Universidad Politécnica
de Madrid

Escuela Técnica Superior de

Ingenieros Informáticos

Grado en Matemáticas e Informática

Trabajo Fin de Grado

Análisis de Datos para Predecir el Desempeño Bursátil de una Empresa

Autor: Juan Hernández Sánchez

Tutor(a): Antonio Jesús Díaz Honrubia

Madrid, junio 2022

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Matemáticas e Informática
Título: Análisis de Datos para Predecir el Desempeño Bursátil de una Empresa
Junio 2022

Autor: Juan Hernández Sánchez

Tutor:
Antonio Jesús Díaz Honrubia
Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

Existen actualmente tres categorías generales de inversión en bolsa: el análisis técnico, el análisis de sentimientos y el análisis fundamental. Las dos primeras categorías se enfocan en la predicción a corto plazo del precio de las acciones, lo cual explica el gran interés existente en ambos campos. La tercera categoría, el análisis fundamental, es la metodología preferida por los mejores inversores del mundo, aunque es menos atractiva, pues promete retornos después de décadas, no de días. A pesar de ser la metodología usada por inversores de la talla de Warren Buffett, el análisis fundamental es una corriente minoritaria y poco estudiada por la literatura académica. Por este motivo, esta metodología de inversión carece de un enfoque objetivo y sistemático que respalde las máximas de la disciplina. La mayoría de los estudios existentes que utilizan el aprendizaje automático y la minería de datos se centran en la creación de modelos de predicción basados en el análisis técnico y el análisis de sentimientos. Sin embargo, la desconexión de estos métodos con los negocios subyacentes y la falta de métodos de valoración de empresas puede afectar negativamente a sus conclusiones y a la efectividad de los modelos propuestos.

En este proyecto, se pretende recolectar, preparar y estudiar datos sobre el mercado de valores en los últimos 40 años; ofreciendo fundamentos objetivos desde la perspectiva del análisis fundamental. El objetivo principal de este estudio es extraer conclusiones respaldadas por los hechos que permitan mejorar los rendimientos de los inversores particulares. Posterior al estudio, se utilizarán los datos financieros de más de 2000 empresas, cada una con décadas de información; para el desarrollo de un modelo de aprendizaje automático que permita predecir el precio de las acciones en el medio y largo plazo, para así poder generar carteras de inversión que logren batir al mercado.

Durante el proyecto se proponen varios modelos de regresión usando XGBoost, que unidos a una cuidadosa selección de los datos fundamentales más importantes, son capaces de predecir sistemáticamente el precio medio de las acciones el año siguiente. Los modelos no solo muestran resultados positivos utilizando las habituales métricas de regresión. Se ha elaborado un sistema de *backtesting* que permite simular la construcción de carteras de inversión basadas en las recomendaciones de los modelos en años anteriores. Esto permite utilizar el rendimiento de las carteras construidas como otra métrica de evaluación. Durante la etapa estudiada todos los modelos propuestos fueron capaces de batir al índice de referencia, generando rentabilidades muy superiores al mercado y al universo de acciones estudiado.

Abstract

There are currently three general categories of stock market investing: technical analysis, sentiment analysis, and fundamental analysis. The first two categories focus on short-term stock price forecasting, which explains the great interest in both fields. The third category, fundamental analysis, is the methodology preferred by the best investors in the world, although it is less attractive, since it promises returns after decades, not days. Despite being the methodology used by investors of the stature of Warren Buffett, fundamental analysis is a minority trend and little studied in academic literature. For this reason, this investment methodology lacks an objective and systematic approach that supports the maxims of the discipline. Most of the existing studies using machine learning and data mining focus on building prediction models based on technical analysis and sentiment analysis. However, the disconnection of these methods with the underlying businesses and the lack of company valuation methods may negatively affect their conclusions and the effectiveness of the proposed models.

In this project, it is intended to collect, prepare, and study data on the stock market in the last 40 years; offering objective foundations from a fundamental analysis perspective. The main objective of this study is to draw conclusions supported by the facts that allow improving the returns of individual investors. After the study, the financial data of more than 2000 companies will be used, each with decades of information; for the development of an automatic learning model that allows predicting the price of shares in the medium and long term, to generate investment portfolios that manage to beat the market.

During the project, several regression models are proposed using XGBoost, which, together with a careful selection of the most important fundamental data, are capable of systematically predicting the average share price for the following year. The models don't just show positive results using the usual regression metrics. A backtesting system has been developed that allows simulating the construction of investment portfolios based on recommendation models in previous years. This allows the performance of the constructed portfolios to be used as another evaluation metric. During the studied stage, all the proposed models were able to beat the reference index, generating returns well above the market and the universe of shares studied.

Índice General

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción y objetivos

Durante el primer capítulo se realizará una exposición de la motivación para dicho trabajo, así como un breve resumen de la metodología, trabajos anteriores y una lista de objetivos a lograr.

1.1 Motivación

La inversión en renta variable es una de las áreas que más interés suscita tanto a inversores individuales como a inversores institucionales. La promesa de conseguir altas rentabilidades con un mínimo esfuerzo es el aliciente perfecto para animar a las personas a participar de uno de los sistemas más complejos y arriesgados que existen en la actualidad. Tanto los inversores individuales sin formación, como los inversores institucionales con décadas de experiencia en materia de inversiones, palidecen al intentar adivinar qué rumbo tomará la bolsa, llevando a la idea popular de que "la bolsa es un casino". Son muchos los factores que contribuyen al desempeño a largo plazo de un inversor, entre ellos factores psicológicos como el efecto Dunning–Kruger, la aversión a la pérdida u otros sesgos cognitivos inherentes al ser humano [4]. Es así como con el paso de los años, debido a la incertidumbre y la inmensa complejidad del mercado, se han creado diferentes escuelas de inversión muy diferentes entre ellas. Sin embargo, las personas que se inicien en la inversión en bolsa hoy en día, lejos de encontrar sistemas probados y contrastados, experimentaran algo distinto: *la regresión del logos al mito.*

Actualmente la ciencia detrás de la inversión se entremezcla con anecdotarios de inversores famosos y experiencias personales, bañados ambos por un aura de misterio y medias verdades. Es en este contexto donde parece necesario examinar detenidamente las realidades que subyacen a la inversión. La intención principal es realizar un análisis de los factores que contribuyen al desempeño bursátil de una empresa en el largo plazo, basándonos en los datos fundamentales del negocio. Esta proyecto también está concebido como una forma de aportar al pequeño inversor datos fiables sobre los que poder construir conocimiento, es decir, *para apartar los mitos y encontrar los hechos*.

Adicionalmente, el área del aprendizaje automático bajo la perspectiva del análisis fundamental permanece relativamente inexplorado [1] [2] [3], por lo que se desea profundizar en este.

1.2 Trabajo previo

Antes de comenzar este trabajo ha sido necesaria una basta labor de investigación en el análisis de datos y sus fundamentos matemáticos, así como su metodología, sus limitaciones y sus capacidades. Se ha requerido también una revisión de trabajos anteriores y el desarrollo de un "conocimiento de

negocio" sobre la inversión en renta variable; y por consiguiente en economía y psicología.

La mayoría de los trabajos existentes que utilizan aprendizaje automático se centran en el análisis técnico y el análisis de sentimientos [1] [2] [3]. El análisis de sentimientos se basa en el análisis de noticias, publicaciones o informes sobre empresas o el mercado, para averiguar la actitud general de los inversores respecto a una empresa o acción. Por otro lado, el análisis técnico toma como entrada solo los precios históricos y el volumen de compra. El análisis técnico se rige por la máxima de que toda la información pública de la empresa está ya reflejada en el precio de mercado; por lo que solo hace falta estudiar el precio de cotización. A esta, le siguen dos premisas más: que el precio sigue tendencias que se pueden predecir y que la historia se repite. Estos enfoques son los preferido por la literatura académica debido a la promesa de retornos rápidos en el corto plazo, y han conseguido un cierto grado de éxito [5] [6].

Las acciones son, en esencia, pequeñas partes de una empresa, sin embargo, ambos enfoques anteriores parecen ignorar este hecho, despreciando el estudio del negocio. Es el análisis fundamental de acciones la única disciplina que contempla el estudio de la empresa para la previsión de los precios de cotización. Son más escasos los trabajos usando análisis fundamental [7] [8] [9] y suelen verse limitados por la falta de datos financieros. Los enfoques más destacados en este campo han sido la conversión del problema de regresión a un problema de clasificación [7], el uso de "*Feedforward neural networks*" (FNN por sus siglas en inglés) [10] para la predicción de tendencias, la mezcla del análisis fundamental con los otros tipos [10] [11] y la solución del problema de regresión con posterior ordenación en forma de ranking para la selección de acciones [8]. Este último es el enfoque escogido para el presente trabajo.

1.3 Objetivos

En base a todo lo anterior el objetivo principal de este trabajo es encontrar hechos comprobables y respaldados por los datos, sobre los cuales sea posible construir una metodología de inversión en bolsa.

Para lograr este objetivo principal se proponen los siguientes objetivos específicos:

- Aplicar metodología CRISP-DM para el estudio de los datos fundamentales de las empresas cotizadas en bolsa.
- Realizar modelos predictivos que ayuden a la construcción de una cartera de inversiones.
- Confeccionar una herramienta que haga uso de las conclusiones extraídas para ayudar a un inversor a tomar mejores decisiones.
- Discutir los resultados obtenidos, evaluando el procedimiento seguido, así como las conclusiones obtenidas.

1.4 Planificación del proyecto

El proyecto se realizará en diferentes etapas de acuerdo con la metodología CRISP-DM para *Knowledge Discovery*, una metodología comúnmente utilizada en análisis de datos. El proyecto se dividirá en las siguientes tareas:

1. **Comprensión del negocio:** Estudio de la inversión en bolsa y de otros conceptos necesarios relacionados con el mundo de la inversión.

2. **Comprensión de los datos:** Recolección, descripción y exploración de los datos iniciales.

3. **Preparación de los datos:** Selección, limpieza y formateo de los datos para su posterior análisis o modelado.

4. **Modelado:** Selección de técnicas de modelado y construcción de los modelos.

5. **Evaluación:** Evaluación del proceso seguido y de la validez de los modelos creados.

6. **Implementación:** Transformar el conocimiento adquirido en herramientas o guías que ayuden a tomar decisiones de negocio, en este caso, que ayuden a la inversión en bolsa.

Sobre la metodología de trabajo se desarrollará más en el capítulo 3.

1.5 Estructura de la memoria

A continuación, se resumirá brevemente cada de los capítulos de este trabajo:

1) Introducción y objetivos. En este primer capítulo se expone la motivación para el trabajo, así como los objetivos a lograr. También se hace un breve repaso de los trabajos previos sobre la cuestión y se describe la metodología a usar.

2) Alcance del proyecto. En el segundo capítulo se exponen las pretensiones de alcance del proyecto, marcando los límites de este.

3) Metodología. En el tercer capítulo se desarrolla en qué consiste la metodología utilizada y las fases necesarias para llevarla a cabo.

4) Desarrollo del proyecto. En el cuarto capítulo se procede a exponer paso a paso lo realizado en el proyecto, desde la etapa de recolección de datos hasta el despliegue de la aplicación.

5) Conclusiones. Este capítulo pretende ser un breve resumen de las principales conclusiones a las que se han llegado elaborando el proyecto, reflexionando sobre el cumplimiento de los objetivos de negocio y planteando mejoras para trabajos futuros.

6) Análisis de impacto. Se procede a evaluar el impacto de este trabajo en relación con los objetivos de desarrollo sostenible.

Capítulo 2

Alcance del proyecto

Debido a las limitaciones temporales no es posible tratar todas las fases de un proyecto de ciencia de datos con la profundidad deseable. Además, el tema de este trabajo es lo suficientemente amplio para que sobre él corran ríos de tinta, por lo que se debe limitar el alcance del proyecto.

2.1 Alcance del proyecto

Ya se ha comentado que la bolsa es un mercado de gran complejidad afectado por una innumerable cantidad de factores como las noticias, los resultados de las empresas, el sentimiento de los inversores o las expectativas a futuro. En palabras de Benjamin Graham [12]:

"A corto plazo el mercado de acciones se comporta como una máquina de votar, pero a largo plazo actúa como una máquina de pesar"

Con estas palabras queda reflejada la incertidumbre de la bolsa en el corto plazo, pues es donde más pesa la psicología de mercado y donde los ciclos eufórico-depresivos son más tangibles. Está es una intuición del padre de la inversión en valor[1] que también ha sido compartida por otros grandes economistas como Keynes y sus "espíritus animales" [13] o Akerlof y Shiller con sus propuestas de incorporar factores psicológicos humanos a los modelos macroeconómicos [14]. A día de hoy ya existe evidencia empírica de la presencia de un factor psicológico en la bolsa [15], lo cual complica sustancialmente la predicción de los precios a corto plazo basándose únicamente en datos financieros. Como ya se ha tratado, existen intentos de predecir el precio de cotización de acciones usando herramientas como el análisis de sentimiento o el análisis técnico, que intentan precisamente hacer uso de la psicología humana o de los datos históricos de precios. Estos enfoques se encuentran ampliamente estudiados y presentan una desconexión del análisis de la acción con el negocio real, por lo que no serán los enfoques del proyecto.

Podría pensarse entonces en intentar predecir la bolsa a largo plazo, sin embargo, aquí también existen dos limitaciones fatales. Primero los "Cisnes negros", nombre que se popularizó gracias a al libro de Nassim Nicholas Taleb [16]; y segundo, el hecho de que la bolsa es un sistema caótico.

Los cisnes negros son eventos impredecibles, con un alto impacto y que tienden a racionalizarse una vez han ocurrido (dando la falsa idea de que podrían haberse predicho). Cualquier estimación a largo plazo está condenada a encontrarse con un cisne negro que arruinará el modelo (¿cuántos modelos tuvieron en cuenta la crisis económica del 2008 o la pandemia mundial de 2020?). Por otro lado, se tiene el problema básico de la acumulación de errores. Cuanto más adelante se intenta predecir más impacto tienen los errores

[1] "*Value investing*" en inglés.

cometidos en los supuestos iniciales, pues los errores se magnifican a cada año y esto puede llevar a que diferencias minúsculas en la entrada de un modelo haga estimaciones completamente diferentes en el largo plazo.

Existe un tercer argumento en contra de la predicción de la bolsa a largo plazo, la hipótesis de los mercados eficientes. Según esta teoría, en cualquier momento dado existen millones de personas buscando información privilegiada que les permita predecir el precio futuro de las acciones. Estos agentes económicos, compran a precio bajos y venden a precios altos; el resultado es que esta información es rápidamente incorporada al mercado y los precios ajustados en consecuencia. La formulación semi-fuerte es un axioma para el análisis técnico, sin embargo, la evidencia empírica de esta hipótesis no respalda ninguna formulación fuerte, además de que existen disputas tanto a nivel empírico como teórico sobre la validez de esta [17] [18] [19] [20]. Este es uno de los motivos por los que durante este proyecto no se utilizará análisis técnico, si no que se optará por un análisis desde la perspectiva fundamental.

