporte valor informativo significativo.

Con cada proceso de eliminación de variable, se actualizaba el modelo de regresión logística para evaluar su desempeño. Sin embargo, no se observó una mejora significativa en los resultados a pesar de la actualización iterativa del modelo. Por tanto, se tomó la decisión de retener el conjunto original de variables consideradas inicialmente y continuar con la evaluación de otros modelos predictivos. Esta decisión se basó en la comprensión de que, aunque la simplificación del modelo es deseable, la exclusión de variables no debería comprometer la integridad del análisis ni la capacidad de capturar la complejidad inherente a los datos.

| Variables |
| --- |
| Acid test |
| DER |
| EM |

*Figura 19. Variables Significativas Elegidas*

## Prueba 13 Modelos

Se procede a la evaluación de trece modelos distintos, los cuales se detallan en la lista siguiente:

| Modelos utilizados |
|---|
| Gradient Boosting Classifier |
| HGBC |
| Random Forest Classifier |
| XGBoost |
| Decision Tree Classifier |
| K-Nearest Neighbors (uniformly weighted) |
| K-Nearest Neighbors (weighted by distance) |
| Support Vector Classifier |
| Multilayer Perceptron |
| Stochastic Gradient Descent |
| Logistic Regression |
| Ridge Classifier CV |
| Linear Discriminant Analysis |

*Figura 20. Lista de modelos a evaluar*

Las pruebas se realizan con los datos modificados de esta manera:

a) Estandarizados, todas las variables
b) No Estandarizados, todas las variables
c) Estandarizado, variables significativas
d) No Estandarizado, variables significativas

Además de estos enfoques, se probaron distintas transformaciones en los 13 modelos, incluyendo técnicas de reducción de la dimensionalidad como el Análisis de Componentes Principales (PCA) y el Análisis Discriminante Lineal (LDA). Estas técnicas buscan simplificar la tarea del algoritmo manteniendo la información relevante; sin embargo, no se llegaron a resultados significativos con estas transformaciones.

A continuación, se presentan los resultados más valiosos de este proceso que fueron los datos Estandarizados con todas las Variables y datos sin Estandarizar con todas las Variables:

***Estandarizado con todas las variables***

Se observa que el mejor resultado es HGBC con un Accuracy en el Train de 85.69% y un Accuracy en la prueba de 73.82%. Aunque el Accuracy Train es mayor en otros modelos, se elige el modelo que tenga mejor Accuracy en la prueba, ya que ello significa que puede predecir mejor nuestros datos, además de que una calificación tan buena en el train sugiere que fue sobreentrenado.

| Model | Accuracy Train | Accuracy Test |
|---|---|---|
| **HGBC** | **85.69%** | **73.82%** |
| Random Forest Classifier | 100.0% | 73.75% |
| Gradient Boosting Classifier | 77.95% | 73.67% |
| XGBoost | 94.93% | 71.35% |
| Decision Tree Classifier | 100.0% | 65.0% |
| Support Vector Classifier | 63.64% | 62.23% |
| Stochastic Gradient Descent | 61.0% | 61.78% |
| K-Nearest Neighbors (weighted by distance) | 100.0% | 61.48% |
| Logistic Regression | 61.43% | 61.33% |
| Linear Discriminant Analysis | 61.39% | 61.33% |
| Ridge Classifier CV | 61.39% | 61.18% |
| Multilayer Perceptron | 61.54% | 61.11% |
| K-Nearest Neighbors (uniformly weighted) | 74.93% | 60.58% |

*Figura 21. Resultado de prueba "Estandarizado con variables significativas"*

En el contexto de la estandarización con variables significativas, se destaca que el mejor desempeño se atribuye al modelo Gradient Boosting Classifier (GBC), que exhibe una precisión (Accuracy) del 77.95% en el conjunto de entrenamiento y del 73.67% en el conjunto de prueba. Aunque otros modelos pueden presentar un Accuracy superior en el conjunto de entrenamiento, se prioriza la capacidad del modelo para generalizar y obtener resultados precisos en situaciones del mercado real.

***Sin Estandarizar con todas las variables***

Se obtiene un rendimiento superior en términos de precisión (Accuracy) tanto en el conjunto de entrenamiento como en el conjunto de prueba para Random Forest. No obstante, la diferencia entre el accuracy en el train y el test sugiere que el modelo está sobreentrenado, lo que mantiene al modelo Gradient Boosting Classifier (GBC) como el mejor clasificador. Es importante destacar que se observan indicios de overfitting en varios modelos (no sólo Random Forest), lo que sugiere una adaptación excesiva a los datos de entrenamiento. Es por ello por lo que se concede prioridad al indicador de Accuracy en el conjunto de prueba, que alcanza un valor del 73.67%. Siguiendo el principio de parsimonia, se toma la decisión de seleccionar el modelo GBC que utiliza todas las variables sin estandarización, en virtud de su desempeño consistente y balanceado en el conjunto de prueba.

| Modelo | Accuracy Train | Accuracy Test |
|---|---|---|
| **Random Forest Classifier** | **100.0%** | **73.75%** |
| Gradient Boosting Classifier | 77.95% | 73.67% |
| HGBC | 85.99% | 73.07% |
| XGBoost | 94.93% | 71.35% |
| Decision Tree Classifier | 100.0% | 65.0% |
| Support Vector Classifier | 61.26% | 61.63% |
| Multilayer Perceptron | 61.78% | 61.41% |
| Logistic Regression | 61.28% | 61.41% |
| Linear Discriminant Analysis | 61.39% | 61.33% |
| Ridge Classifier CV | 61.34% | 61.18% |
| K-Nearest Neighbors (uniformly weighted) | 74.04% | 59.76% |
| K-Nearest Neighbors (weighted by distance) | 100.0% | 58.71% |
| Stochastic Gradient Descent | 44.39% | 43.38% |

*Figura 22. Resultado de prueba "Sin estandarizar con todas las variables"*

***Stacking***

Con la finalidad de alcanzar un modelo de mayor rendimiento, se optó por la implementación de una técnica denominada "Stacking". El proceso de Stacking consiste en la combinación de múltiples modelos base en un enfoque jerárquico, buscando mejorar la precisión predictiva del modelo final. Se realizaron dos de estas pruebas, una con todos los modelos; y otra en donde se seleccionaron los modelos en los que se observó mejor desempeño en el conjunto de prueba: Random Forest, HGBC, GBC y XGboost. Si bien esta estrategia puede conllevar un aumento sustancial en la exactitud de las predicciones, en este caso específico, no ha logrado superar el desempeño observado en el modelo Gradient Boosting Classifier (GBC). En consecuencia, se descarta la utilización de esta técnica y se procede a comparar los valores de Accuracy para sustentar esta decisión:

| Modelo | Accuracy Train | Accuracy Test |
|---|---|---|
| **Stacking de todos los Modelos** | **90.83%** | **74.5%** |
| Stacking Manual | 100.0% | 73.74% |

*Figura 23. El resultado de prueba sugiere un sobreajuste de ambos modelos a los datos de entrenamiento.*

## Elección de Modelo Gradient Boosting Classifier (GBC)

Inicialmente, se estableció la regresión logística como el punto de referencia (benchmark) en nuestro estudio. Este modelo, aunque simplista y funcional, a menudo se ve limitado por su incapacidad para capturar relaciones complejas y no lineales en los datos. Sin embargo, tras pruebas y análisis comparativos, se ha concluido que el modelo Gradient Boosting Classifier (GBC) sobresale como la elección óptima para abordar nuestro problema específico.

El modelo Gradient Boosting Classifier (GBC) se distingue por su utilización de gradientes, esenciales en el campo de data science, para optimizar su desempeño. Al emplear el descenso del gradiente, el GBC ajusta sistemáticamente los parámetros de los árboles de decisión subsecuentes con el fin de minimizar la función de pérdida, una métrica que cuantifica el error del modelo. Esta aproximación aprovecha la información del gradiente para saber en qué dirección ajustar los pesos de las características, y así mejorar la predicción con cada nuevo árbol añadido al modelo.