¿Qué se debe hacer pues si no se puede predecir exactamente el rumbo que tomará la bolsa? Que no se pueda predecir con exactitud no significa que no merezca la pena estudiar los factores que afectan a su comportamiento. Véase el ejemplo de inversores como Joel Greenblat con su "fórmula mágica" [21] o el profesor Joseph D. Piotroski [22] quienes han conseguido elaborar estrategias sencillas y basadas en los datos fundamentales, capaces de batir al mercado. Estas estrategias de inversión no vienen marcadas por modelos estocásticos complejos ni modelos de series temporales de última generación, sino por un conocimiento profundo de las empresas y de los factores relevantes en sus estados financieros. Es decir, su superior conocimiento de la importancia de los datos fundamentales de los negocios se traduce en superiores retornos en bolsa.

Por esta razón el alcance del proyecto entonces queda limitado al análisis de los factores fundamentales que contribuyen al desempeño bursátil a medio y largo plazo, ignorando el análisis técnico y de sentimientos. Se buscará hacer uso del conocimiento adquirido para intentar mejorar las rentabilidades del inversor individual. Aunque vayan a implementarse modelos predictivos, no es el objetivo predecir la bolsa con exactitud, pues aquello es un trabajo muy complejo y a menudo estéril, que además se sale del alcance de este proyecto. Basta con ser capaces de predecir, *grosso modo,* su precio futuro y diseñar portfolios con las mejores ideas disponibles a cada momento.

Debido a la enorme complejidad del sistema, se vuelve necesario enumerar algunas consideraciones previas que han de tenerse en cuenta, entre las que se cuentan: los sesgos del autor, las premisas utilizadas, las limitaciones de la información y las decisiones que se han tomado de manera arbitraria; aunque estas se irán exponiendo a lo largo del desarrollo.

Capítulo 3

Metodología

Uno de los principales problemas que debe afrontar un analista de datos es el de identificar información válida y relevante en grandes volúmenes de datos, utilizando lo importante y descartando lo superfluo. En palabras de Vijay Kotu y Bala Deshpande: *"El valor de los datos almacenados es cero a menos que se actúe en consecuencia."* [23]. Para conseguir este objetivo, se harán uso de metodologías ya establecidas para el análisis de datos.

3.1 Knowledge discovery in databases

Con el objetivo de poder operar con volúmenes tan bastos y abrumadores de información existe "*Knowledge discovery in databases*" (KDD), un proceso organizado que ayuda a encontrar patrones válidos, novedosos o potencialmente útiles dentro de un gran conjunto de datos, enfocado a tomar decisiones importantes.

KDD es un proceso iterativo, donde no solo es posible, sino que es habitual volver a pasos anteriores. Existen variantes sobre el proceso, pero en esencia todas llevan a cabo el mismo recorrido. Una formulación de los pasos del proceso es [24]:

1. **Comprender los dominios de aplicación involucrados y el conocimiento que se requiere para la tarea.**
   Es necesario entender el dominio del negocio para tomar decisiones sobre representación, algoritmos, transformaciones, etc. Este conocimiento es necesario para comprender y especificar los objetivos.

2. **Seleccionar el conjunto de datos sobre el que se llevará a cabo el descubrimiento.**
   Habiendo definido los objetivos, se ha de seleccionar los datos que se usarán en el proceso; usualmente teniendo que crear de cero el conjunto de datos a partir de varias fuentes.

3. **Limpiar y preprocesar los datos.**
   Existen infinidad de mecanismos para limpiar los datos y mejorar la calidad de estos, requisito crucial para un análisis exitoso. Es en este paso es cuando debe lidiarse con los datos faltantes, erróneos y los valores atípicos.

4. **Transformar los datos.**
   Desde simplificar los conjuntos de datos eliminando variables indeseadas hasta aplicar métodos de reducción de dimensionalidad. Este paso depende mucho del proyecto en específico que se está intentando realizar, pero el objetivo es el de preparar los datos de acuerdo con el problema a resolver.

5. **Elegir objetivos adecuados a los que aplicar técnicas de minería de datos.**
   En este paso se debe elegir qué tipo de problema se quiere solucionar (problema de predicción, de asociación, …). Esto depende principalmente de los objetivos definidos en el paso 1.

6. **Elegir técnicas de minería de datos adecuadas para descubrir patrones ocultos.**
   Una vez escogida la tarea, se debe seleccionar entre todos los métodos de minería de datos disponibles aquel que se adapta a nuestras necesidades, siendo conscientes de las ventajas y desventajas de cada uno.

7. **Utilizar la técnica de minería escogida.**
   Este paso se refiere a la implementación y uso del método seleccionado. En este paso será necesario ejecutar el algoritmo varias veces cambiando los parámetros hasta encontrar una solución satisfactoria.

8. **Evaluar los resultados.**
   Una vez ejecutados los algoritmos se deben evaluar e interpretar los resultados, dictaminando la utilidad de los modelos y si es necesario modificarlos o incluso desecharlos.

9. **Usar el conocimiento descubierto.**
   De nada sirve descubrir patrones y elaborar modelos si no se lleva a la práctica. Esta etapa consiste en utilizar el conocimiento adquirido en otro sistema para conseguir un impacto.

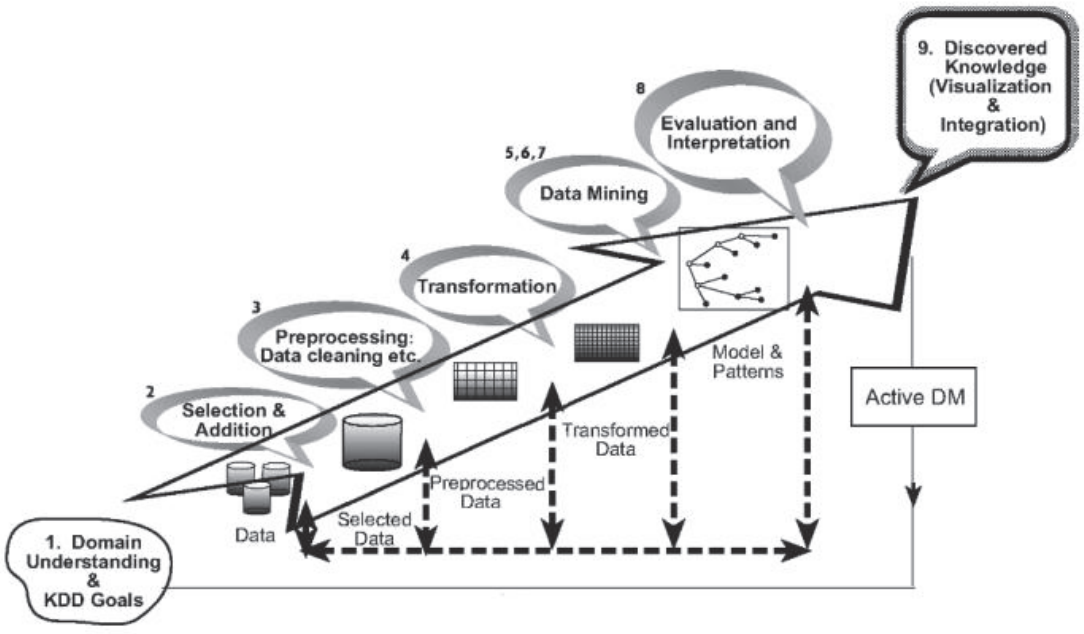

Figura 1: Proceso de Knowledge Discovery in Databases [24]

3.2 CRISP-DM

***CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining* (CRISP-DM) es un proceso de seis fases que describe el ciclo de vida de un proyecto en ciencia de datos. Es decir, se trata un proceso organizativo que pretende ayudar a completar de manera ordenada un proyecto. Se puede resumir la metodología CRISP-DM como un proceso iterativo y bidireccional.

Las seis etapas en las que se divide son:

1. **Comprensión del negocio**
   Se centra en comprender los objetivos del proyecto, así como los requerimientos desde una perspectiva de negocio. Posteriormente traduce el problema de negocio a un problema de minería de datos.

   - Determinar objetivos de negocio
   - Valorar la situación
   - Determinar objetivos de minería de datos
   - Producir planes de proyecto

2. **Comprensión de los datos**
   Empieza con la recolección de datos, así como la elaboración de un conjunto de datos sobre los que trabajar. También incluye la familiarización con los datos para identificar problemas en su calidad o formular hipótesis sobre los mismos.

   - Recolectar los datos iniciales
   - Describir los datos
   - Explorar los datos
   - Verificar la calidad de los datos

3. **Preparación de los datos**
   En esta etapa se incluye todo el proceso de conversión de los datos sin procesar hasta la creación de un conjunto de datos limpio al final. Aquí se encuentran la limpieza de datos, la selección de variable o la aplicación de diferentes técnicas como las de reducción de dimensionalidad.

   - Seleccionar datos
   - Limpiar datos
   - Construir datos
   - Integrar datos
   - Formatear los datos

4. **Modelado**
   En esta fase se seleccionan e implementan diferentes modelos, eligiendo las técnicas en función del objetivo a cumplir. Además, durante esta fase se calibran los parámetros de los modelos para encontrar los valores óptimos para cada problema.

   - Seleccionar las técnicas de modelado
   - Generar el diseño para las pruebas

- Construir los modelos
- Valorar los modelos

5. **Evaluación**
   Se revisan de manera detenida los modelos, verificando si son válidos y si cumplen con los objetivos y requisitos de negocio.

   - Evaluar los resultados
   - Revisar el proceso
   - Determinar los siguientes pasos

6. **Despliegue**
   El paso final es la implementación de los modelos o el conocimiento adquirido en otros sistemas, otorgando valor real al proceso.

   - Planear el despliegue
   - Planear la monitorización y el mantenimiento
   - Realizar un reporte final
   - Revisar el proyecto

Una vez visto KDD, CRISP-DM resultará familiar, esto es debido a que se podría considerarse a este último como una implementación del primero [25]. Por tanto, CRISP-DM será la metodología seguida en este trabajo, pues proporciona unas guías sólidas sobre cómo proceder de manera ordenada a la hora de realizar un proyecto de ciencia de datos.

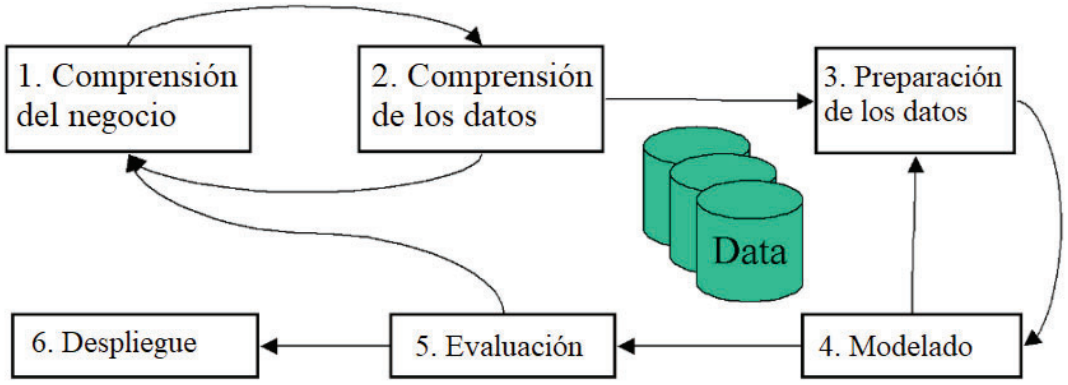

Figura 2: El ciclo de CRISP-DM [24]

Capítulo 4

Desarrollo del Proyecto

En las siguientes secciones se presentan todas las fases de un proyecto de análisis de datos inspirado en la metodología CRISP-DM. Para un mejor estudio de los factores que contribuyen en el largo plazo al desempeño de las empresas en bolsa, se ha decidido realizar en análisis de los datos en un apartado independiente, posterior a la limpieza y preparación.

4.1 Herramientas utilizadas

El lenguaje de programación utilizado ha sido Python, haciendo uso de diferentes librerías para cada tarea. Las librerías utilizadas según su funcionalidad han sido:

- Obtención de datos: Selenium [26], Beautiful Soup [27], Yfinance [28]
- Concurrencia: Multithreading [29], Multiprocessing [30].
- Manejo de datos: Pandas [31], NumPy [32], SciPy [33]
- Visualización de datos: Matplotlib [34], Seaborn [35], Plotly [36]
- Machine Learning: Scikit-learn [37], XGBoost [38], Featurewiz [39]
- Frontend: Streamlit [40]

Otras herramientas utilizadas han sido Notepad++ y Excel, para la manipulación puntual de pequeñas cantidades de datos.

4.2 Comprensión de negocio

La comprensión del área en el que se está trabajando es de vital importancia para entender los datos en su contexto y alinear los objetivos de la investigación con los objetivos de los stakeholders. Por este motivo se procede a exponer algunos puntos importantes en lo que respecta a la inversión en bolsa.

El principal objetivo de los inversores es el incremento del capital. Para este objetivo son fundamentales dos componentes: *el retorno de la inversión y el riesgo de la inversión.*

Retorno sobre la inversión

El retorno sobre la inversión es la rentabilidad que se ha conseguido con la inversión. Si se compran las acciones de una empresa a un precio de 50€ la acción y se venden más tarde a 100€ la acción, el beneficio ha sido de 50€ y el retorno del 100% sobre la inversión inicial. Sin embargo, no es suficiente con considerar solo el retorno, pues está claro que puede ser preferible una inversión que retorne el 50% en un año, frente a una que retorne el 100% en un plazo de

cuarenta años. Esto da lugar a dos conceptos claves*, la preferencia temporal y el coste de oportunidad.*

Es preferible una inversión que reporte beneficios antes a otra que lo haga más tarde a igualdad de circunstancias. Esto se debe a la *preferencia temporal*, el ser humano preferirá antes 100€ entregados hoy que 100€ entregados dentro de un año. Supóngase entonces que se debe elegir entre 100€ entregados hoy o 200€ entregados el año que viene; es aquí donde diferentes personas darán diferentes respuestas, pues cada persona tiene una preferencia temporal distinta. Por este motivo debe tenerse en cuenta el retorno ajustado al tiempo, usando normalmente la métrica del *retorno anual de la inversión*.

Otro factor relevante es el concepto de *coste de oportunidad*. Para saber si el desempeño de una inversión ha sido bueno debe compararse con el resto de las opciones, pues siempre se busca el mejor rendimiento de entre todas las opciones. Esto implica que no hay un valor sobre el cual el retorno de una inversión es siempre bueno, pues depende del resto de opciones, es relativo. El desempeño de una empresa en bolsa puede ser considerado como excelente o nefasto según la coyuntura. Obtener un retorno del 10% anual pudiendo haber conseguido un retorno del 50% con igual riesgo, es un fracaso; de la misma manera, un retorno del 10% cuando el resto de las alternativas rinden un 5%, es un éxito. Debido a esto, es común comparar las rentabilidades de las carteras de valores frente a un índice de referencia o frente al "mercado" (normalmente refiriéndose al S&P500[2]).

Riesgo

Respecto a la definición de riesgo existe más controversia. En la Teoría de Porfolio Moderna (MPT[3] por sus siglas en inglés) se define como riesgo de mercado a la probabilidad de variaciones en el precio de una empresa [41], lo que comúnmente se conoce como volatilidad. Sin embargo, grandes inversores como Warren Buffet o Seth Klarman han expresado su disconformidad con la definición, pues para ellos el riesgo es la **pérdida permanente de capital** [42] [43]. Las críticas a la MPT dieron lugar a la Teoría de Porfolio Posmoderna (PMPT[4] por sus siglas en inglés), que hace mayor énfasis en el riesgo de pérdida. Para el estudio se tendrán en cuenta ambas definiciones de riesgo.

Valoración de acciones

Comprender cómo los inversores entienden el retorno y el riesgo es crucial si se quiere proporcionar información útil que mejore estos factores. Otra pieza fundamental para la inversión en bolsa que se debe comprender es la valoración de acciones.

A nivel teórico se dice que el valor intrínseco de un activo financiero son los flujos de caja futuros descontados a un tipo de interés adecuado. Es decir, el valor de una empresa es el dinero que generará en el futuro aplicando una penalización a los flujos de caja más lejanos en el tiempo. Esta penalización es

[2] El S&P500 es el principal índice de referencia de Estados Unidos y está compuesto por algunas de las 500 empresas más grandes del país.
[3] MPT hace referencia a Modern Portfolio Theory.
[4] PMPT hace referencia a Post-modern Portfolio Theory

el tipo de interés, que teóricamente se debe a factores como la preferencia temporal, el riesgo, la inflación o el coste de oportunidad. En resumen, el trabajo de un inversor consiste en averiguar cuáles van a ser los flujos de caja futuros de una empresa, descontarlos a un tipo de interés adecuado (teniendo en cuenta el riesgo, la inflación, etc.) obteniendo así el valor intrínseco de la empresa; y comprar el activo más barato que ese precio teórico.

Si el valor de un activo financiero viene determinado por sus flujo de caja futuros, esto hace que sea preferible, *ceteris paribus*, una empresa que vaya a durar en el tiempo (generando más flujos de caja) a otra que no. Es de aquí de donde nace la preocupación de la salud financiera de las empresas.

4.3 Recolección de datos

La recolección de datos es parte vital de todo proyecto de ciencia de datos. En algunos proyectos es posible comenzar con un dataset[5] ya existente sobre el cual añadir o eliminar características de acuerdo con el proyecto. Sin embargo, al comienzo de este proyecto no se disponía de ningún dataset y, por tanto, ha tenido que elaborarse desde cero.

El primer problema al realizar un proyecto relacionado con la inversión es la disponibilidad de los datos. Los datos financieros de calidad son un bien preciado y solo pueden encontrarse a cambio de una buena cuantía de dinero. Puesto que no se contaba con datos previos y se requiere de una gran cantidad de estos, se ha tenido que recolectar los datos de manera automática. Como ya se ha adelantado previamente, la recolección se ha realizado por medio de scripts de Python usando Selenium.

Se debe destacar la gran cantidad de datos a recolectar (del orden de siete u ocho millones), por lo que es necesario acortar los tiempos de recolección. El principal método para esto es le ejecución multihilo o multiproceso de scripts de larga duración. Los procesos multihilo y concurrentes añaden una capa adicional de complejidad a cambio de disminuir los tiempos de significativamente.

4.3.1 Obtención de la lista de empresas

Lo primero a conseguir es un listado de empresas sobre las cuales se van a recolectar datos. Toda empresa cotizada cuenta con un identificador único para el mercado en el que cotiza, este identificador se conoce como "*ticker*". El *ticker* consiste habitualmente en una serie de 1 a 5 caracteres alfanuméricos, siendo algunos ejemplo "AAPL" para Apple o "TSLA" para Tesla.

La página web que se usará para la obtención de la mayor parte de los datos es *roic.ai* [44]. Debido a que se quiere obtener la mayor cantidad de datos posibles se ha optado por hacer uso de todos los datos disponibles en la página web. Para obtener la lista de tickers a descargar se ha optado por descargar el *sitemap*

[5] Conjunto de datos

de la web y de ahí, extraer la lista de tickers disponibles en base a las direcciones web disponibles. Esto es posible debido a que, en la web, las consultas a empresas siguen el patrón:

https://roic.ai/financials/{TICKER}.

La decisión de recolectar información sobre la mayor cantidad de acciones viene dada por dos principales motivos.

1. Se desea utilizar la mayor cantidad de datos para tener una imagen más fiel de la realidad y del mercado en su conjunto.
2. Debido a que la información mostrada en la página es gratuita se espera que no sea de buena calidad y, por tanto, haga falta eliminar una gran cantidad de empresas para acabar con un dataset limpio.

Siguiendo este método se pudo extraer una lista de 7726 empresas, aunque a fecha de elaboración de este documento, se tiene constancia que la página dispone ahora de una mayor cantidad de información.

Cabe destacar que este método de recolección de datos sesga la muestra, pues es posible que los datos disponibles sean aquellos más deseados por la gente o los datos de las empresas más grandes y, por tanto, no correspondan a un reflejo fiel del mercado. Estos sesgos se tendrán en cuenta durante el análisis y las conclusiones.

4.3.2 Recolección de precios

La obtención de precios se llevó a cabo utilizando una página web diferente a la ya mencionada, debido a la no disponibilidad de los precios históricos de las empresas. La nueva página web utilizada es Yahoo Finance [45], una página comúnmente utilizada para propósitos de *web scrapping* en el mundo de las finanzas[6].

Esta web permite el acceso a un historial de precios de cotización de las acciones desde su inicio, además de aportar información sobre los dividendos y las divisiones de acciones. Los precios que se han recolectado comienzan en la fecha de inicio de la cotización de la acción y acaban el 27 de febrero de 2022. Cabe destacar que el último punto temporal es importante pues decide en gran medida la rentabilidad acumulada de las acciones a estudiar. Sin embargo, no existe ningún momento objetivamente mejor que otro para realizar esta comparación, por lo que se opta por el último dato recopilado.

La recolección de los datos se logra haciendo uso de Beautiful Soup, una librería de Python que permite acceder al HTML de la página y encontrar la tabla que se desea descargar, todo de una manera sencilla y sin realizar excesivas peticiones a la página.

[6] Con el objetivo de no presentar problemas al servicio que ofrece la página, se optó por limitar la velocidad de descarga simulando actividad normal. Esto ralentiza enormemente la obtención de datos, pero permite la recolección de estos sin presentar inconvenientes a la página.

Al recolectar datos de varias fuentes diferentes cabe la posibilidad de que existan acciones de empresas en una página que no estén disponibles en la otra. Sin embargo, esto no supuso un gran problema para la mayoría de las compañías.

En el proceso se descargan las tablas de precios históricos (descritas en detalle en el capítulo 4.4) y se guardan en ficheros CSV. Si no es posible encontrar los precios para el ticker en cuestión, se omite y no se genera el documento CSV que contiene su precio. En el caso de que la recolección de un precio falle de manera no prevista, se apunta el ticker en una lista de errores y se revisa posteriormente de manera manual.

Al final de la recolección se obtuvieron 7635 tickers con precios.

4.3.3 Recolección de datos básicos

Por datos básicos se hace referencia a la información sobre la empresa no representada en los estados financieros. Algunos ejemplos de estos datos son nombre, sector, industria, año de salida a bolsa (IPO por sus siglas en inglés), porcentaje de acciones que poseen los directivos, etc. Estos datos se recolectan de la primera web ya mencionada[7].

Debido al uso de JavaScript y a que los elementos que se quieren descargar no están ordenados, esta vez se realiza la recolección con el propio Selenium. Se accede a los elementos deseados por su XPath y se descargan. Los elementos descargados son ordenados en una tabla y guardados en un fichero CSV.

En caso de error se aplica la misma técnica que con el precio, se omiten las acciones para las cuales no se ha podido conseguir datos y se apuntan en una lista los fallos inesperados. Posteriormente se revisa manualmente la lista de fallos.

4.3.4 Recolección de los estados financieros

Los estados financieros se recolectan de manera diferente a ambos casos anteriores, haciendo uso de la misma web de la que se extraen los datos básicos. Se accede primero a la página correspondiente al ticker sobre el cual se quiere descargar información. Como en el caso anterior la presencia de JavaScript dificulta la descarga usando Beautiful Soup, pero esta vez existen demasiadas entradas de datos para escoger elementos específicos según su XPath. La manera de recolección seleccionada es haciendo uso de funcionalidades que ofrece la propia web, pues existe un botón que copia la tabla de los estados financieros al portapapeles. Copiando la tabla al portapapeles es posible rescatarla en Python con el método de Pandas pd.read_clipboard() que convierte aquello que se tenga en el portapapeles a un dataframe[8]. Una vez se tiene el dataframe se procede a guardar en formato CSV como en los otros casos, haciendo uso del método de pandas pd.to_csv().

[7] www.roic.ai
[8] Se trata de una estructura de datos similar a una tabla.

El atento observador que acceda a la web verá la presencia de otro botón que permite descargar directamente los estados financieros en fichero xlsx. Durante la realización de esta actividad no fue posible descargar ficheros mediante ese método, pues la descargar no iniciaba al hacer clic durante la ejecución automatizada (aunque si era posible realizar la tarea de manera manual). Se realizaron otros intentos haciendo uso de la API que surte al propio botón, recolectando primero los end-points[9] y posteriormente haciendo peticiones; sin embargo, este método tampoco dio resultado pues los end-points para muchas de las acciones no funcionaban.

Los errores fueron gestionados de manera análoga a los otros procesos. Finalmente se obtuvieron los estados financieros de 7372 tickers de empresas de las cuales ya se han recolectado información básica e histórico de precios.

4.3.5 Recolección de datos macroeconómicos

Al contrario que los datos sobre empresas, la recogida de datos macroeconómicos se realizó de manera manual. Estos datos se recogieron de diversas fuentes [46] [47] [48] [49] [50] [51] y se organizaron en un documento CSV para su posterior uso.

Son diversos los datos macroeconómicos escogidos y se expondrán durante la descripción de los datos. Al seleccionar los datos macroeconómicos debe tenerse en cuenta se han utilizado las medias de los periodos para conseguir estimaciones anuales (excepto indicación contraria) y se han utilizado estimaciones de expertos para algunos datos recientes sobre 2021 y 2022.

4.4 Descripción de datos

En este apartado se procede a la descripción de los datos recolectados, con el fin de documentar y comprender los datos disponibles.

Cabe destacar que tanto la información básica, como el histórico de precios, como los estados financieros se tienen para cada una de las empresas, aunque todas mantienen el mismo formato.

Histórico de precios:

- **Date**: Fecha del precio en formato “Mes día, año”.
- **Open**: Precio de apertura del periodo.
- **High**: Precio más alto del periodo.
- **Low**: Precio más bajo del periodo.
- **Close**: Precio de cierre del periodo, ajustado para tener en cuenta división de acciones[10].

[9] Punto final de comunicación de la API

[10] La división de acciones o “stock splits” es cuando la compañía divide sus acciones existentes en acciones múltiples para aumentar la liquidez de las acciones.

- **Adj Close**: Precio de cierre del periodo, ajustado para tener en cuenta división de acciones y dividendos y/o distribuciones de ganancias de capital.
- **Volume**: Volumen de acciones que se intercambian.

Granularidad de los datos[11]: Mensual

Existen algunas filas correspondientes a dividendos u otros eventos especiales. Si se trata de un dividendo aparece en "Open" el valor del dividendo y en "High" la palabra "Dividend", los campos siguientes aparecen vacíos.

Información básica

- **Name:** Nombre de la compañía.
- **Currency:** Moneda en la que están representados los estados financieros.
- **Sector:** Sector en el que opera la compañía.
- **Industry:** Industria en el que opera la compañía.
- **Country:** País al que pertenece la compañía.
- **IPO:** Fecha de salida a bolsa de la compañía.
- **Insider Percentage:** Porcentaje de las acciones que pertenecen a directores, oficiales o ejecutivos de la compañía.
- **Institution Percentage:** Porcentaje de las acciones que pertenecen a compañías u organizaciones que invierten el dinero en nombre de otras personas.

Granularidad de los datos: Solo una muestra, basada en los últimos datos disponibles.

Estados financieros

Existen tres estados financieros: la cuenta de resultados, el balance y el estado de los flujos de caja. Para cada acción se disponen de todos los campos de los tres estados en un único fichero, donde las columnas son los años y donde las filas son las filas de cada estado financiero (además de alguna que otra fila no deseada).

Se tiene en este fichero un total de 102 filas correspondientes a entradas de cuentas contables. El contenido de estas filas está perfectamente definido en el ámbito de la contabilidad financiera por lo que no se hará una descripción de cada una de estas filas [52].

La unidad de los estados financieros viene definida por la moneda de cada empresa. Aquellas que no estén representadas en dólares cuentas con una fila extra que presenta el tipo de cambio a dólares. A la hora de trabajar con los estados financieros serán convertidas todas las monedas a dólares.

Algunos de los elementos disponibles en los estados contables son: los ingresos ("*Revenue*"), el margen de beneficios ("*Net income ratio*"), los activos totales ("*Total Assets*"), la deuda a largo plazo ("*Long-Term Debt*"), etc.

Granularidad de los datos: Anual.

[11] La granularidad de los datos hace referencia a la medida temporal usada para medir los datos. Esto podrían ser días, horas, semanas, meses, etc.

Datos macroeconómicos

- **Year:** Año correspondiente al dato.
- **Inflation Rate US:** Tasa de inflación en EE. UU. Representado en porcentaje.
- **Inflation Rate China:** Tasa de inflación en China. Representado en porcentaje.
- **Inflation Rate World:** Tasa de inflación a nivel mundial. Representado en porcentaje.
- **Inflation Rate Euro Area:** Tasa de inflación en la zona euro. Representado en porcentaje.
- **US GDP:** PIB de EE. UU. Representado en dólares estadounidenses.
- **US GDP Per Capita:** PIB per cápita de EE. UU. Representado en dólares estadounidenses.
- **Implied US Population:** División de "US GDP" entre "US GDP Per Capita". Cantidad de población de EE. UU. que implican los datos.
- **China GDP:** PIB de China. Representado en dólares estadounidenses.
- **China GDP Per Capita:** PIB per cápita de China. Representado en dólares estadounidenses.
- **Implied China Population:** División de "China GDP" entre "China GDP Per Capita". Cantidad de población de China que implican los datos.
- **Euro Area GDP (USD):** PIB de la zona euro. Representado en dólares estadounidenses.
- **Euro Area GDP Per Capita:** PIB per cápita de la zona euro. Representado en dólares estadounidenses.
- **Implied Euro Area Population:** División de "Euro Area GDP" entre "Euro Area GDP Per Capita". Cantidad de población de la zona euro que implican los datos.
- **World GDP:** PIB a nivel mundial. Representado en dólares estadounidenses.
- **World GDP Per Capita:** PIB per cápita a nivel mundial. Representado en dólares estadounidenses.
- **Implied World Population:** División de "World GDP" entre "World GDP Per Capita". Cantidad de población del mundo que implican los datos.
- **US Unemployment Rate:** Tasa de desempleo de EE. UU. Representado en porcentaje.
- **China Unemployment Rate:** Tasa de desempleo de China. Representado en porcentaje.
- **Euro Area Unemployment Rate:** Tasa de desempleo de la zona euro. Representado en porcentaje.
- **World Unemployment Rate:** Tasa de desempleo a nivel mundial. Representado en porcentaje.
- **Global price of Agricultural Raw Material Index:** Precio global del Índice de Materias Primas Agrícolas, usado como indicador representativo del mercado global de los productos agrícolas. Representado en dólares estadounidenses.
- **CRB Commodity Index:** Precio del índice CRB de materias primas, usado como indicador representativo del mercado global de materias primas. Representado en dólares estadounidenses.