Este proceso iterativo y su enfoque en la optimización basada en gradientes permiten al GBC adaptarse con precisión a las complejidades del conjunto de datos, lo que es crítico para obtener una alta capacidad predictiva en presencia de datos complejos y ruidosos. Además, el GBC es particularmente eficaz en situaciones donde los datos presentan un desequilibrio, ya que se centra en las instancias más difíciles de clasificar, ajustando de manera incremental los modelos para mejorar en las áreas donde los modelos previos se equivocaban más. Por ende, el empleo de gradientes y su importancia en el ajuste preciso de modelos en data science, convierte al GBC en una herramienta potente para la generación de modelos predictivos con altos niveles de precisión y robustez. Para optimizar aún más el rendimiento de nuestro modelo, se explorarán enfoques adicionales, como el uso de Optuna para el fine tunning de hiperparámetros, con el propósito de mejorar y afinar nuestras métricas.

## Métricas Train

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.61 | 0.68 | 2076 |
| 1 | 0.78 | 0.89 | 0.83 | 3270 |
| accuracy | | | 0.78 | 5346 |
| macro avg | 0.78 | 0.75 | 0.76 | 5346 |
| weighted avg | 0.78 | 0.78 | 0.77 | 5346 |

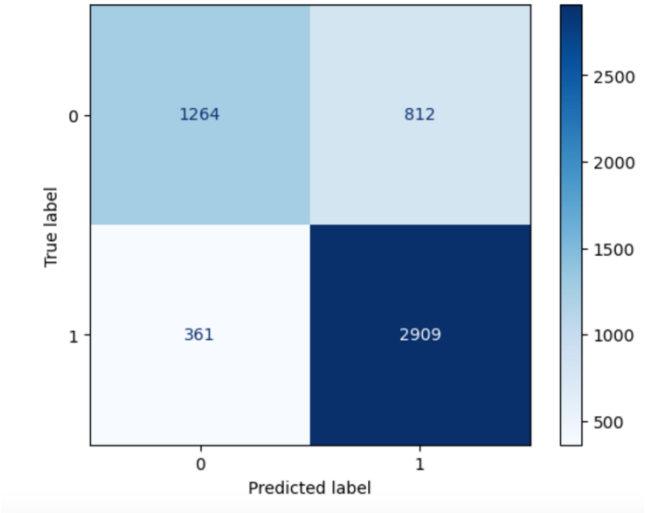


*Figura 24. Métricas de GBC en Train*

## Métricas Test

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.71 | 0.53 | 0.61 | 510 |
| 1 | 0.75 | 0.87 | 0.80 | 827 |
| accuracy | | | 0.74 | 1337 |
| macro avg | 0.73 | 0.70 | 0.71 | 1337 |
| weighted avg | 0.74 | 0.74 | 0.73 | 1337 |

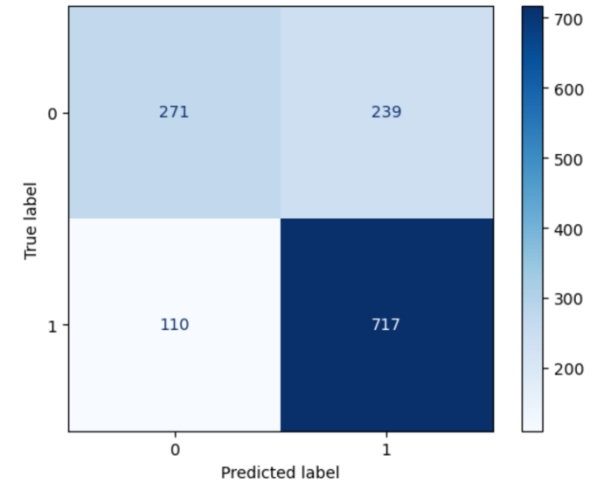


*Figura 25. Métricas de GBC en Test*

## GBC optimizado por Optuna

Optuna es una biblioteca de optimización de hiperparámetros diseñada con un enfoque en la automatización y eficiencia en la búsqueda del mejor conjunto de parámetros para modelos de aprendizaje automático. Lo que distingue a Optuna es su capacidad para realizar una búsqueda guiada utilizando técnicas de optimización avanzadas, como el TPE (Tree-structured Parzen Estimator), que es un enfoque bayesiano. La utilización de métodos bayesianos le permite aprender de los resultados de evaluaciones anteriores para dirigir la búsqueda hacia las combinaciones más prometedoras de parámetros, lo cual puede ser más eficiente que los enfoques de búsqueda aleatoria (random search) o de cuadrícula (grid search).

La idea detrás de la optimización bayesiana, y cómo Optuna la incorpora, es construir un modelo probabilístico de la función objetivo, que en este caso sería el f1-score del modelo Gradient Boosting Classifier. Este modelo probabilístico se actualiza a medida que Optuna evalúa más y más configuraciones de parámetros. El f1-score es una métrica particularmente útil para la optimización cuando se trata de problemas de clasificación con conjuntos de datos desequilibrados o cuando el equilibrio entre precisión y recuperación (precision and recall) es crucial. Optuna utiliza la información acumulada sobre el rendimiento de las configuraciones pasadas para decidir inteligentemente qué conjunto de parámetros probar a continuación, lo que a menudo lleva a encontrar mejores hiperparámetros más rápidamente que otros métodos de optimización.

El uso de Optuna comienza con la definición de una función objetivo que evalúa el modelo de aprendizaje automático con un conjunto dado de hiperparámetros y devuelve el f1-score que desea maximizar. Durante el proceso de optimización, Optuna itera sobre diferentes combinaciones de hiperparámetros, llamando a esta función objetivo y ajustando su modelo probabilístico basado en los resultados obtenidos. A medida que se realizan más ensayos, Optuna se vuelve más 'inteligente' en la selección de hiperparámetros que son

más propensos a mejorar la métrica de rendimiento, guiando así una búsqueda eficiente en el espacio de hiperparámetros.

A continuación se muestran los resultados de esta optimización.

**Métricas Train**

| Metric | Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| **0** | Class 0 | 95.59% | 97.28% | 91.18% | 94.13% |
| **1** | Class 1 | 95.59% | 94.62% | 98.38% | 96.46% |

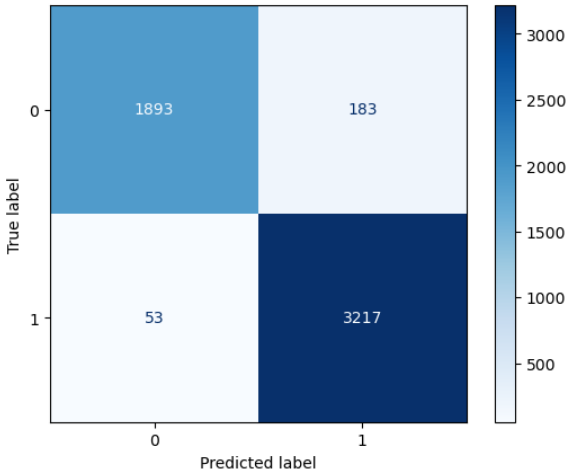


*Figura 26. Métricas de GBC optimizado por Optuna en Train*

**Métricas Test**

| Metric | Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| **0** | Class 0 | 74.57% | 71.36% | 55.69% | 62.56% |
| **1** | Class 1 | 74.57% | 75.93% | 86.22% | 80.75% |

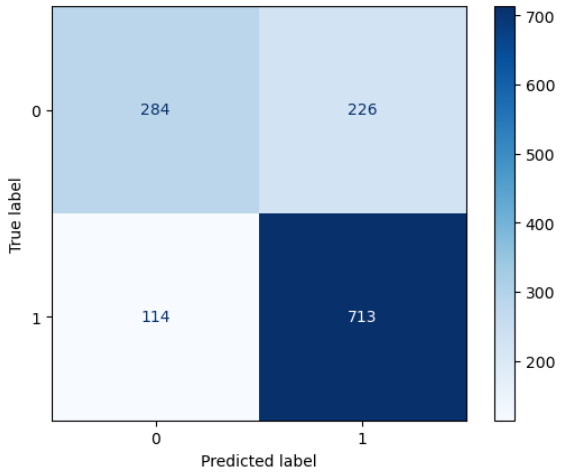


*Figura 27. Métricas de GBC optimizado por Optuna en Test*

## Conclusiones de Modelaje

El análisis comparativo entre el modelo Gradient Boosting Classifier (GBC) y su versión optimizada con Optuna revela una cuestión crítica de overfitting en la versión optimizada. Este fenómeno se evidencia claramente en las diferencias significativas entre las métricas de desempeño en los conjuntos de entrenamiento y prueba.

Mientras que el GBC optimizado por Optuna muestra métricas excepcionales en el conjunto de entrenamiento con precisión y recall superiores al 94% para ambas clases, estos números caen drásticamente en el conjunto de prueba, donde el recall de la clase 0 se reduce a 55.69% y la precisión de la misma clase a 71.36%. Este descenso es indicativo de que el modelo está excesivamente ajustado a los datos de entrenamiento, lo que compromete su capacidad para generalizar sobre datos no vistos.

En contraste, el modelo GBC sencillo, aunque presenta métricas ligeramente inferiores en el entrenamiento, mantiene un nivel de rendimiento más equilibrado y consistente entre los conjuntos de entrenamiento y prueba. Esto sugiere una mejor generalización y una robustez superior frente a variaciones en los datos, lo que es esencial en aplicaciones prácticas donde los modelos se enfrentan a datos reales y a menudo ruidosos.

El riesgo de tomar decisiones de inversión basadas en un modelo sobreajustado como el GBC optimizado por Optuna es considerable, dado que podría llevar a decisiones erróneas al confiar en patrones que no son aplicables universalmente, sino que están demasiado adaptados a las particularidades del conjunto de entrenamiento. En contextos de inversión, donde las consecuencias de una mala decisión pueden ser significativas, es preferible optar por un modelo que ofrezca una garantía más sólida de rendimiento consistente y fiable en diversos escenarios, minimizando así el riesgo asociado a decisiones erróneas basadas en un entendimiento distorsionado de las tendencias del mercado.

Por estas razones, la decisión de no elegir el GBC optimizado por Optuna es prudente y está alineada con el objetivo de mantener la integridad y la fiabilidad en la toma de decisiones de inversión.

**Clustering para ver el performance del modelo en cada industria**

Con el objetivo de buscar la generalización del modelo, realizamos el análisis del desempeño del modelo en diferentes sectores del índice S&P 500. En los sectores "Consumer Discretionary" e "Industrials" sobresalen al exhibir valores elevados de F1-Score, lo que sugiere un equilibrio efectivo entre precisión y exhaustividad en las predicciones. En contraste, el sector "Energy" presenta un F1-Score relativamente inferior, indicando un posible desequilibrio entre precisión y exhaustividad en sus resultados. Este desbalance podría deberse a la naturaleza específica de los datos o a la necesidad de ajustes adicionales en el modelo para mejorar su capacidad predictiva en ese contexto.

| Sector | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Health Care | 0.76 | 0.77 | 0.86 | 0.81 |
| Industrials | 0.82 | 0.82 | 0.91 | 0.86 |
| Consumer Discretionary | 0.80 | 0.79 | 0.91 | 0.85 |
| Information Technology | 0.80 | 0.81 | 0.89 | 0.85 |
| Financials | 0.77 | 0.81 | 0.86 | 0.83 |
| Consumer Staples | 0.74 | 0.75 | 0.86 | 0.80 |
| Utilities | 0.72 | 0.71 | 0.90 | 0.79 |
| Materials | 0.77 | 0.77 | 0.89 | 0.82 |
| Real Estate | 0.79 | 0.76 | 0.94 | 0.84 |
| Energy | 0.67 | 0.70 | 0.79 | 0.74 |
| Communication Services | 0.77 | 0.76 | 0.90 | 0.82 |

*Figura 28. Análisis del desempeño del modelo en diferentes sectores del índice S&P 500*

Además, al realizar un análisis general del modelo, se observa una variación en la precisión (Accuracy) según el sector. En particular, se destaca el alto Accuracy en los sectores de "Industrials", "Consumer Discretionary" y “Information Technology”, con valores de 0.82, 0.80 y 0.80, respectivamente. Estos resultados reflejan la capacidad del modelo para realizar predicciones precisas en dichos sectores. Por otro lado, el sector "Energy" muestra un Accuracy ligeramente inferior, alcanzando un valor de 0.67, lo que sugiere que el modelo podría beneficiarse de ajustes específicos en ese ámbito.

## Backtesting

El backtesting, en el contexto de este proyecto, es un procedimiento fundamental que implica la evaluación y prueba de una estrategia, su propósito es determinar la eficacia y viabilidad del modelo de aprendizaje automático (ML) seleccionado en el contexto de su aplicación práctica. A continuación, se detalla el procedimiento seguido para documentar exhaustivamente la estrategia adoptada.

En la fase inicial, integramos una base de datos que incluyera las empresas seleccionadas junto con su correspondiente información financiera. Este proceso fue importante, ya que la calidad y la precisión de los datos ingresados serían determinantes en la eficacia del modelo de Machine Learning (ML) aplicado posteriormente. Una vez alimentado el modelo con estos datos, se logró generar un vector de predicciones compuesto por ceros y unos. Este vector nos proporciona indicaciones claras sobre la conveniencia de invertir en cada empresa a lo largo del tiempo. Esta metodología, basada en la inteligencia artificial, nos permite realizar análisis predictivos con un grado de precisión alto, según las métricas de ajuste del modelo mostradas en la sección anterior.

La ponderación de los activos en el portafolio se determinó utilizando una optimización basada en el índice Omega. La estrategia detrás de este índice busca equilibrar dos objetivos fundamentales: minimizar el riesgo de pérdida (downside risk) y maximizar el potencial de ganancia (upside risk). Aspiramos a contar con activos que exhiban una volatilidad favorable en términos de ganancias, pero que al mismo tiempo mantengan una volatilidad reducida en las pérdidas. Este enfoque es particularmente relevante en contextos de mercado inciertos, donde la gestión eficaz del riesgo es primordial.

En la etapa de implementación, se partió con un capital inicial de USD $1'000,000. Con este capital, se procedió a la creación de un portafolio de inversión, aplicando un filtro para seleccionar los activos más adecuados según las recomendaciones del modelo de ML. Para cada "fiscal date", que corresponde a cada período (trimestre) en el que se simuló la

estrategia de compra y venta de activos, se seleccionaron al azar cinco de estos activos recomendados. Esta selección aleatoria añade un elemento de diversificación y mitigación de riesgo al proceso de inversión.

En el caso de que el algoritmo no recomendara la inversión en al menos cinco activos, se desarrolló una estrategia alternativa que consistía en asignar la ponderación correspondiente a bonos del tesoro de Estados Unidos, específicamente a aquellos que cotizan bajo el ticker "^IRX" o 13 Week Treasury Bills. Aquí, es esencial tener en cuenta que, tal y como en la inversión directa en tasas como las de Banxico o la FED donde no hay comisión, la compra de estos bonos no conlleva un costo transaccional.

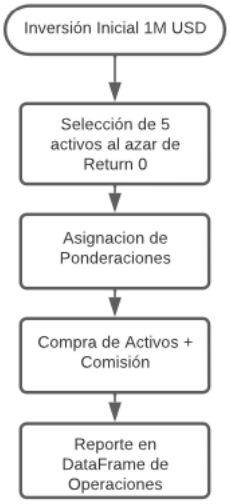


*Figura 29. Diagrama de flujo estrategia*

Por ejemplo, si el algoritmo selecciona 5 activos convenientes a invertir, el portafolio óptimo por el índice de Omega calcula las ponderaciones óptimas y se invierte el 100% en ellos, es decir, el 100% en activos recomendados y 0% en bonos estadounidenses a la tasa vigente para ese período; por otro lado, en dado caso de que solo se seleccionaran 4 activos convenientes de inversión el 80% se invertiría en activos y 20% se invertiría en los bonos estadounidenses y así sucesivamente. Siguiendo el ejemplo anterior, si solo se seleccionan 3 activos, las ponderaciones proporcionadas por la optimización de la asignación de activos

basada en Omega se multiplicarán por 0.60 (o 60%), evitando así una sobreexposición a un número limitado de activos y mejorando la diversificación del portafolio.