- **All-Transactions House Price Index US:** Índice del precio de la vivienda de EE. UU. Sirve como indicador representativo del mercado de vivienda en EE. UU. Representado en dólares estadounidenses.
- **30 year us fixed mortgage rates average:** Promedio de tasas hipotecarias fijas de EE. UU. a 30 años. Representado en porcentaje.
- **FED Funds Rate Average Yield:** Rendimiento promedio de la tasa de los fondos de la FED. La tasa de fondos federales es la tasa de interés a la que las instituciones de depósito (bancos y cooperativas de crédito) prestan saldos de reserva a otras instituciones de depósito al día siguiente, sin garantía. Representado en porcentaje.
- **FED Funds Rate Year Close:** Rendimiento a final de año de la tasa de los fondos de la FED. Representado en porcentaje.
- **10-Year Treasury Average Yield:** Rendimiento promedio de los bonos del Tesoro de EE. UU. a 10 años. Representado en porcentaje.
- **10-Year Treasury Year Close:** Rendimiento a fin de año de los bonos del Tesoro de EE. UU. a 10 años. Representado en porcentaje.
- **US Wage Growth:** Crecimiento de los salarios en EE. UU. Representado en porcentaje.
- **Real Wage Growth:** Crecimiento real de los salario en EE. UU. Diferencia entre "Inflation Rate US" y "US Wage Growth". Hace referencia al crecimiento de los salaries una vez se ha tenido en cuenta la inflación. Representado en porcentaje.

Granularidad de los datos: Anual.

4.5 Limpieza y preparación de datos

4.5.1 Arreglos básicos

Debido a la enorme cantidad de datos que se manejan se vuelve necesario aplicar arreglos en lotes que se apliquen de manera generalizada, reservando la revisión manual para casos concretos. Se procede a nombrar los arreglos más relevantes realizados a cada grupo de ficheros.

Arreglos en los ficheros de precios:

- Eliminación de tablas vacías, correspondientes a acciones sin información sobre los precios.
- Conversión de la columna "Date" a tipo datetime. Este tipo es más apropiado y cómodo a la hora de trabajar con fechas.
- Eliminación de filas correspondientes a eventos especiales. Algunas de las filas marcan eventos como el reparto de un dividendo o la división de acciones; se eliminan estas filas de la tabla principal, aunque se guardan temporalmente en una tabla aparte.
- Conversión de las columnas "Open", "High", "Low", "Close", "Adj Close" y "Volume" a tipo numérico.
- Agrupación de los datos por años. Se debe tener la misma granularidad temporal en todos los datos, por lo que se convierten los datos mensuales a datos anuales. Para la agrupación se utilizada la mediana,

con el fin de disminuir la influencia de precios atípicos y de las variaciones cortoplacistas del mercado.
- Adición de una nueva columna llamada "Dividend" que contendrá los dividendos anuales, extraídos de la tabla anteriormente apartada.
- Adición de una nueva columna llamada "Dividend Cum" que contendrá la suma cumulativa de los dividendos. El objetivo de esta columna es proporcionar una idea de la rentabilidad adicional que viene dada por los dividendos en un plazo concreto. [12]
- Adición de una nueva columna "Price with cum Dividends" que contiene la suma del precio ajustado (columna "Adj Close") y la suma cumulativa de dividendos (columna "Dividend Cum"). Esta columna será necesaria para calcular el retorno real de los inversores, es decir, revalorización del precio más retorno en forma de dividendo.
- Eliminación de las filas con valores *NaN*[13] en el precio.

Arreglos en los ficheros de los estados financieros:

- Uso de las entradas de los documentos financieros (por ejemplo "Ingresos") como índices, borrando el índice numérico antiguo.
- Trasposición de los datos, convirtiendo las filas en columnas y viceversa.
- Algunas entradas de los estados financieros tienen el mismo nombre, por lo que se renombran aquellas diferentes y se procede a eliminar las duplicadas.
- Renombramiento de entradas con errores gramaticales o inconsistencias.
- Eliminación de entradas sin utilidad como "*INCOME STATEMENT*", "*BALANCE SHEET*", "*CASH FLOW STATEMENT*", "*Sec Link*" y otros.
- La página de la cual se han extraído los datos no diferencia entre valores nulos y ceros, utilizando para ambos casos la cadena de caracteres "- - " o la cadena "*undefined*". Se sustituye la cadena de caracteres por valores NaN, para indicar campos vacíos o faltantes.
- Conversión de todas las columnas a tipo numérico.
- Eliminación de estados financieros vacíos.

Después de la aplicación de los arreglos sobre los ficheros de precio y de estados financieros, se juntan ambos ficheros para cada empresa. Se obtiene así un solo fichero con la información financiera y de precios, al cual se pasa a referirse como fichero de datos fundamentales.

Se busca obtener un índice con información útil sobre las empresas. Para ello se crea un nuevo fichero llamado "basic_information.csv" donde se guardará toda la información básica (sacada sus respectivos ficheros) de las empresas en una misma tabla. Las columnas serán las mismas que ya se tenían ("Name", "Currency", "Sector", etc.) mientras que las filas serán los tickers de cada empresa.

[12] Más adelante durante la limpieza de los estados financieros algunas marcas temporales serán eliminadas, haciendo el cálculo de la suma cumulativa incoherente. No es necesario entonces, realizar el cálculo de esta columna en este momento.
[13] NaN hace referencia a "Not a Number" y será utilizado para referirse a valores incorrectos o nulos.

Arreglos en el fichero de *basic_information* recién creado:

- Se quitan las entradas de acciones sin datos.
- Se reemplazan los campos con la cadena de caracteres "- -" por NaN.
- Algunas empresas no tenían datos sobre su IPO[14] y en su lugar tenían la cadena de carácteres "Invalid Date". Se sustituye esta cadena por valores NaN.
- Conversión de la columna "IPO" a datetime.
- Adición de la columna "Years since IPO", obtenida restando al año actual el año de su salida a bolsa. Esto proporciona una referencia del tiempo de vida mínimo de la empresa.
- Se eliminan las acciones sin información sobre su sector.

Después de los arreglos el fichero basic_information.csv contiene información de 6684 empresas.

4.5.2 Reconstrucción de datos financieros

Los datos más valiosos durante este proyecto serán los estados financieros, pues son estos los datos que van a representar a cada empresa y sobre los cuales se intentará buscar una relación con el precio de cotización. Debido a la importancia de estos datos, se dedica una parte de la limpieza de datos a la reconstrucción y mejora de la calidad de estos.

El método de reconstrucción hace uso de la coherencia interna de los estados financieros y de sus reglas contables. Por ejemplo, si se tiene el dato de beneficios y el dato de ingresos se puede inferir el margen neto, pues se conoce su fórmula:

$$margen\ neto\ = \frac{beneficios}{ingresos}$$

En otro caso es posible que solo se tengan los ingresos y el margen neto, pudiendo así reconstruir los beneficios y, pudiendo reconstruir cualquier elemento si se tienen los componentes restantes de la ecuación. Sin embargo, existen fórmulas para las cuales no se tienen la mayoría de los elementos, al menos no de manera directa.

Si se quisiera calcular el margen neto, pero no se dispusiera de los ingresos, se puede intentar la reconstrucción de estos antes de intentar resolver la ecuación. Si se tienen los costes de bienes vendidos y el beneficio bruto, se pueden reconstruir los ingresos con la siguiente fórmula:

$$Ingresos - \ Costes\ de\ bienes\ vendidos = Beneficio\ bruto$$

De esta manera es posible reconstruir primero los ingresos y posteriormente el margen neto. Además, se cuenta con la ventaja del cálculo de los elementos puede realizarse de diferentes maneras, por lo que la falta de un dato no suele ser bloqueante para la reconstrucción.

[14] Significa "Initial Public Offering" (en español "Oferta pública de venta") y hace referencia a la fecha de la salida a bolsa de las empresas.

Con este enfoque se ha implementado un método automático de arreglo de los estados financieros, que aprovecha más de 90 reglas contables para intentar reconstruir los datos faltantes. La reconstrucción se aplica de manera iterativa (a todas las columnas) y de manera recursiva para intentar reparar todos los datos necesarios. El algoritmo conserva memoria de las columnas visitadas y las reparadas para optimizar el cómputo, y debido a la recursividad, la ejecución adopta forma de árbol (ver Figura 3), pues a cada iteración se presenta la posibilidad de reconstruir los datos faltantes. Si se quisiera reconstruir el margen neto, pero no se dispusiera de los beneficios o los ingresos, se ejecutaría la misma función de reconstrucción sobre el campo "Beneficios", para tratar de reconstruir los beneficios antes de seguir con la reconstrucción del margen neto.

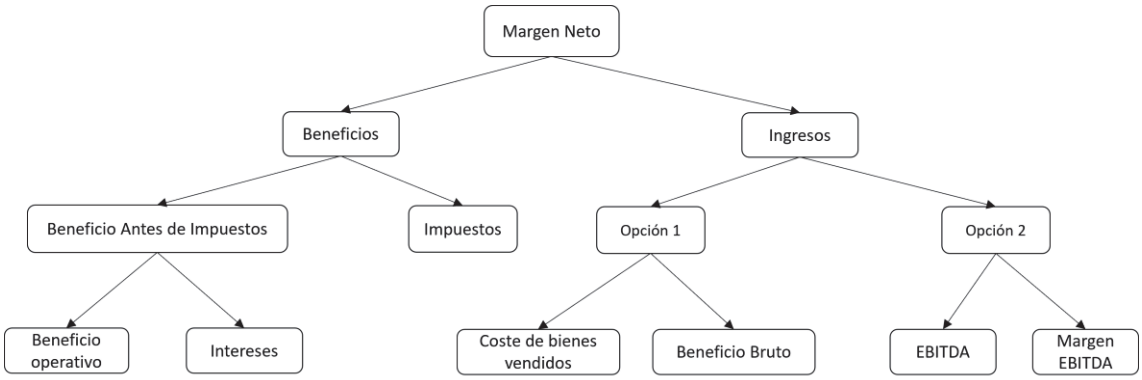

Figura 3: Visualización simplificada de la reparación del margen neto

Durante la exploración de los estados financieros para la reconstrucción también se aplican otros arreglos básicos.

También se ha elaborado una lista de datos que se consideran de gran importancia[15], la lista es la siguiente:

- "Revenue"
- "Gross Profit"
- "Operating Income"
- "Income Tax expense (Gain)"
- "Net Income"
- "Total Assets"
- "Total Liabilities"
- "Total Stockholders Equity"

El objetivo de la lista es garantizar una calidad mínima de los datos, por lo que durante la exploración se eliminan aquellas entradas que no tengan por lo menos estos datos (después de la reconstrucción). Los datos listados serán denominados "Datos Core", debido a que se consideran integrales para conocer la empresa.

[15] Estos datos corresponden a entradas importantes en los estados financieros descargados. Es altamente improbable que una empresa carezca de estas entradas, por lo que, si estos datos faltan, es bastante seguro asumir que sus estados financieros no son correctos y no deben tenerse en cuenta.

Durante la exploración también se eliminan entradas de datos vacías y, posteriormente, se realiza el cálculo de la suma cumulativa de dividendos y del precio junto con la suma cumulativa. Por último, debido al enfoque en el largo plazo son necesarias acciones con varios años de vida y un número significativo de entradas, por lo que se eliminan aquellas con menos de 4 años de datos.

Este proceso debe realizarse para 6684 empresas y ya se ha expuesto la complejidad del algoritmo, por lo que ha implementado mediante un proceso multihilo para acelerar la ejecución.

Una vez acabada la reconstrucción formal siguen existiendo campos vacíos, pues la página utilizada para extraer los datos no distinguía de datos con valor 0 y datos nulos. No es lo mismo no tener datos sobre una entrada que el valor de esta sea 0, por lo que se debe tener cuidado a la hora de reconstruir estos datos. Por este motivo se ha seleccionado cuidadosamente un conjunto de entradas que, a falta de datos, lo más probable es que el valor sea 0. Este segundo arreglo que se realiza se contempla aparte de la reconstrucción y se ha denominado "*Fix trivial*".

Lista de columnas afectadas por el *fix trivial*:

- "Other Liabilities"
- "Preferred Stock"
- "Common Stock"
- "Tax Assets"
- "Tax Payable"
- "Capital Lease Obligations"
- "Common Stock Repurchased"
- "Common Stock Issued"
- "Dividends Paid"
- "Interest Income"
- "Other Expenses"
- "Stock Based Compensation"
- "Deferred Income Tax"
- "Interest Expense (Gain)"
- "Other Investing Activities"
- "Deferred Tax Liabilities"
- "Purchases of Investments"
- "Dividend"
- "Other Non-Current Liabilities"
- "Accounts Payable (Cash Flow)"
- "Accounts Receivable"
- "Intangible Assets"
- "Goodwill"
- "Sales/Maturities of Investments"
- "Investments"
- "Other Assets"
- "Acquisitions Net"
- "Inventory (Balance)"
- "Inventory (Cash Flow)"
- "Short-Term Debt"

- "Debt Repayment"
- "Long-Term Debt"
- "Research and Development Exp."
- "Effect of Forex Changes on Cash"
- "Other Liabilities"
- "Other Working Capital"
- "Short-Term Investments"
- “Dividend Cum”

La reconstrucción y el *fix trivial* ofrecen la posibilidad de reconstruir estados financieros sin apenas información, pero tienen un problema que debe tenerse en cuenta. Si se dispone de poca información de buena calidad no hay problema, sin embargo, si se dispone de poca información y, además, es de mala calidad, esto se propaga a los nuevos campos reconstruidos. Esto puede dar lugar a datasets de empresas aparentemente correctos, pero completamente errados cuando se contrastan con la realidad. Para el presente proyecto se ha intentado minimizar este riesgo y se han aplicado medidas de limpieza conservadoras para evitar realizar el análisis con datos incorrectos.

Tras reparar los datos y eliminar aquellos sin suficientes filas se actualiza la lista que se tenía en el fichero “Basic_information.csv” y se guarda en un nuevo fichero “Basic_information_repaired.csv”. Así se obtiene así una lista de 3337 empresas; siendo este el paso que más ha reducido la muestra (45% de los datos originales).

Evaluación de la reconstrucción

Se procede ahora a evaluar la efectividad de los arreglos, tanto la reconstrucción normal como el *fix trivial*. Para ello se añade al fichero “Basic_information_repaired.csv” columnas con el porcentaje de errores de cada acción antes y después de los arreglo. En la Tabla 1 se visualizan los resultados agregados.