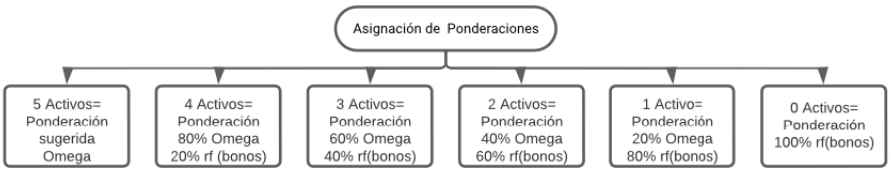


*Figura 30. Asignación de ponderaciones*

Para continuar, la estrategia descrita anteriormente se diseñó con un enfoque en la eficiencia de ahorro de comisiones, considerando una comisión del 0.025% + IVA (16%) por cada transacción de compra y venta. En cada "fiscal date", se evalúa primero si los activos en el portafolio siguen siendo recomendaciones de inversión según el algoritmo, si no lo son, se liquida la posición; de lo contrario, se lleva a cabo un reequilibrio de las posiciones de acuerdo con las ponderaciones sugeridas por la estrategia Omega.

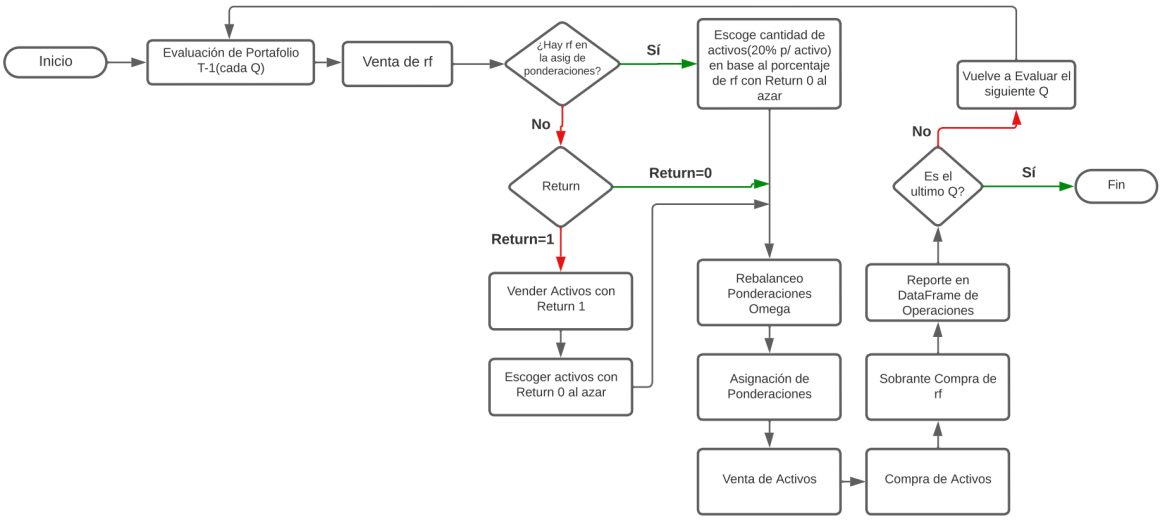


*Figura 31. Diagrama de flujo del código de la estrategia*

Por último, cabe subrayar que todas las transacciones efectuadas en el marco de nuestra estrategia de inversión se basaron en el precio vigente de las acciones en el momento de la ejecución. Además, es relevante señalar que la selección inicial de acciones ejerce una influencia significativa en el algoritmo, ya que estas constituyen la base para la toma de

decisiones subsiguientes respecto a la composición del portafolio. En aras de validar la robustez y la eficiencia del algoritmo bajo distintas condiciones de mercado, se llevaron a cabo 100 simulaciones de Montecarlo. Estas simulaciones, aplicando la ley de los grandes números, nos permiten asegurar que las métricas utilizadas para evaluar el rendimiento del portafolio promedio son el resultado de una convergencia estadística de los resultados, ofreciendo así una confirmación de la solidez de la estrategia de inversión implementada.

## Resultados de Backtesting

### *Condiciones del Mercado (Referencia)*

Establece un marco comparativo para el análisis, utilizando como índice de referencia el S&P 500, con un retorno de benchmark del 16.3321%. Este indicador es esencial para evaluar el rendimiento de nuestro modelo en relación con el mercado en general. La Tasa Libre de Riesgo (Risk-Free Rate) se fijó en un 5.1530% al inicio de la implementación de nuestra estrategia, proporcionada por la Reserva Federal (FED). Este parámetro se utiliza como punto de referencia para establecer el rendimiento sin riesgo, permitiendo comparar el exceso de retorno obtenido por el modelo sobre una inversión libre de riesgo.

| Metric | Value |
|---|---|
| Benchmark Return | 16.3321% |
| Risk Free (FED) | 5.1530% |

*Figura 32. Métricas de condiciones del mercado*

### *Estadísticas del Portafolio*

La Rentabilidad Anual del portafolio en de 71.5602%, lo que indica una superación tanto del rendimiento del mercado como de la tasa libre de riesgo.

La Volatilidad Anual se sitúa en el 67.3916%, proporcionando una medida de la variabilidad de los retornos del portafolio.

El Beta de -0.1798 señala una correlación inversa reducida con los movimientos del mercado, implicando que el portafolio es menos sensible a las fluctuaciones del mercado en comparación con el índice de referencia.

| Metric | Value |
|---|---|
| Annual Return | 71.5602% |
| Annual Volatility | 67.3916% |
| Beta | - 0.1798 |

*Figura 33. Métricas de estadísticas del portafolio*

***Ratios Financieros***

**Sharpe Ratio**

El Sharpe Ratio mide la rentabilidad excedente por unidad de riesgo en comparación con la tasa libre de riesgo. Se calcula utilizando la siguiente fórmula:

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

*Figura 34. Sharpe Ratio Fórmula*

donde ( $R_p$) es el retorno del portafolio, ($R_f$) es la tasa libre de riesgo, y ( $\sigma_p$ ) es la desviación estándar de los retornos del portafolio.

Nuestra Métrica: Con un valor de 0.9854, este ratio destaca la capacidad del modelo para generar rendimientos ajustados al riesgo por encima de la tasa libre de riesgo.

**Information Ratio**

El Information Ratio mide la habilidad del modelo para generar retornos excesivos en relación con un índice de referencia, ajustados por la dispersión de esos excesos de retorno. Su fórmula es:

$$\text{Information Ratio} = \frac{R_p - R_b}{\sigma_{p-b}}$$

*Figura 35. Information Ratio Fórmula*

donde $R_p$) es el retorno del portafolio, ($R_b$) es el retorno del índice de referencia, y ( $\sigma_{p-b}$ ) es la desviación estándar de la diferencia entre los retornos del portafolio y del índice.

Nuestra Métrica: Con un valor de 0.3984, este ratio refleja la capacidad del modelo para superar al índice de referencia.

**Jensen's Alpha**

Jensen's Alpha indica el rendimiento de un portafolio sobre lo esperado por el CAPM, ajustado por el riesgo del mercado. Se calcula como:

$$\alpha = R_p - \left(R_f + \beta\left(R_m - R_f\right)\right)$$

*Figura 36. Jensen's Alpha Fórmula*

donde ( $\alpha$) es el Alpha de Jensen, ( $R_p$) es el retorno del portafolio, ( $R_f$ ) es la tasa libre de riesgo, ($\beta$ ) es el beta del portafolio, y ( $R_m$ ) es el retorno del mercado.

Nuestra Métrica: Un valor de 0.6842 sugiere un rendimiento superior al predicho por el CAPM.

**Treynor Ratio**

El Treynor Ratio mide el retorno por cada unidad de riesgo de mercado asumido. Se define como:

$$\text{Treynor Ratio} = \frac{R_p - R_f}{\beta_p}$$

*Figura 37. Treynor Ratio Fórmula*

donde ( $R_p$) es el retorno del portafolio, ( $R_f$ ) es la tasa libre de riesgo, y ( $\beta_p$ ) es el beta del portafolio.

Nuestra Métrica: Se reporta un -369.2869% reflejando principalmente la peculiaridad de un beta negativo, lo que puede indicar una estrategia defensiva o contraria a las tendencias del mercado. A pesar de este valor negativo extremo en el Treynor, otros indicadores como los ratios de Sharpe y Sortino, que son positivos, demuestran que el portafolio maneja efectivamente el riesgo no sistémico.

**Sortino Ratio**

El Sortino Ratio mide el exceso de rendimiento sobre un umbral mínimo en relación con la volatilidad a la baja. Se calcula como:

$$\text{Sortino Ratio} = \frac{R_p - R_t}{\sigma_d}$$

*Figura 38. Sortino Ratio Fórmula*

donde ($R_p$ ) es el retorno del portafolio, ( $R_t$) es el retorno mínimo aceptable, y ( $\sigma_d$ ) es la desviación estándar de los retornos negativos del portafolio.

Nuestra Métrica: Un Sortino Ratio de 15.4521 sugiere que la desviación estándar de las pérdidas en el portafolio es relativamente baja, lo que explica por qué este ratio es

considerablemente alto. Esto es positivo porque indica que el portafolio maneja eficazmente las caídas, proporcionando una buena referencia sobre las pérdidas esperadas y permitiendo estrategias efectivas para mitigarlas.

**Omega Ratio**

El Omega Ratio compara la probabilidad y magnitud de obtener rendimientos por encima de un umbral con aquellos por debajo de él. Se expresa como:

$$\text{Omega Ratio} = \frac{\int_{R_t}^{\infty} \big(1 - F(R)\big)\, dR}{\int_{-\infty}^{R_t} F(R)\, dR}$$

*Figura 39. Omega Ratio Fórmula*

donde ($R_t$ ) es el umbral de rendimiento mínimo aceptable y ( $F(R)$ ) es la función de distribución acumulativa de los rendimientos.

Nuestra Métrica: Un Omega Ratio de 1.0000 indica que, para el portafolio en cuestión, la probabilidad y magnitud de obtener rendimientos positivos son equitativas con la probabilidad y magnitud de rendimientos negativos. Este valor sugiere que el portafolio apenas está generando rendimientos que compensan el riesgo asumido, ofreciendo una recompensa balanceada por cada unidad de riesgo a la baja.

| Metric | Value |
|---|---|
| Sharpe Ratio | 0.9854 |
| Information Ratio | 0.3984 |
| Jensen´s Alpha | 0.6842 |
| Treynor Ratio | -369.2869% |
| Sortino Ratio | 15.4521 |
| Omega Ratio | 1.0000 |

*Figura 40. Métricas de ratios financieros*

***Medidas de Riesgo***

La siguiente tabla presenta el Valor en Riesgo (VaR) y el Shortfall Esperado, que son medidas de la pérdida potencial en condiciones de mercado adversas. Estas métricas son vitales para entender la cantidad máxima que se podría perder en un porcentaje dado del tiempo, ayudando así a evaluar la exposición al riesgo del portafolio.

Un ejemplo de interpretación del Value at Risk (VaR) al 95% sería que en el 5% de los casos, la pérdida esperada podría superar los $6,296,881.63, asumiendo un capital inicial de $1,000,000. Esto implica que, bajo las condiciones normales de mercado y la escala del portafolio durante el período de análisis, existe una probabilidad del 5% de que la pérdida exceda esa cantidad. Además, el Expected Shortfall al 95% indica que, en esos casos donde la pérdida excede los $9,557,907.25, la pérdida media esperada sería del 9.5579% del valor del portafolio, lo cual es equivalente a $9,557,907.25.

| Metric | Value % |
|---|---|
| VaR at 95% | -6.2969% |
| VaR at 99% | -8.9057% |
| Expected Shortfall at 95% | -9.5579% |
| Expected Shortfall at 99% | -9.5579% |

*Figura 41. Métricas de medidas de riesgo*

**Comparación del Portafolio Promedio y Benchmark a lo Largo del Tiempo**

La gráfica ilustra el progreso del valor de un portafolio promedio, resultante de 100 simulaciones distintas, en comparación con el índice de referencia S&P 500 a lo largo del tiempo. Se observa claramente que el portafolio muestra supera al benchmark de forma consistente, destacando especialmente hacia el final del período evaluado donde el valor del portafolio muestra un incremento notable.

Este crecimiento sostenido del portafolio muestra puede reflejar una selección estratégica de activos o la aplicación de una táctica de inversión que ha sabido aprovechar oportunamente las condiciones de mercado. El análisis sugiere que el modelo de inversión empleado ha sido capaz de gestionar los portafolios de tal manera que ha superado regularmente al mercado de referencia, demostrando la eficacia de la estrategia de inversión utilizada.

En resumen, el análisis basado en esta visualización sugiere que la estrategia de inversión, aplicada a lo largo del tiempo, no solo ha sido efectiva, sino que ha mostrado una capacidad para capturar y potencializar los momentos de crecimiento del mercado, lo cual se refleja en el incremento significativo en el valor del portafolio hacia el final del período analizado.

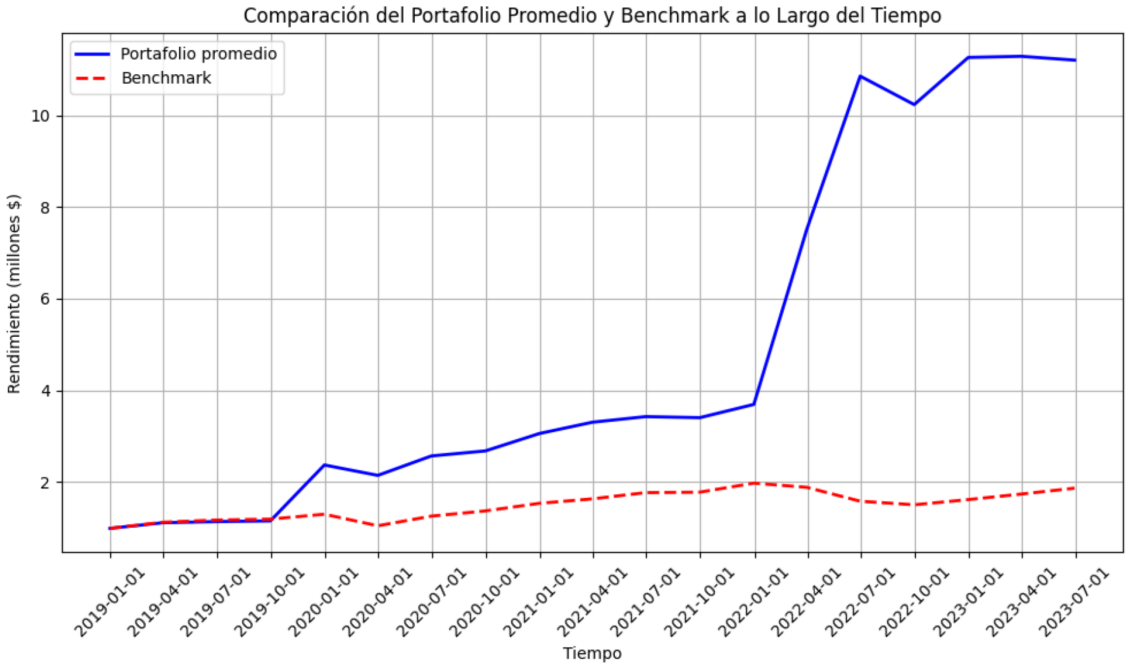


*Figura 42. Gráfica de la evolución del dinero en el tiempo*

**Análisis de la Distribución de Rendimientos en Efectivo al Final de las Simulaciones de Montecarlo mediante Boxplot**

El diagrama de caja y bigotes presentado muestra la distribución de los rendimientos trimestrales de un portafolio de inversiones. La mediana de los rendimientos, indicada por la línea roja central de la caja, se encuentra aproximadamente en 0.1. Esto sugiere que el 50% de los rendimientos están por encima de este nivel y el 50% por debajo, lo que representa un nivel medio de rendimiento que la mayoría de las simulaciones alcanzan o superan.