Tabla 1: Datos estadísticos relevantes sobre los datos faltantes según el arreglo aplicado

	Sin reconstrucción	**Con reconstrucción**	**Con reconstrucción y *fix trivial***
count	3308.00	3239.00[16]	3308.00
mean	0.23	0.15	0.01
std	0.07	0.04	0.01
min	0.06	0.05	0.00
25%	0.19	0.12	0.00
50%	0.23	0.15	0.00
75%	0.27	0.18	0.01
max	0.57	0.47	0.14

[16] Si la reconstrucción normal no es suficiente para llegar al mínimo de filas no nulas requeridas se eliminan. Esto provoca que existan acciones que cumplan con el número mínimo de filas al aplicar reconstrucción y *fix trivial*, pero no solo reconstrucción.

La reconstrucción mejora significativamente la calidad de los datos, teniendo el top 75% de los datos reconstruidos el mismo porcentaje de datos faltantes que el top 25% de los datos sin reconstruir. Usando la mediana (por ser un indicador más robusto) se aprecia una mejora significativa de 0.23 a 0.15, así como se observa mejoría en el resto de los parámetros.

Al añadir el *fix trivial* el salto en calidad de los datos es mayor sobre el papel, sin embargo, esto no aporta demasiada información útil, como ya se ha comentado. Aun así, cabe destacar la mejora sustancial que se da en cuanto a datos faltantes, llegando incluso a tener empresas sin ningún dato faltante.

Se puede apreciar la diferencia en la calidad de los datos mediante el histograma en la Figura 4, donde la distribución se desplaza a la izquierda conforme se aplican reparaciones en los datos.

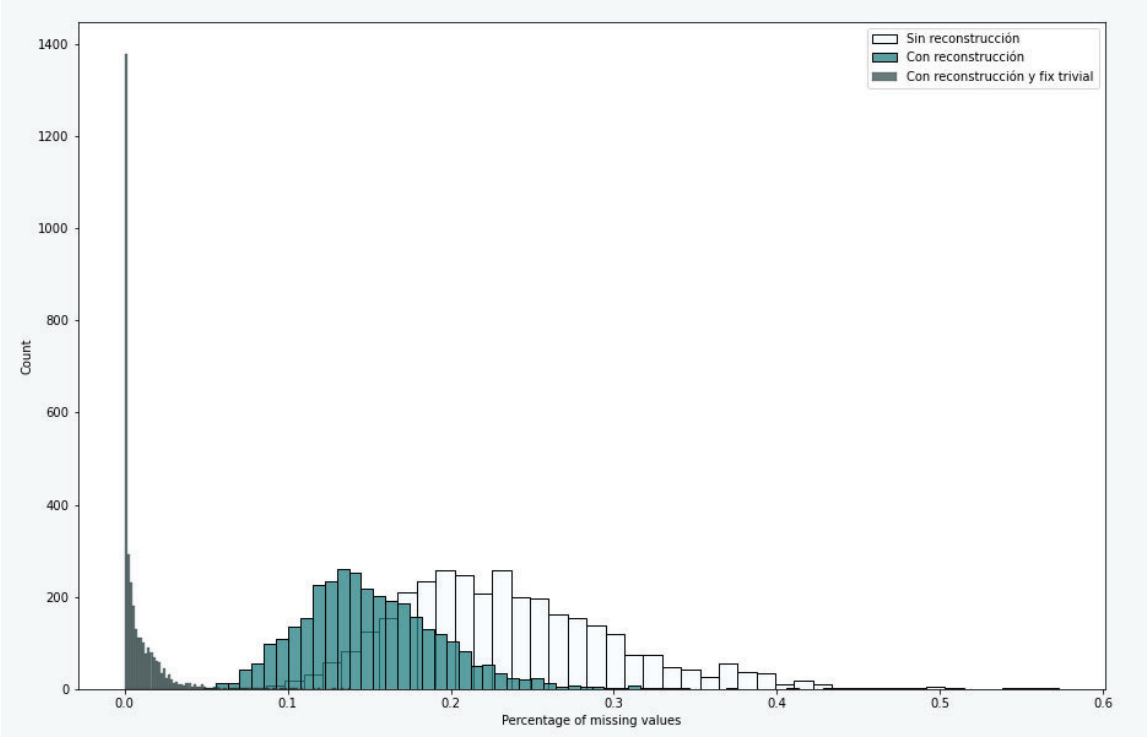

Figura 4: Histograma de los datos faltantes según las correcciones aplicadas

Sin embargo, no solo son de interés el agregado de los fallos, sino también los fallos por parámetro. El objetivo de esta evaluación es obtener en qué columnas se observan más fallos e intentar repararlos, modificando la reconstrucción del paso anterior. Este proceso es iterativo y a nivel de proyecto se realiza un ciclo de reparación y visualización de la corrección característico de la metodología CRISP-DM. Por este motivo las figuras mostradas a continuación son el producto de varias iteraciones de reparaciones y representan el estado final, pero no el proceso intermedio.

En la Figura 5 ya se ha realizado la reconstrucción, pero todavía no se ha aplicado el *fix trivial*, por lo que algunas columnas que deberían estar rellenadas a cero aparecen como error. Esta gráfica sin embargo es útil para saber qué columnas son más problemáticas e intentar desarrollar nuevos arreglos para la etapa anterior. El resultado que se observa es después de varias iteraciones de

este proceso, por lo que las columnas más problemáticas han sido ya solucionadas.

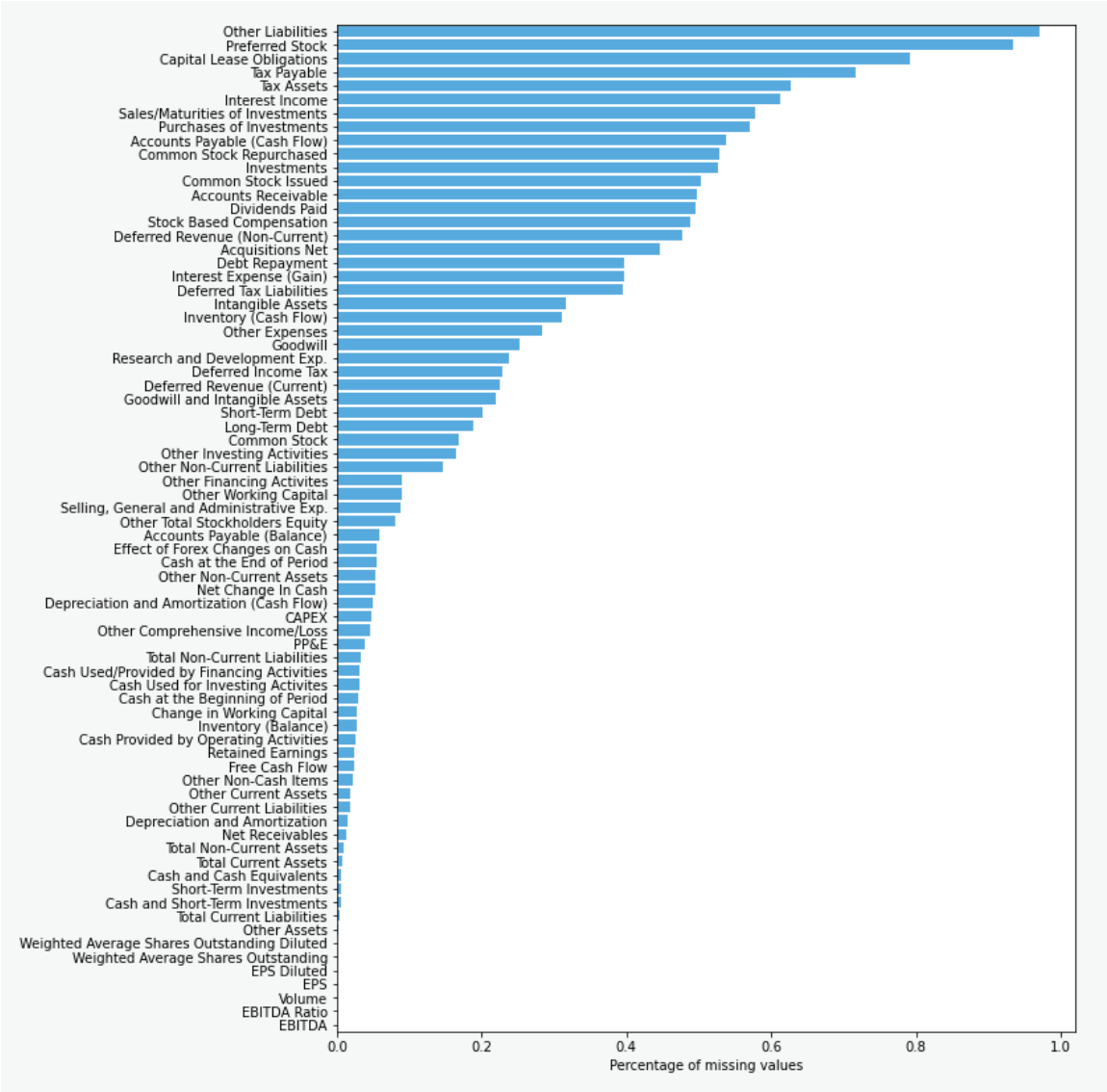

Figura 5: Porcentaje de datos faltantes por columna (Después de reconstruir)

En la Figura 6 se observa el porcentaje de errores por columna después de haber aplicado la reconstrucción y el *fix trivial*, por lo que esos datos permanecerán con valores nulos.

Estos campos no se han incluido en el *fix trivial* debido a que son comunes en los estados financieros y, el hecho de que no existan es más probable que se deba la falta de los datos que a que el valor sea 0. Por ejemplo, es imposible que la empresa no tenga acciones en circulación o poco probable que no tenga unas oficinas (representados en "PP&E") o que no tenga activos corrientes (representados en "Total Current Assets").

Sin embargo, este arreglo es arbitrario y podría solucionarse con una mejor fuente de datos, mejorando la estimación arbitraria con la ayuda de un

profesional en contabilidad o mejorando la habilidad de reconstrucción del algoritmo.

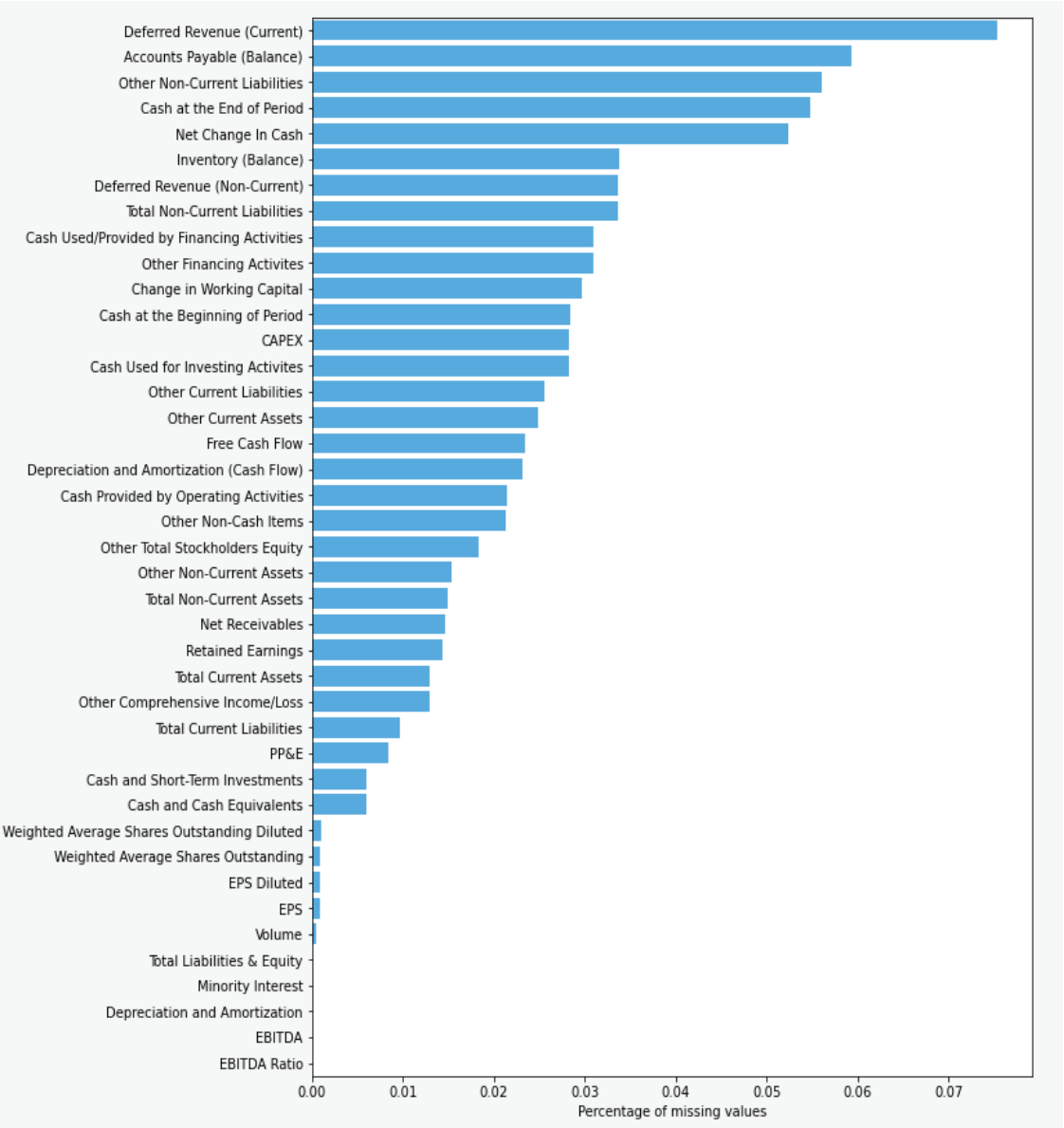

Figura 6: Número de datos faltantes por columna (después de reconstruir y aplicar el fix trivial)

Los resultados finales pueden apreciarse en la Figura 7. Es esta figura el color más blanco representa el porcentaje de errores antes de ningún arreglo, el color azul-verdoso representa el porcentaje de errores después del primer arreglo y el color azul claro el porcentaje de errores después de todos los arreglos.