La caja, que se extiende desde el primer cuartil hasta el tercer cuartil, cubre un rango de rendimientos desde aproximadamente 0.0 hasta 0.2. Este rango intercuartílico muestra que los rendimientos del portafolio se concentran en un espectro relativamente estrecho, indicando una variabilidad moderada. La altura de la caja refleja una distribución compacta de los datos, lo cual es indicativo de que las estrategias de inversión del portafolio generan resultados consistentes y predecibles en el período observado.

Los bigotes del diagrama se extienden desde cerca de -0.05 hasta aproximadamente 0.2, señalando los valores mínimos y máximos no atípicos dentro de los rendimientos. Esto demuestra que, mientras la mayoría de los resultados de inversión se mantienen dentro de un rango más controlado, hay posibilidades de alcanzar rendimientos ligeramente más bajos o más altos, dependiendo de las condiciones del mercado y las decisiones de inversión tomadas.

En la gráfica también se observan un par de valores atípicos, situados alrededor de 0.5 y 1.0, respectivamente. Estos puntos destacan rendimientos excepcionalmente altos que no siguen la norma de distribución general del portafolio. La presencia de estos valores atípicos sugiere que existen períodos en los cuales el portafolio es capaz de generar rendimientos significativamente superiores, posiblemente debido a condiciones de mercado favorables o decisiones de inversión particularmente acertadas.

En resumen, la interpretación del diagrama de caja y bigotes para este portafolio indica que, generalmente, los rendimientos son moderados y están concentrados dentro de un rango específico, con episodios esporádicos de rendimientos altamente positivos. Este perfil podría ser atractivo para inversores que buscan una combinación de estabilidad y la oportunidad de capturar ganancias superiores en momentos de alta rentabilidad en el mercado.

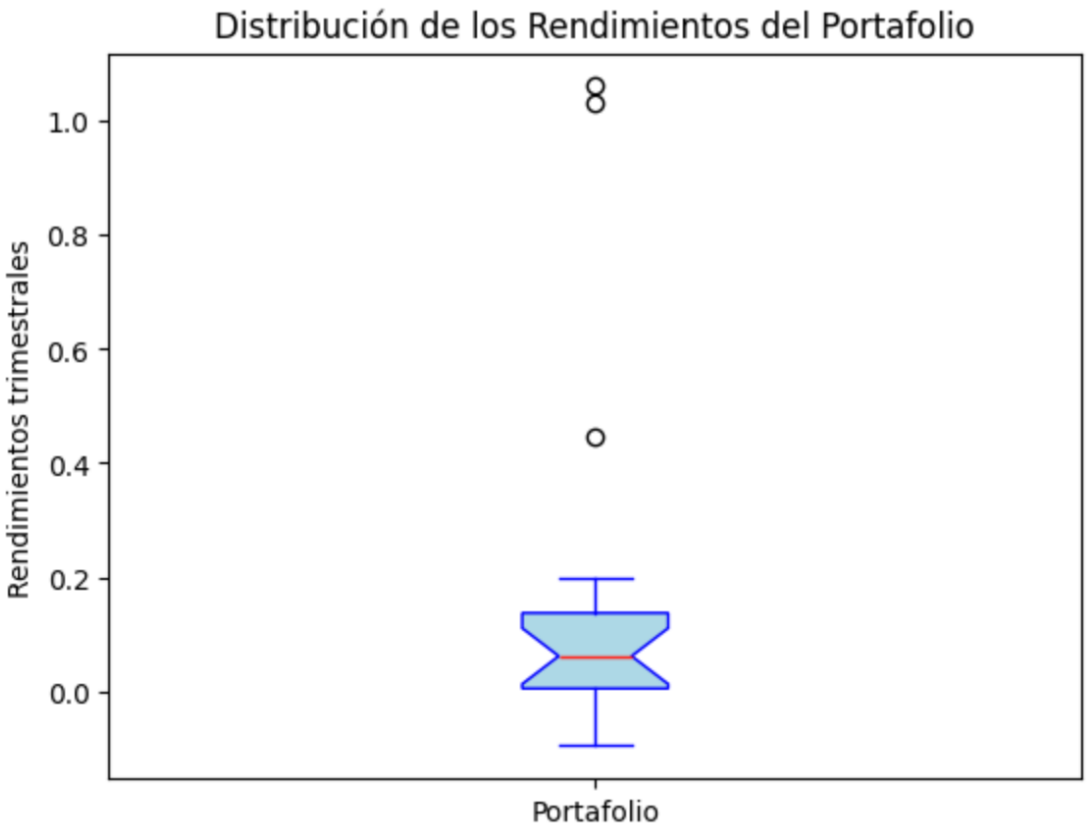


*Figura 43. Gráfica Boxplot de Distribución de Rendimientos en Efectivo.*

## 3. Resultados del trabajo profesional

Este proyecto se enfocó en desarrollar un algoritmo de programación diseñado para descargar y analizar los estados financieros trimestrales de todas las empresas que pertenecen al índice S&P 500. El objetivo principal consistió en emplear estos estados financieros para calcular diversos ratios financieros con el propósito de predecir el rendimiento futuro de las empresas en el próximo trimestre.

El algoritmo clasifica las empresas en función de si se espera un rendimiento positivo o negativo, emitiendo recomendaciones de inversión según sea el caso. En el caso de un rendimiento positivo, el algoritmo devuelve un valor de 1, indicando la sugerencia de invertir; mientras que, en caso de un rendimiento negativo, devuelve un valor de 0, recomendando abstenerse de la inversión.

Para llegar al modelo final, se llevó a cabo un exhaustivo proceso de selección y evaluación de modelos de machine learning, considerando diversas técnicas con el objetivo de maximizar la precisión predictiva. Se exploraron métodos como Gradient Boosting Classifier, HGBC, Random Forest Classifier, XGBoost, Decision Tree Classifier, K-Nearest Neighbors, Support Vector Classifier, Multilayer Perceptron, Stochastic Gradient Descent, Logistic Regression, Ridge Classifier CV, y Linear Discriminant Analysis. Probar con varios modelos se obtuvieron varios enfoques permitiendo una evaluación integral de las capacidades predictivas de cada modelo en el contexto financiero.

Adicionalmente, se experimentó con diferentes estrategias de preprocesamiento de datos, incluyendo la estandarización y no estandarización de los datos, técnicas de reducción de la dimensionalidad (Análisis de Componentes Principales, o PCA; y Análisis Discriminante Lineal, o LDA), así como la aplicación de técnicas de ensamblado como bagging y stacking. Estas exploraciones se llevaron a cabo con el objetivo de identificar la configuración óptima para el conjunto de datos específico, considerando la complejidad y la variabilidad inherente a la información financiera.

El proceso iterativo de prueba y error, junto con la evaluación rigurosa de diferentes modelos y estrategias, finalmente condujo a la selección del modelo que demostró el mejor desempeño en la predicción del rendimiento financiero de las empresas del S&P 500. Este enfoque metódico y exhaustivo respalda la robustez del modelo final y su capacidad para abordar las complejidades inherentes a la predicción financiera en un entorno dinámico y cambiante.

Para evaluar la efectividad del algoritmo, se llevaron a cabo 100 simulaciones Montecarlo sobre la estrategia propuesta para el backtesting. Ello llevó a que se realizaran pruebas utilizando una diferente "condición inicial", recordando que el algoritmo da preferencia a las acciones que ya contiene el portafolio por temas de ahorro de comisiones. El algoritmo está basado en la filosofía de Value Investing, por lo mismo, el horizonte mínimo de inversión recomendado es de 1 año.

A partir de estas simulaciones, se calcula nuestro portafolio muestra, un promedio simple del valor del portafolio en las distintas simulaciones a lo largo del periodo de análisis. En los resultados se obtuvieron métricas de Rendimiento, Volatilidad, VAR (Value at Risk), Expected Shortfall, Alpha de Jensen, Treynor, Sortino, Information Ratio y Ratio de Sharpe positivas. Estas métricas respaldan la capacidad del algoritmo para generar recomendaciones de inversión efectivas, proporcionando una base sólida para la toma de decisiones financieras. La combinación de la descarga de datos, el análisis financiero y el posterior backtesting contribuyó al desarrollo de un enfoque integral y exitoso para la predicción del rendimiento financiero, con implicaciones significativas para la toma de decisiones estratégicas en el ámbito de las inversiones.