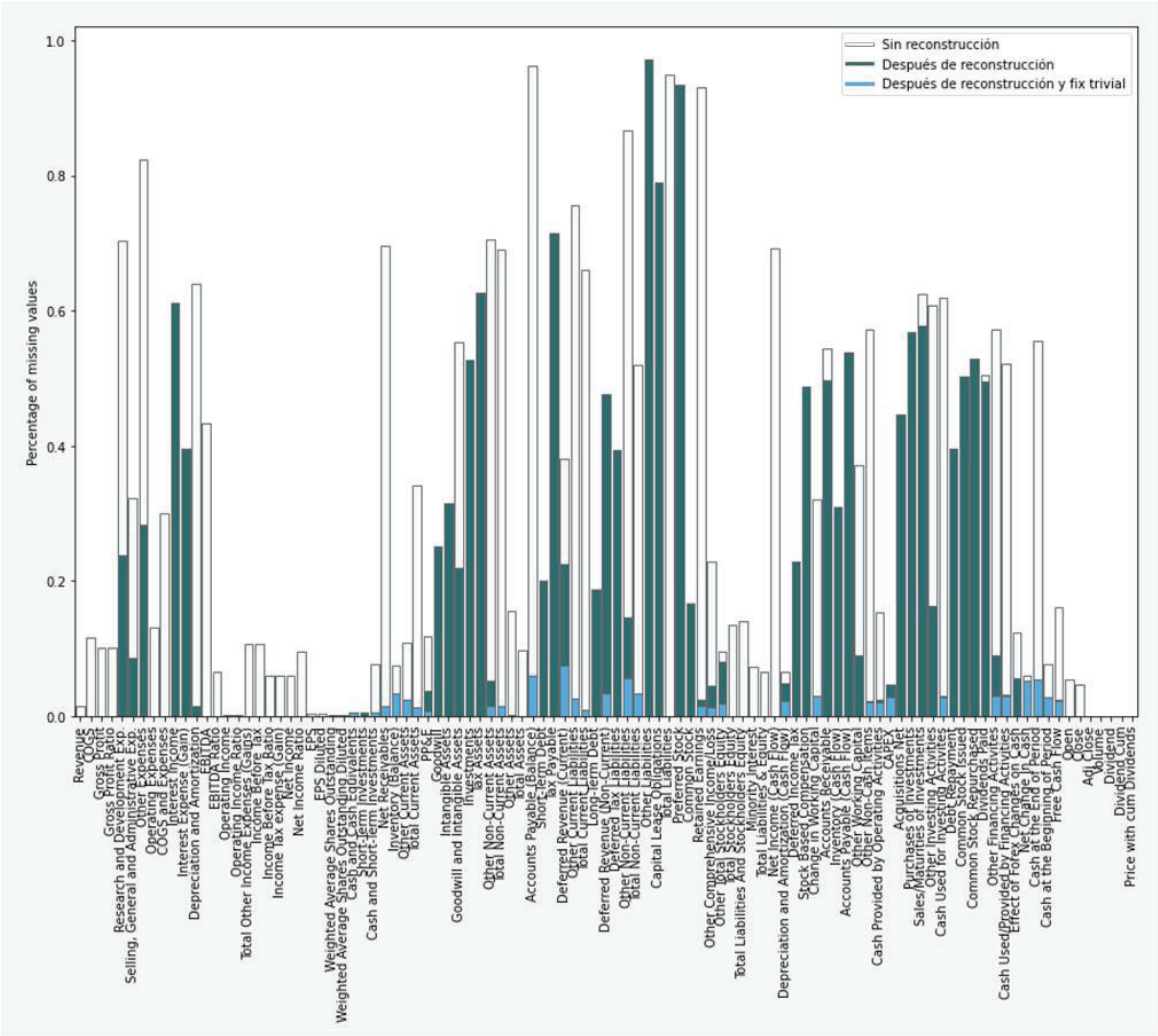

Figura 7: Porcentaje de valores faltantes por columna y tipo de arreglo

Por último, antes de proseguir, se revisan las acciones con más fallos de manera manual para evitar posibles casos atípicos, aunque esta revisión se hará en profundidad más adelante en su correspondiente apartado.

4.5.3 Adición de características

A la hora de elaborar modelos añadir características adicionales puede ir en detrimento de los intereses del proyecto, sin embargo, para el análisis de datos, estas nuevas variables pueden aportar información valiosa.

Existen dos ficheros donde se quieren añadir características:

1. El fichero de “Basic_information_repaired.csv”. Aquí se guarda la lista de las empresas junto con información básica sobre las mismas (sector, nombre, número de años, porcentaje de errores, etc.).
2. El fichero de datos fundamentales de cada empresa, donde se guardan los estados financieros y el precio por años.

Características añadidas al fichero “Basic_information_repaired”:

- Se añade la capitalización de mercado de cada acción en una nueva columna llamada “Market Cap”. La capitalización de mercado es la medida comúnmente utilizada para juzgar el tamaño de una empresa cotizada y se calcula multiplicando el número de acciones por el precio de cada acción.
- Se calcula la tasa de crecimiento anual compuesta histórica de todas las acciones y se guardan en las columnas “CAGR[17]” y “CAGR with divs” dependiendo de si se tiene en cuenta o no el dividendo. Puesto que “CAGR with divs” es igual a la columna “CAGR” pero con algunos ingresos extra por dividendos, el histograma es muy parecido, movido ligeramente a la derecha como se puede ver en la Figura 8.
- Se calcula la diferencia de rentabilidad entre “CAGR with divs” y “CAGR” y se guarda en “Difference CAGR”. Puesto que la rentabilidad con dividendos siempre va a ser mayor a la rentabilidad sin dividendos, es buena señal que no existan valores negativos para “Difference CAGR”.
- Se añade una nueva columna “Years”. Esta columna se obtiene calculando los años transcurridos desde la primera entrada de datos fundamentales hasta este año.

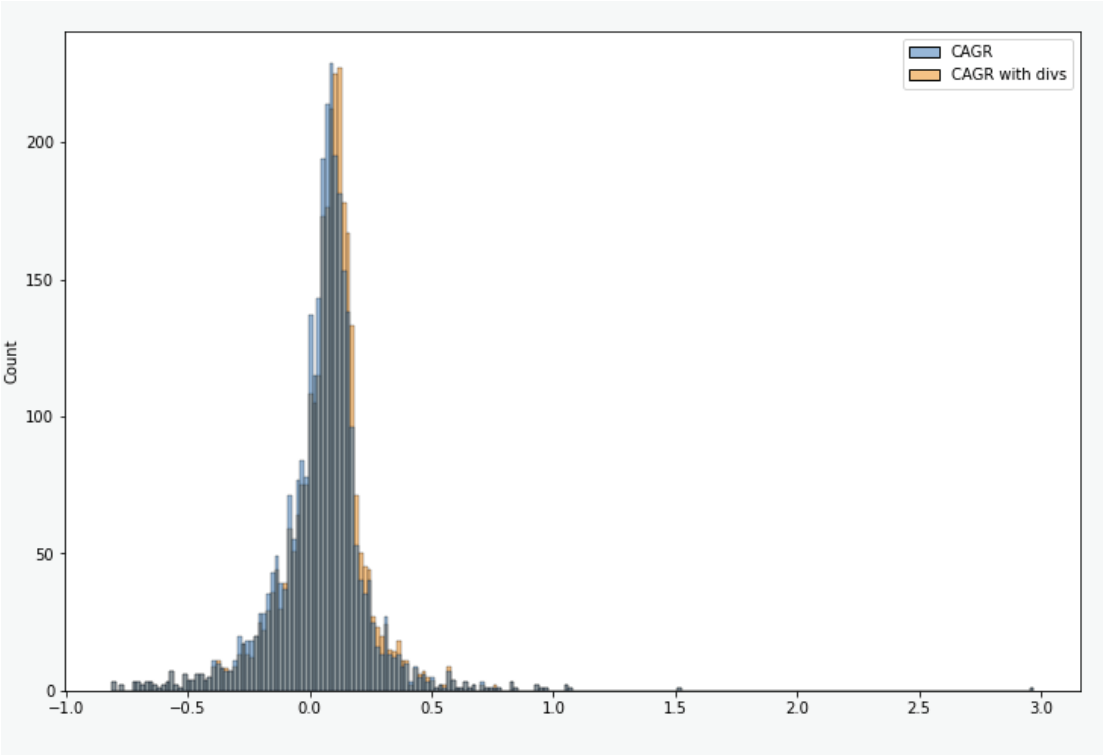

Figura 8: Histograma del crecimiento anual compuesto con y sin dividendos.

Al añadir características se encuentran más acciones con datos incoherentes, que se proceden a eliminar o arreglar.

- Se eliminan las acciones con capitalización de mercado nula, negativa o igual a 0.

[17] CAGR hace referencia a “Compounded Annual Growth Rate”, es decir, “Tasa de crecimiento anual compuesta”

- Se eliminan las acciones cuyo “Insider Percentage” no este comprendido entre 0 y 1.
- Se eliminan las acciones cuyo “Institution Percentage” no este comprendido entre 0 y 1.
- Se eliminan las empresas con “CAGR” infinito (consecuencia de dividir entre 0).
- Se revisa manualmente los valores más extremos según las columnas “CAGR”, “CAGR with divs” y “Difference CAGR”. Se eliminan aquellas que presenten datos erróneos.
- Algunas fechas de la columna “IPO” son erróneas y presentan fechas más tardías a su salida a bolsa; esto afecta a la columna “Years since IPO”. Para evitar edades de empresas más jóvenes de lo que realmente son, se comprueba si se disponen de datos desde antes de la fecha y, si es así, se pone la fecha desde la primera entrada de datos (columna “Years”).
- Se elimina la columna “IPO” dado que ya no aporta información.

Se añaden a los estados financieros más de 100 métricas usadas comúnmente en el campo de la inversión, que serán de interés durante el análisis. El cálculo de estas métricas puede ser encontrado en el anexo.

Algunas de las empresas no son originarias de EE.UU. y por tanto la información de sus estados financieros está en otra moneda. Antes de realizar el cálculo de las nuevas características, se realiza la conversión de moneda para todas las columnas que representen unidades monetarias en otra moneda distinta al dólar estadounidense, es decir, todas menos:

- "Gross Profit Ratio"
- "EBITDA Ratio"
- "Operating Income Ratio"
- "Income Before Tax Ratio"
- "Net Income Ratio"
- "Weighted Average Shares Outstanding"
- "Weighted Average Shares Outstanding Diluted"
- "Open"
- "Close"
- "Adj Close"
- "Volume"
- "Dividend"
- "Dividend Cum"
- "Price with cum Dividends"
- "Forex Rate"

Una vez realizada la conversión de moneda, se procede a calcular los nuevos ratios.

Características añadidas a los estados financieros:

- Tax Rate
- Deferred Revenue
- Net Interest Income
- Free Cash Flow to the Firm
- Tangible Assets

- Adjusted Operating Income
- EBIT
- Operating Cash Flow
- Common Book Value
- Tangible Book Value
- Common Tangible Book Value
- Free Cash Flow Ratio
- Selling, General and Administrative Exp. Ratio
- Research and Development Exp. Ratio
- Other Expenses Ratio
- Net Interest Income Ratio
- Depreciation and Amortization Ratio
- EBIT Margin
- Adjusted Operating Margin
- Operating Cash Flow Margin
- Cost-to-Income Ratio
- Operating Expense Ratio
- Revenue per share
- Operating Income per share
- FCF per share
- CAPEX per share
- Book value per share
- Dividends per share
- Total Debt
- Financial leverage
- Cash to Debt Ratio
- Cash & Investments
- Cash & Investments to Debt Ratio
- Net Debt
- Net Debt w/Investments
- Working Capital
- Current Ratio
- Quick Ratio
- Cash to Current Assets
- Cash to Assets
- Debt to Equity
- Debt to Assets
- Interest Coverage
- Current Liability Coverage Ratio
- Cash Ratio
- Net Working Capital to Assets
- Long Term Debt Ratio
- Total Debt/ Assets
- Net Debt/ Assets
- Total Debt/ Equity
- Total Liabilities/ Equity
- Net Debt/ Equity
- Equity Ratio
- Equity Multiplier

- Debt/ Tangible Book Value
- Net Debt/ EBITDA
- Cash to Debt
- Cash Flow Coverage Ratio
- Free Cash Flow/ Long Term Debt
- Debt Leverage Ratio
- Interest Expense to Debt Ratio
- Cash Sales
- Cash Revenue Adjustment
- CFO/ Net Income
- Return on Equity
- NOPAT
- Invested Capital
- Return on Invested Capital
- Capital Employed
- ROCE
- CFROI
- Asset Turnover Ratio
- Inventory Turnover
- Receivables Turnover
- Cash Turnover Ratio
- Gross Profitability Ratio
- Tangible Gross Profitability Ratio
- Cash Return on Invested Capital
- Adjusted Return on Capital Employed
- Cash Return on Capital Employed
- EBIT Return on Assets
- EBIT Return on Tangible Assets
- Return on Assets
- Return on Tangible Equity
- Cash Return on Equity
- Return on Retained Earnings
- R&D / Assets
- R&D / Book
- CapEx / Assets
- CapEx / Fixed Assets
- Retained Earnings / Total Assets
- Inventory / Assets
- Accounts Receivable / Assets
- Plowback Ratio
- Dividend & Repurchase / FCF
- Dividend & Repurchase / EBITDA

Por cada una de las columnas añadidas o ya existentes, se calcula su crecimiento cada año, además de su media de crecimiento de 3, 5 y 10 años.

En este punto se conservan datos de 3022 empresas con 197 columnas (sin contar el cálculo de los crecimientos).

4.5.4 Valores atípicos (Outliers)

La calidad de los datos es la principal prioridad, por lo que antes de pasar al análisis se va a realizar una última comprobación sobre los datos atípicos. Esta comprobación se realizará por medio de dos métodos:

1. Revisión manual utilizando diagramas de dispersión (comparando todas las columnas frente a su capitalización de mercado)[18].
2. Uso de K-means para la agrupación automática en clusters[19] y revisión manual de estos.

La elección de K-means como método de detección de datos atípicos es resultado de un proceso iterativo con diferentes algoritmos, tras el cual se determina que la K-mean presenta la mejor solución.

Diagramas de dispersión

El objetivo es comparar todos los datos fundamentales contra la capitalización de mercado, para así detectar entradas de datos anómalas. Para ello se crea un dataset que contiene la información básica y los datos fundamentales (usando la media de todos los años[20]). A nivel técnico cabe destacar que se crea primero el dataset en un array de Numpy y es posteriormente convertido a un dataframe de pandas. Esto se hace para mayor eficiencia de cómputo y menor tiempo de ejecución. Acto seguido se sustituyen los valores "infinitos" y los valores nulos por cero.

Se realizan dos exploraciones normalizando los datos de maneras diferentes:

- En la primera exploración cada fila se reescala de manera independiente para que la norma de la fila sea igual a 1.
- En la segunda exploración, cada columna se reescala asumiendo una distribución normal de media 0 y desviación 1.

En la

Figura 9 se pueden observar ejemplos de los diagramas de dispersión generados por la primera exploración. La única diferencia radica en que a la hora de hacer el análisis también se dibujaban los tickers de las acciones para poder identificarlas.

Los datos atípicos en los diagramas de dispersión son aquellos alejados inusualmente del resto y que presentan valores anormales para su capitalización de mercado. La anormalidad de estos valores es contextual a cada variable estudiada, y no existen unas pautas definidas para detectarlos, por lo que la detección es subjetiva y por tante, al observar algún dato atípico se procedía a la revisión manual estos.

---

[18] Se ha optado por el uso de diagramas de dispersión frente a los diagramas de cajas y bigotes dado que los valores a visualizar sólo toman contexto frente a su capitalización de mercado (100 millones de ganancia pueden ser normales para una empresa pequeña, pero anormales para una empresa grande). Además, el uso de los diagramas de dispersión permite comenzar a explorar las relaciones entre los factores fundamentales y su cotización.

[19] Denominación común para referirse a agrupaciones o grupos.

[20] El uso de la media es intencional, se busca destacar los resultados anómalos.