## 4. Reflexiones del alumno o alumnos sobre sus aprendizajes, las implicaciones éticas y los aportes sociales del proyecto

El RPAP tiene también como propósito documentar la reflexión sobre los aprendizajes en sus múltiples dimensiones, las implicaciones éticas y los aportes sociales del proyecto para compartir una comprensión crítica y amplia de las problemáticas en las que se intervino.

**Juan Carlos Gutiérrez Valdivia**: Durante el desarrollo de nuestro proyecto, tuvimos la oportunidad de desarrollar muchas competencias disciplinarias, sociales y universitarias. Respecto a las disciplinarias, fortalecimos habilidades como la responsabilidad, el compromiso, la amabilidad, entre otras que pusimos en práctica día con día; en la parte social, la colaboración y la comunicación constante fueron factores clave que nos ayudaron a trabajar en conjunto y optimizar las actividades, dependiendo de las habilidades de cada uno, y así poder obtener un buen resultado; finalmente, en el ámbito universitario, nos permitió gestionar mejor los tiempo y priorizar actividades para integrar la teoría junto con la práctica, adentrándonos más en la vida profesional diaria y demostrando nuestra capacidad para cumplir con las responsabilidades universitarias de manera exitosa.

Por último, los aprendizajes fueron muchos, profesionalmente adquirimos los conocimientos prácticos necesarios que van más allá de la parte teórica que durante la carrera obtenemos, enfrentándonos a un problema del mundo real y aplicando soluciones prácticas. Socialmente, mejoramos el trabajo en equipo adaptándonos en todo momento y valorando las atribuciones y/o aportaciones de cada integrante para gestionar el tiempo adecuadamente. Para concluir, a nivel personal aprendí a valorar mis aportaciones, mi autoeficacia y a enfrentar desafíos que al inicio parecían muy complicados, demostrando mi disposición a aprender y crecer en todo momento.

**Carlos Daniel Ponce Anguiano**: Desde la perspectiva universitaria, este proyecto nos proporcionó una valiosa lección en la gestión del tiempo y en la integración de la teoría con la práctica. Nos enfrentamos a retos que simularon escenarios reales, lo que nos preparó

mejor para nuestras futuras carreras profesionales. Aprendimos a equilibrar nuestras responsabilidades académicas y laborales con las demandas del proyecto, una habilidad esencial para nuestro desarrollo profesional.

A nivel profesional, cada uno de nosotros ganó una comprensión más profunda de cómo aplicar conocimientos teóricos en escenarios prácticos. Aprendimos a aplicar soluciones creativas y a adaptarnos rápidamente a los cambios. Socialmente, fortalecimos nuestras habilidades de trabajo en equipo, aprendiendo a valorar y capitalizar las fortalezas individuales para el éxito colectivo.

Finalmente, a nivel personal, este proyecto fue una jornada de autodescubrimiento y crecimiento. Nos enfrentamos a desafíos que inicialmente parecían abrumadores, pero aprendimos a abordarlos con confianza y determinación. Descubrimos nuestra capacidad para adaptarnos, aprender y prosperar en ambientes desafiantes, lo que aumentó nuestra autoeficacia y nos preparó para futuros retos profesionales y personales.

## 5. Conclusiones

En nuestro recorrido por el proyecto, nos hemos enfrentado a la dura realidad de la cultura del ahorro y la educación financiera en México. A través de interacciones directas con nuestro entorno, hemos visto de primera mano cómo la falta de conocimiento financiero afecta las oportunidades a largo plazo de las personas.

Desde la perspectiva de nuestra futura profesión, la experiencia en el PAP subraya la urgencia de abordar estas cuestiones de educación financiera. Estamos conscientes de que cada paso que damos hacia la promoción de la educación financiera puede tener un impacto transformador en las comunidades con las que trabajamos.

Este entendimiento nos ha empujado a buscar soluciones que no solo sean viables, sino que también sean éticamente sólidas y socialmente responsables. A través de nuestro enfoque

colaborativo, aspiramos a dejar una huella que trascienda el ámbito académico y genere un cambio tangible en la sociedad como una forma más accesible para mejorar sus fondos de retiro en un futuro.

## 6. Bibliografía


- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

- Buffett, W. (1984). *The Superinvestors of Graham-and-Doddsville*. Hermes: The Columbia Business School Magazine.

- Graham, B., & Dodd, D. L. (1934). *Security Analysis*. McGraw-Hill.

- Graham, B. (1937). *The Interpretation of Financial Statements*. Harper & Brothers.

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

- Lo, A. W., & MacKinlay, A. C. (1999). *A Non-Random Walk Down Wall Street*. Princeton

- Rist, M., & Pizzica, A. J. (2015). *Financial Ratios for Executives: How to Assess Company Strength, Fix Problems, and Make Better Decisions*. Apress

# Anexos

## Glosario

Con el propósito de establecer una base sólida para comprender las generalidades y alcances del proyecto, se presenta a continuación una sección dedicada a la definición de conceptos fundamentales en el ámbito financiero y tecnológico. Estos conceptos servirán como cimientos para una comprensión más profunda de la implementación programática del modelo matemático y el análisis fundamental en la toma de decisiones de inversión. Es entonces que a continuación se definen los siguientes tecnicismos:

- Análisis Fundamental: Representa un enfoque exhaustivo que busca calcular el valor real de un título financiero a través de la evaluación detallada de los balances y estados financieros de una entidad. Este método compara este valor calculado con el precio actual de mercado del activo, con el fin de identificar posibles situaciones de infravaloración en un momento dado. Es importante tener en cuenta que tanto los resultados financieros como los precios experimentan cambios constantes, lo que implica que la cotización de un activo puede ser influenciada por una diversidad de factores, incluyendo los de naturaleza política y económica. El análisis fundamental se erige como una herramienta de gran relevancia para evaluar el riesgo financiero asociado a empresas, haciendo uso de un análisis minucioso del entorno en el que operan, la determinación de ratios financieros clave y la valoración intrínseca de las propias compañías. Es importante destacar que este enfoque se diferencia del análisis técnico, que se centra principalmente en el estudio de tendencias de mercado y gráficos bursátiles.

- Activo: En el ámbito bursátil, se refiere a un instrumento que representa un valor económico y puede ser objeto de negociación en los mercados financieros. Estos activos representan derechos de propiedad o de deuda sobre una entidad, como una empresa, gobierno u otra entidad emisora. Los activos financieros pueden variar en naturaleza y

características, incluyendo acciones, bonos, opciones, futuros, divisas y otros instrumentos financieros.

- Acción: Representa una parte de la propiedad en la empresa emisora y otorga al titular el derecho a participar en las decisiones de la empresa y recibir una parte proporcional de las ganancias en forma de dividendos.

- Índices bursátiles: medidas estadísticas que reflejan el rendimiento general de un grupo de activos financieros, como acciones o bonos, en un mercado financiero específico. Calculados mediante fórmulas ponderadas, estos brindan una visión general del comportamiento del mercado y ayudan a los inversores a comprender cómo se está desempeñando un grupo de acciones en relación con un período anterior o un mercado más amplio. Ejemplos incluyen el S&P 500 y el Dow Jones en EE. UU.
- Liquidez: Se refiere a la facilidad con la que un activo financiero puede ser comprado o vendido en el mercado sin afectar significativamente su precio. Los activos líquidos son aquellos que pueden ser convertidos rápidamente en efectivo sin una disminución sustancial en su valor.

- Riesgo de mercado: Se define como la posibilidad de pérdidas en el valor de inversiones debido a cambios generales en los mercados financieros, como fluctuaciones en precios de activos, tasas de interés o tipos de cambio. Es un riesgo no diversificable y se vincula a factores macroeconómicos que afectan ampliamente a las inversiones, independientemente de su naturaleza. Este riesgo es una preocupación inherente para los inversores, ya que los movimientos en los precios de los activos financieros pueden impactar negativamente el valor de una inversión, incluso si los activos en cuestión tienen fundamentos sólidos. El riesgo de mercado puede influir en activos de diferentes clases y sectores, lo que subraya la importancia de gestionarlo de manera efectiva. Los inversores abordan este riesgo mediante estrategias de asignación de activos,

diversificación y, en algunos casos, mediante el uso de instrumentos financieros de cobertura, como futuros o opciones.