Algunos ejemplos de los casos más frecuentemente observados eran empresas con erratas en los estados financieros, donde habitualmente existía una fila que mostraba datos artificialmente altos[21] o directamente, valores erróneos. Al observar que una empresa de pequeño tamaño genera más ingresos que las más grandes (detectado como un punto anormalmente alejado del resto en los diagramas de dispersión), se procede a la revisión manual; que consiste en contrastar los datos financieros de dicha empresa frente a otras fuentes de información gratuitas. Si el error podía arreglarse, se aplicaba el arreglo correspondiente.

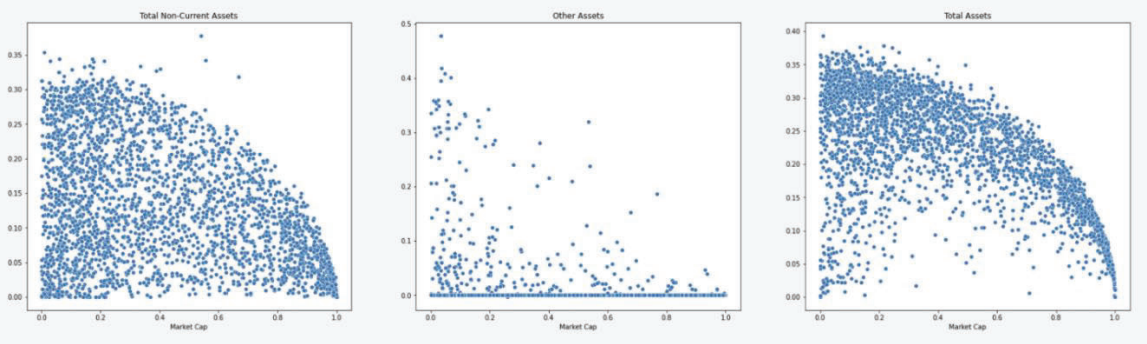

Figura 9: Gráficas de dispersión de los activos no corrientes, otros activos y el total de activos (eje Y) y la capitalización de mercado (eje X). Usando el primer método de exploración.

En base a lo observado en los diagramas de dispersión se realiza una nueva serie de arreglos:

- La acción "BYDDF" tiene durante los años 2010 y 2015 más acciones normales que diluidas, lo cual es imposible. Se soluciona asumiendo que las acciones diluidas son iguales al número de acciones normales.
- Conversión de moneda errónea en las acciones japonesas, por lo que es necesario eliminarlas de la muestra. Debido a esto se procede a revisar todas las conversiones de moneda, confirmando que el resto de las conversiones son correctas.
- Se detectan multitud de acciones con datos inservibles, por lo que se quitan del dataset.
- Se detectan datos atípicos en acciones individuales, por lo que se eliminan las filas con datos erróneos.

Se observa durante el proceso de limpieza manual, una gran parte de las anomalías y los casos atípicos correspondían a empresas de biotecnología, lo cual puede sesgar la muestra de análisis y se tendrá en cuenta.

K-means para detección de *outliers*

El uso de K-means para la detección de *outliers* aprovecha el hecho de que estos se encuentran generalmente a mayor distancia del grupo. Utilizando el algoritmo con un solo cluster y midiendo la distancia de cada punto al centroide

[21] Comúnmente se encontraba una fila cuyos valores estaban multiplicados por un millón, errata de la página de la que se extraían los datos.

es posible detectar valores atípicos, pues corresponderán a aquellos puntos más alejados.

Para el uso de k-means existen multitud de parámetros que se deben tener en cuenta, entre ellos:

- Las columnas utilizadas**.** Puesto que la intención es encontrar errores en los estados financieros, se han utilizado solo las columnas correspondientes a estos.

- El número de clusters. Como la intención es la de detección de *outliers*, se ha utilizado un solo cluster.

- La métrica de distancia. La métrica de distancia utilizada ha sido la distancia euclídea.

Como se aprecia en la Figura 10 existen una gran cantidad de elementos que podrían ser *outliers* (aquellos en color naranja). Además, si se coge el conjunto del último decil en cuanto a distancia (aquellos más alejados), se observa que la lista obtenida guarda algunas similitudes con la lista de tickers que se ha decidido eliminar en la revisión manual[22].

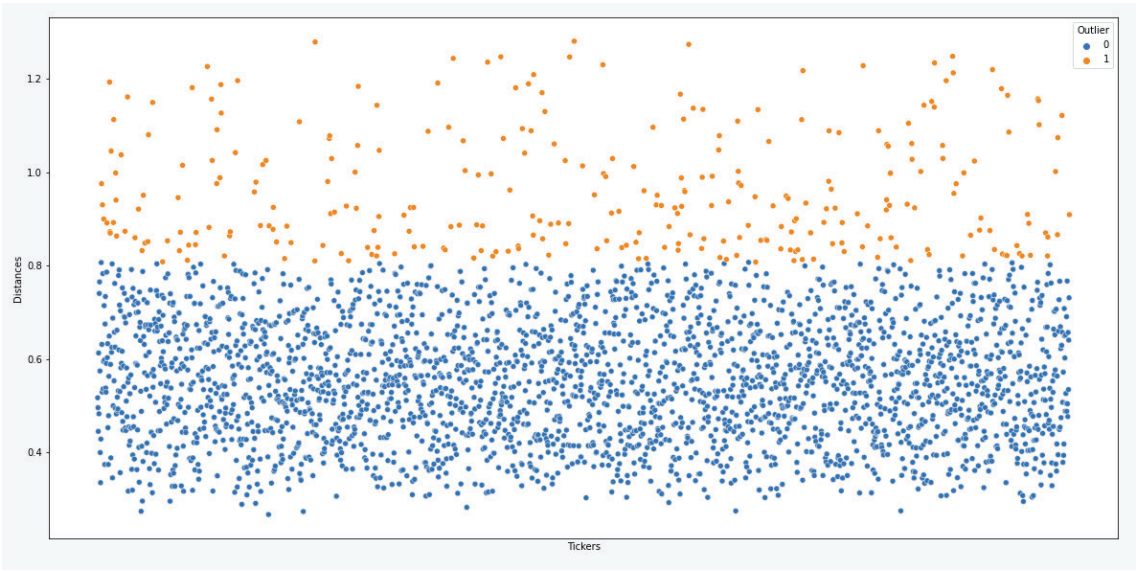

Figura 10: Distancias de cada punto respecto del centroide.

Se procede entonces a la revisión manual de algunos de los elementos considerados como *outliers*. Siguiendo el método utilizado anteriormente y a modo de referencia se exploran los gráficos de dispersión anteriormente mostrados, pero esta vez, marcando los posibles *outliers* detectados mediante el clustering (ver Figura 11).

[22] La aplicación de este algoritmo se está realizando sobre el conjunto de datos antes de la eliminación manual de datos, para poder observar similitudes entre los outliers detectados por el algoritmo y por un humano.

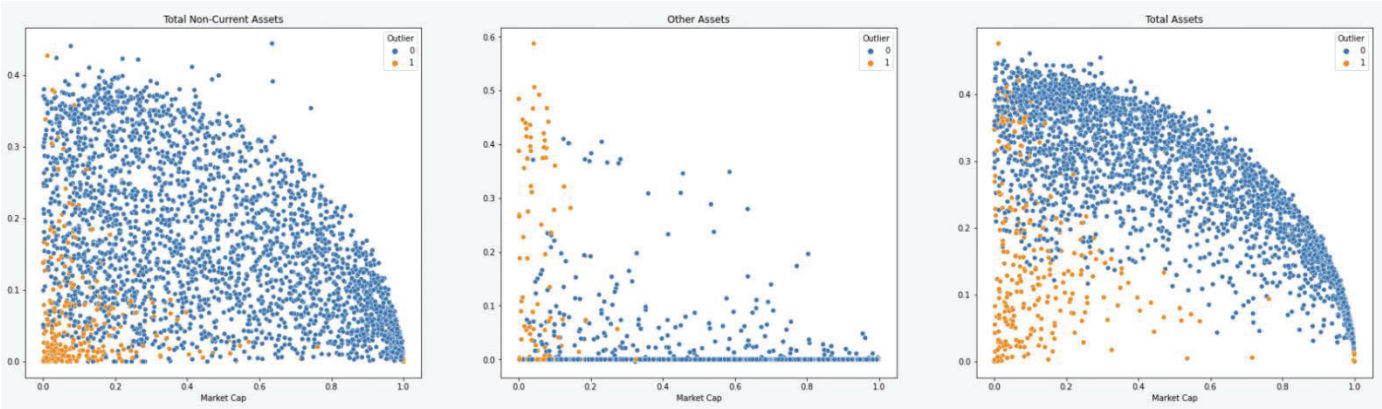

Figura 11: Gráficas de dispersión de los activos no corrientes, otros activos y el total de activos (eje Y) y la capitalización de mercado (eje X). Outliers marcados en naranja.

El conjunto de *outliers* cogido ha sido conformado por los puntos en el último decil de distancia respecto del centroide. Ser un *outlier* no implica que el punto sea diferente en todas las métricas, por lo que es normal ver diagramas de dispersión donde los puntos naranjas no parecen *outliers*. Cabe destacar la presencia generalizada de *outliers* en las empresas de menor capitalización, siendo menos frecuente la presencia de *outliers* en empresas de mucha capitalización.

Una vez realizada la revisión manual de una parte del conjunto, se decide que es suficientemente representativo de datos atípicos como para eliminar el conjunto de la muestra, teniendo la mayoría errores grandes en los datos.

4.6 Análisis de datos

Se procede ahora a realizar el análisis de los datos, con la intención de entender la composición del dataset, la distribución de las variables y las relaciones de estas con el precio de cotización.

Existen multitud de empresas de diferentes tipos y características, por lo que un análisis agregado del mercado no ofrecería un imagen realista del mercado de valores. Se ha optado por el estudio desagregado en categorías más específicas, como son el tamaño, la antigüedad o el sector en el que opera la empresa.

4.6.1 Análisis por sectores

La división por sectores ayuda a separar los negocios según factores comunes, dando una imagen más granular que la vista agregada del mercado. Es posible que aquellas empresas bajo un mismo sector compartan características similares y se trata de una manera de agruparlas según su tipo de actividad económica.

Composición

En la Figura 12 se puede ver la composición del dataset por sectores, donde se observa preponderancia de los sectores industriales, tecnológicos, médicos y consumo cíclico. Se aprecia entonces que el dataset no está balanceado en cuanto a sectores se refiere, por lo que los datos agregados estarán sesgados hacía los sectores que más pesen. Esto es relevante a la hora de elaborar los modelos predictivos, pues se debe tener en cuenta con qué datos se entrenará el modelo.

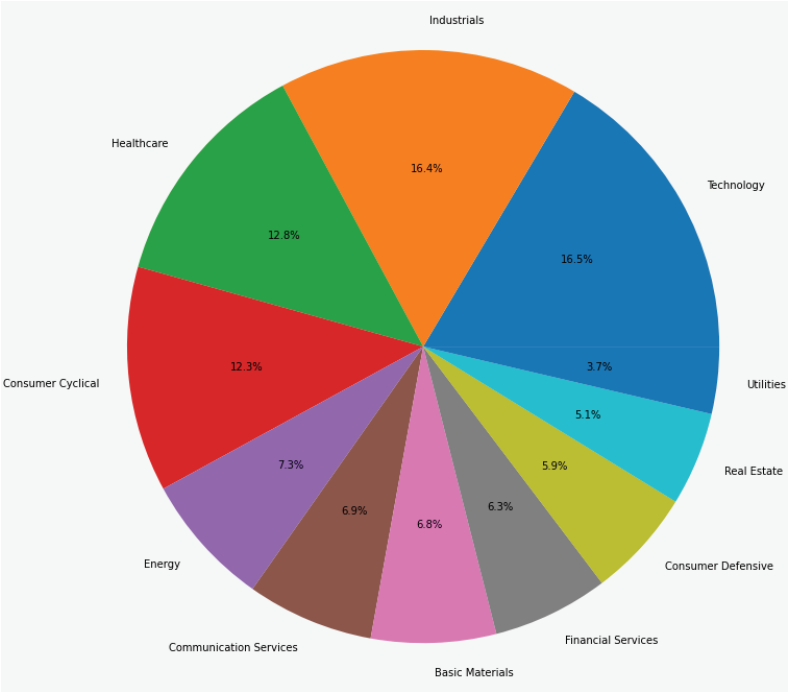

Figura 12: Gráfico de sectores. Composición del dataset por sectores.

Errores

Es de utilidad conocer la distribución de errores según la categoría estudiada. En la Tabla 2 se observa la descripción estadística de la variable "R - Percentage Missing", que contiene el porcentaje de datos faltantes después de efectuar la reconstrucción (sin el *fix trivial*). Se aprecia que no existen grandes diferencias respecto a la distribución de errores entre los sectores, a excepción del sector del "Real Estate" (bienes raíces) que tiene el valor máximo considerablemente más elevado; lo cual puede indicar la presencia de datos atípicos.

Se debe tener en cuenta que se está estudiando un dataset ya procesado, por lo que no es representativo de los datos recolectados. Existe la posibilidad de que algún sector fuera especialmente problemático pero que, debido a la limpieza y a que ya se han eliminado esos datos, no se aprecie la diferencia respecto del resto de sectores en este momento.

Tabla 2 Descripción estadística del porcentaje de fallos por sector después de la reconstrucción.

Sector	count	mean	std	min	25%	50%	75%	max
Basic Materials	181.0	0.14	0.04	0.06	0.11	0.14	0.17	0.25
Communication Services	184.0	0.13	0.03	0.07	0.11	0.13	0.15	0.23
Consumer Cyclical	330.0	0.14	0.04	0.06	0.12	0.14	0.17	0.25
Consumer Defensive	159.0	0.14	0.04	0.06	0.12	0.14	0.17	0.29
Energy	183.0	0.16	0.04	0.08	0.13	0.15	0.19	0.28
Financial Services	157.0	0.17	0.05	0.08	0.14	0.16	0.19	0.34
Healthcare	343.0	0.15	0.04	0.06	0.12	0.14	0.17	0.31
Industrials	436.0	0.15	0.04	0.06	0.12	0.15	0.18	0.26
Real Estate	135.0	0.17	0.05	0.08	0.13	0.17	0.19	0.47
Technology	442.0	0.14	0.04	0.05	0.12	0.14	0.17	0.28
Utilities	97.0	0.14	0.04	0.05	0.12	0.15	0.17	0.21

Rentabilidad

Respecto al desempeño en bolsa de los diferentes sectores se observa en la Tabla 3 que todos cuentan con un crecimiento anual compuesto promedio positivo. El crecimiento anual compuesto mediano también es positivo para todos los sectores con rentabilidades entre el 7% y el 12% anual. Esto puede considerarse como una diferencia fundamental de la bolsa respecto de los casinos, pues no hay casino con una rentabilidad agregada positiva.

Desde el punto de vista de la rentabilidad mediana, los sectores más atractivos en nuestro dataset han sido históricamente los sectores de “Utilidades”, “Servicios Financieros” y “Bienes raíces”. Los menos rentables han sido los sectores de “Energía”, “Servicios de comunicación” y “Materiales básicos”.

Recordando las dos definiciones de riesgo expuestas durante el apartado de investigación, se pueden también definir los sectores más seguros.