- Sentimiento de mercado: Actitud general o percepción emocional de los inversores y participantes de mercados financieros. Puede ser positivo, negativo o neutral, y a menudo influye en las decisiones de compra y venta de los inversores. Se basa en factores como noticias, tendencias económicas y opiniones públicas, y puede tener un impacto en la volatilidad de los precios y en las tendencias del mercado.

- Ciclo económico: Es la fluctuación periódica de la actividad económica en una economía, que incluye fases de expansión y contracción. Las fases incluyen el auge, la recesión, la depresión y la recuperación, y están influenciadas por factores económicos y políticos.

- Parsimonía: En el contexto del análisis y modelado, alude al principio de que un modelo simple que explica los datos de manera efectiva es preferible a un modelo más complejo si no proporciona una mejora sustancial en la comprensión o la predicción.

- Aprendizaje Supervisado: Es un enfoque dentro del campo del machine learning donde se entrena un modelo utilizando un conjunto de datos etiquetados, es decir, datos donde la respuesta o el resultado deseado está previamente conocido. El modelo utiliza esta información para aprender a hacer predicciones o tomar decisiones sobre nuevos datos no vistos previamente. En el contexto financiero, esto podría implicar la creación de un modelo que aprende a hacer recomendaciones de inversión al comparar los datos históricos de activos financieros con sus resultados reales de rendimiento.

- EDA (Exploratory Data Analysis)/ Análisis Exploratorio de Datos: Implica examinar y visualizar un conjunto de datos para obtener una comprensión inicial de su estructura,

patrones, distribuciones y posibles relaciones entre variables. Este enfoque inicial permite tomar decisiones informadas sobre el preprocesamiento de datos y la elección de técnicas de modelado adecuadas.

- Algoritmo K-Nearest Neighbors (KNN): El algoritmo K-Nearest Neighbors (KNN) es un método de aprendizaje automático supervisado utilizado para la clasificación y la regresión. En KNN, se asigna una etiqueta a un punto de datos desconocido basándose en la mayoría de las etiquetas de sus vecinos más cercanos en un conjunto de datos de entrenamiento. La "K" en KNN representa el número de vecinos más cercanos que se toman en consideración al realizar la clasificación. En esencia, el algoritmo encuentra los puntos de datos más cercanos al punto de consulta y toma una decisión basada en la etiqueta más común entre esos vecinos, lo que lo convierte en un enfoque simple pero efectivo para problemas de clasificación y regresión en el aprendizaje automático.

- Matriz de Correlación por Método de Pearson: Es una medida estadística que se utiliza para evaluar la relación lineal entre dos variables cuantitativas continuas. Es importante destacar que el coeficiente de correlación de Pearson solo mide asociaciones lineales y puede no ser adecuado para capturar relaciones no lineales entre las variables. También es importante tener en cuenta que la correlación no implica causalidad, es decir, el hecho de que dos variables estén correlacionadas no significa necesariamente que una causa la otra. El coeficiente de correlación de Pearson puede tener valores en el rango de -1 a 1:
    - Si r es 1, significa que hay una correlación positiva perfecta, lo que indica que a medida que una variable aumenta, la otra también aumenta en una relación lineal perfecta.
    - Si r es -1, significa que hay una correlación negativa perfecta, lo que indica que a medida que una variable aumenta, la otra disminuye en una relación lineal perfecta.
    - Si r es 0, significa que no hay una relación lineal entre las dos variables.

- Matriz de correlación por método de Spearman: Es una herramienta estadística que evalúa la relación entre las variables en un conjunto de datos mediante el coeficiente de correlación de Spearman. Este método calcula la correlación entre las variables al considerar no solo las relaciones lineales, sino también las relaciones monótonas, lo que lo hace especialmente adecuado para datos ordinales o no paramétricos. La matriz resultante muestra las asociaciones entre las variables, lo que permite identificar patrones y dependencias entre ellas, contribuyendo así a comprender mejor la estructura subyacente de los datos y su potencial influencia en un análisis o modelo.

- Accuracy: Es una medida comúnmente utilizada en el campo del aprendizaje automático y la estadística para evaluar el rendimiento de un modelo de clasificación. Representa la proporción de predicciones correctas que un modelo hace en relación con el total de predicciones realizadas.

- Métrica de Precision (Precisión): La precisión mide la proporción de predicciones positivas que son verdaderamente positivas.

- Métrica de Recall (Exhaustividad o Sensibilidad): La exhaustividad mide la proporción de casos positivos que el modelo identifica correctamente. Una alta exhaustividad indica que el modelo es bueno para evitar falsos negativos.

- Métrica de F1-Score: El F1-Score es una métrica que combina precisión y exhaustividad en una sola medida. Un F1-Score alto indica un equilibrio entre precisión y exhaustividad.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

- Overfitting (Sobreajuste): es una situación en la que un modelo de machine learning se ajusta demasiado a los datos de entrenamiento y, como resultado, tiene un rendimiento deficiente en datos nuevos o de prueba. El modelo puede ser tan complejo que memoriza los datos de entrenamiento en lugar de aprender patrones generales, lo que lo hace ineficaz para hacer predicciones precisas en situaciones no vistas.

- Backtesting: El backtesting constituye una metodología esencial en el ámbito financiero y estadístico, diseñada para evaluar la efectividad y el rendimiento de estrategias de inversión, modelos predictivos o sistemas de trading. Esta técnica implica la aplicación retrospectiva de dichas estrategias o modelos sobre datos históricos relevantes, con el objetivo de recrear operaciones o inversiones pasadas. Esta práctica permite discernir, de manera objetiva y cuantitativa, el comportamiento que habría tenido una determinada estrategia o modelo en condiciones de mercado anteriores. Además de evaluar la eficacia general de las estrategias, el backtesting revisado incluye la estimación de la generación de utilidades o pérdidas (P&L) que se habrían obtenido con la estrategia o modelo. Esto implica simular los resultados financieros bajo diferentes escenarios y calcular intervalos de confianza para las utilidades o pérdidas generadas. Estos intervalos de confianza proporcionan una medida de la variabilidad esperada en los resultados, ofreciendo una perspectiva más amplia sobre el riesgo y la incertidumbre asociados con las estrategias evaluadas.

## Productos

1. **Base de Datos Financiera:**

Base de datos que contiene información financiera detallada de compañías listadas en el S&P 500, recopilada durante un período de 5 años. Incluye un análisis en profundidad de los datos financieros para identificar tendencias y patrones.
Es una herramienta para la toma de decisiones informadas en inversiones, estudio de mercados y análisis económico.

2. **Reporte de Operaciones de Simulaciones:**

Conjunto de reportes generados a partir de 100 simulaciones ejecutadas utilizando la estrategia de inversión propuesta. Proporciona una visión detallada del rendimiento y comportamiento esperado de la estrategia bajo diferentes condiciones de mercado.
Permite a los inversores y analistas evaluar la viabilidad de la estrategia de inversión propuesta.

3. **Modelo GBC (archivo .pkl):**

Un modelo de machine learning Gradient Boosting Classifier (GBC) optimizado con la técnica de Bagging, almacenado en un archivo .pkl para su uso y aplicación inmediata en análisis predictivos.
El archivo facilita la implementación rápida del modelo en aplicaciones de análisis financiero, sin la necesidad de reconstruir o reentrenar el modelo desde cero.

4. **Repositorio de Proyecto en GitHub:**

URL: https://github.com/carlos-p11/stock_picking/tree/main?tab=readme-ov-file#install-dependenencies

Este URL te redirige al repositorio de GitHub que alberga todos los archivos, código fuente y documentación asociados con el proyecto de modelo de valoración.

El propósito de esto es proporcionar un acceso centralizado para colaboradores y partes interesadas para revisar, descargar o contribuir al proyecto.