Entendiendo riesgo como volatilidad, los sectores con menos riesgo han sido los sectores de “Utilidades”, “Bienes raíces” y “Consumo Cíclico”, por su menor desviación típica.

Utilizando la otra definición de riesgo expuesta (riesgo como la probabilidad de pérdida permanente de capital), podría interpretarse que aquellos sectores más seguros son aquellos en los que el mínimo es más alto. Es decir, que los sectores donde los percentiles más bajos presenten mayor rentabilidad son aquellos que presentan un menor riesgo. Mediante esta interpretación los sectores más seguros son “Utilidades”, “Servicios Financieros”, “Bienes raíces” y “Consumo cíclico”. El sector de “Servicios de comunicación” tiene uno de los mínimos con mejor rentabilidad, sin embargo, palidece en la rentabilidad de su percentil 25.

Para un inversor experto que fuera capaz de seleccionar de manera continuada acciones en el percentil 75, los sectores más rentables habrían sido “Servicios financieros”, “Salud” y “Tecnología”.

Según los datos recopilados los mejores sectores para invertir habrían sido “Utilidades”, “Bienes raíces”, “Consumo Cíclico” y “Servicios financieros” teniendo en cuenta su alta rentabilidad mediana y promedio, así como su seguridad (desde ambas definiciones de riesgo).

Tabla 3 Descripción estadística de la variable "CAGR with divs" por sectores.

	count	mean	std	min	25%	50%	75%	max
Sector								
Financial Services	170.0	0.12	0.17	-0.52	0.06	0.12	0.19	1.06
Utilities	98.0	0.11	0.07	-0.27	0.10	0.12	0.14	0.31
Real Estate	136.0	0.11	0.13	-0.33	0.05	0.11	0.16	0.83
Technology	441.0	0.10	0.19	-0.70	0.01	0.09	0.17	0.97
Consumer Cyclical	330.0	0.09	0.13	-0.46	0.03	0.09	0.16	0.84
Industrials	440.0	0.09	0.19	-0.56	0.03	0.10	0.14	2.97
Healthcare	343.0	0.09	0.22	-0.61	-0.02	0.09	0.17	1.52
Communication Services	186.0	0.08	0.18	-0.31	0.01	0.08	0.15	1.05
Consumer Defensive	159.0	0.08	0.16	-0.58	0.02	0.10	0.14	0.72
Basic Materials	182.0	0.07	0.14	-0.57	0.01	0.09	0.14	0.54
Energy	195.0	0.02	0.16	-0.63	-0.06	0.07	0.12	0.58

Factores Fundamentales

Durante el análisis de los factores fundamentales se usará el top 20% de las empresas ordenadas por capitalización de mercado según la categoría correspondiente. Esto se debe principalmente a dos motivos: el primero es que según la teoría económica el mercado es más eficiente para empresas de mayor capitalización; la segunda es que en un apartado posterior se comprobará como las acciones de mayor capitalización tienen menos datos faltantes.

Se procederá a estudiar la correlación de los factores fundamentales con el precio de cierre, esto es, ignorando dividendos. Cabe destacar que correlación no implica causalidad, por lo que el estudio de la relación causa-efecto queda para trabajos posteriores. Se omitirán de la memoria aquellos resultados que no sean relevantes, sin embargo, todos los resultados pueden reproducirse por el usuario particular con las instrucciones de esta memoria. Estas correlaciones por sectores pueden encontrarse en la Tabla 4.

Nota: Se consideran correlaciones fuertes aquellas con coeficientes superiores a 0.75 o menores a -0.75.

Tabla 4 Correlaciones de fundamentales con el precio por sector

Sector	Correlaciones Positivas Fuertes	Correlaciones Negativas Fuertes
Utilidades	- Inmovilizado Material (Propiedades, plantas y equipos) - Activos Tangibles - Activos Totales No Corrientes - Valor en Libros por Acción - Capital Empleado - Total de Pasivos y Patrimonio - Activos Totales - Ganancias Retenidas - Valor en Libros Tangible - PIB de Estados Unidos - Índice de precios de la vivienda para todas las transacciones de EE. UU	- Gastos de capital - Dividendos pagados - Promedio de tasas hipotecarias fijas de EE. UU. a 30 años - Rendimiento promedio del Tesoro a 10 años
Consumo Cíclico	- Ingresos por acción - Activos Tangibles - Activos Totales - Ingresos - Total de Pasivos y Patrimonio - "Cash Sales" - Total de Pasivos - Capital Invertido - Beneficio Bruto - PIB de Estados Unidos - PIB de China - Población de Estados Unidos - Población Mundial - Población de China - Índice de precios de la vivienda para todas las transacciones de EE. UU.	- Ninguna significativa
Servicios Financieros	- Ingresos - "Cash Sales" - Beneficio operativo por acción - Ingresos Retenidos - Capital Invertido - Patrimonio Total de los accionistas - Total de Pasivos y Patrimonio - Ingresos por acción - Activos Tangibles - Beneficio Operativo - Beneficio Neto - EBITDA - Valor en libros - Capital Empleado - Flujo de caja libre - Beneficios por acción (diluidos) - NOPAT - Beneficio Bruto - Flujo de caja libre por acción	- Dividendos por acción.
Salud	- Ingresos por acción - Beneficio Bruto - Activos totales - PIB de Estados Unidos	- Gasto de capital

Materiales Básicos	- Ingresos por acción - Beneficio Operativo - EBITDA - Beneficio Neto por acción - Activos Totales	- Ninguna significativa
Consumo Defensivo	- Activos Totales - PIB de Estados Unidos	- Gastos en capital - Promedio de tasas hipotecarias fijas de EE. UU. a 30 años - Rendimiento promedio del Tesoro a 10 años
Industriales	- Valor en libros por acción - Ingresos por acción - Beneficio operativo por acción - Activos totales - EBITDA - Beneficio Neto por acción - PIB de Estados Unidos - Índice de precios de la vivienda para todas las transacciones de EE. UU.	- Promedio de tasas hipotecarias fijas de EE. UU. a 30 años - Rendimiento promedio del Tesoro a 10 años - Dividendos por acción
Tecnología	- Ingresos por acción - Beneficio Bruto - Beneficio Operativo Ajustado - Flujo de caja libre por acción - Activos totales	- Ninguna significativa
Bienes Raíces	- Beneficio Bruto - Ingresos - EBITDA - Índice de precios de la vivienda para todas las transacciones de EE. UU. - Beneficio Neto - PIB de Estados Unidos - Depreciación y amortización - Activos Tangibles - Activos Totales	- Dividendos por acción - Promedio de tasas hipotecarias fijas de EE. UU. a 30 años
Servicios de Comunicación	- Activos Totales - Índice de precios de la vivienda para todas las transacciones de EE. UU. - Ingresos - Activos Tangibles - Beneficio Bruto - PIB de Estados Unidos	- Ninguna significativa
Energía	- Valor en libros por acción	- Gastos en capital

Según las correlaciones aquí expuestas, se puede observar la alta correlación de algunas variables compartida en casi todos los sectores. Estas variables es posible que tomen un papel fundamental a la hora de desarrollar los modelos predictivos, debido a su estrecha relación con el precio independientemente del sector. Las variables más destacadas son:

- Activos totales (o variables semejantes)
- Beneficio operativo (o variables semejantes)
- Dividendos
- Gastos en capital
- PIB de EEUU
- Índice de precios de la vivienda para todas las transacciones de EE. UU.
- Promedio de tasas hipotecarias fijas de EE. UU. a 30 años

En resumen, parece que las principales variables correlacionadas al precio de las acciones son los activos que posee la empresa, los ingresos que genera, el dinero que se reparte en dividendos, sus gastos en capital y la macroeconomía. Cuando se realicen los modelos se revisarán estas relaciones, pues es probable que haya factores fundamentales que no se estén teniendo en cuenta; además se debe recordad que casualidad no implica causalidad.

Factores macroeconómicos

Se ha podido comprobar la alta correlación de la macroeconomía con los precios de cotización de las acciones, por lo que parece sensato dedicar una sección a este análisis. Debido a complejidad de las relaciones macroeconómicas se recomienda mayor cautela al sacar conclusiones, sin embargo, teniendo siempre presente la posibilidad de que las explicaciones aportadas no sean suficientes y dando por hecho que se requiere de un mayor estudio; se proceden a formular las hipótesis sobre las relaciones macroeconómicas.

Llama la atención la relación de varios sectores con las viviendas, tanto positiva con el precio de la vivienda, como negativa con los tipos fijos para las hipotecas. Una bajada de los tipos fijos hipotecarios incentiva a la compra de casas, aumentando la demanda y por tanto los precios; por lo que la relación es coherente. Sin embargo, la relación de varios de los sectores y la vivienda puede explicarse mejor por medio del rendimiento promedio del tesoro a 10 años. La causa que explica todas las relaciones es el abaratamiento del crédito y la facilidad para adquirir activos en un entorno de bajos tipos de interés. Si la hipótesis es correcta, son los bajos tipos lo que estimula la compra de activos y el aumento de la capitalización. El inversor individual debería tener cuidado en entornos con subidas de tipos de interés, pues un endurecimiento del crédito perjudicará a los sectores relacionados, es decir, aquellos más endeudados.

Para desarrollar la hipótesis sería positivo comprobar el rendimiento de las acciones de estos sectores en épocas de tipos de interés crecientes. Esta etapa podría corresponder a los años desde 1963 hasta aproximadamente 1982 como se puede ver en la Figura 13. Sin embargo, no se dispone de acciones suficientes en estos sectores sobre las que se tengan datos en ese periodo, por lo que queda para un trabajo futuro demostrar o falsar la teoría.

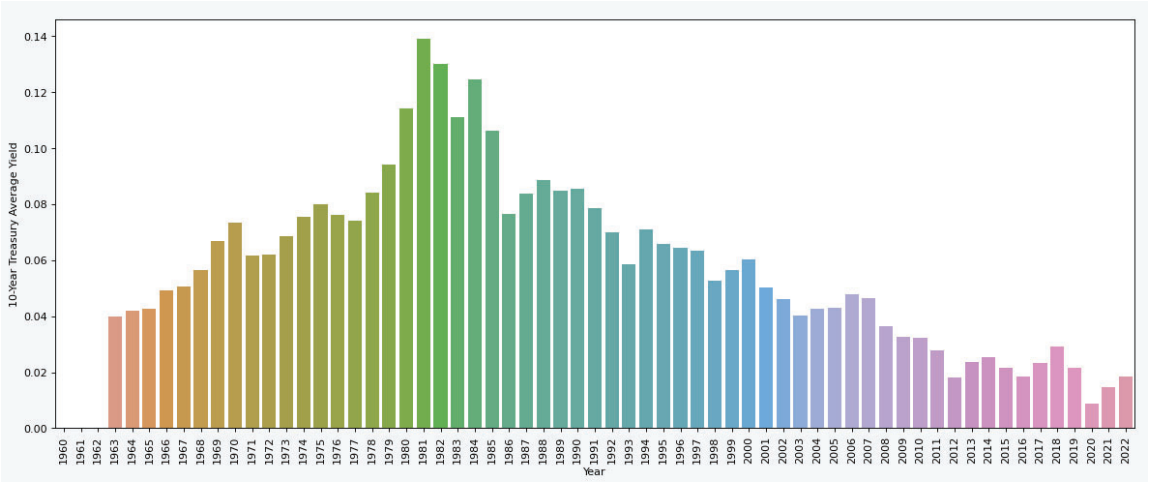

Figura 13: Gráfico de barras del rendimiento promedio del Tesoro a 10 años

Ineficiencias de mercado

La literatura actual sostiene que el mercado es más eficiente para las grandes empresas, es decir, que se fija su precio correcto con mayor velocidad y precisión. Por el contrario, sostiene que en las empresas pequeñas pueden existir ineficiencias de mercado, y se tarda más tiempo en que estas lleguen a su valor intrínseco. Esto sucede debido a que la mayoría de los bancos y grandes instituciones financieras no pueden invertir en pequeños mercados, por lo que se centran solo en comprar y vender en los grandes.

Durante el cálculo de las correlaciones por sector, se ha observado una gran diferencia respecto de las correlaciones al coger las empresas más pequeñas o tan solo las más grandes. En el conjunto de las empresas más pequeñas las correlaciones altas desaparecen en gran medida, mientras que al coger el conjunto más grandes si se pueden vislumbrar factores de alta correlación. Este fenómeno puede explicarse por las ineficiencias de mercado ya mencionadas. Puede que el precio de las acciones más pequeñas, al no ser eficiente, esté más relacionado con factores psicológicos de mercado que con los factores fundamentales aquí estudiados. También cabe destacar que la falta de correlaciones en los datos de empresas con menor capitalización venga dada por una peor calidad de los datos.

Este hecho no es definitivo, pero presenta una posibilidad atractiva. Mayores ineficiencias de mercado pueden permitir al inversor individual (que sí puede invertir en acciones pequeñas) comprar empresas excepcionalmente baratas durante largos periodos de tiempo.

Principales conclusiones:

- El mercado parece ser más ineficiente con las empresas pequeñas.
- Existen una serie de datos fundamentales (ingresos, activos, dividendos…) con alta correlación con el precio que podría ser sensato tener en cuenta a la hora de investigar empresas.
- La macroeconomía parece tener un papel importante en el desempeño de algunos sectores (normalmente aquellos con su precio altamente correlacionado con los activos).

- Según los datos recopilados los mejores sectores para invertir habrían sido “Utilidades”, “Bienes raíces”, “Consumo Cíclico” y “Servicios financieros” teniendo en cuenta su alta rentabilidad mediana y promedio, así como su seguridad (desde ambas definiciones de riesgo).

4.6.2 Análisis por industrias

El estudio según industrias es más complicado, pues existen 146 industrias diferentes en nuestro conjunto de datos y no se dispone de una muestra de empresas suficientemente grande para el estudio de todas. Un estudio individual de cada industria sería enormemente relevante y seguramente revelador, sin embargo, esto escapa al alcance del proyecto y requeriría de más datos y tiempo.

Composición

Como se puede observar en la Tabla 5 las industrias más representadas en los datos son el software (tanto aplicación como infraestructura) y la biotecnología, representando la más común tan solo un 3.5% del total. La distribución de industrias está menos concentrada que la distribución en sectores y se tiene una muestra más pequeña para el estudio individual de cada una. El número de acciones por industria no es una muestra significativa para extraer conclusiones de cada industria individual (más si se tiene en cuenta que de algunas industrias la muestra es de tan solo 1 o 2 acciones), por lo que dificulta un estudio estadísticamente relevante.

Tabla 5 Top 20 industrias por número de acciones en el dataset.

Industry	
Software—Application	93
Biotechnology	87
Software—Infrastructure	74
Oil & Gas E&P	60
Medical Devices	59
Specialty Industrial Machinery	57
Telecom Services	52
Specialty Chemicals	50
Communication Equipment	50
Information Technology Services	48
Packaged Foods	48
Asset Management	46
Semiconductors	45
Aerospace & Defense	44
Oil & Gas Equipment & Services	44
Utilities—Regulated Electric	42
Drug Manufacturers—Specialty & Generic	40
Oil & Gas Midstream	40
Restaurants	39
Diagnostics & Research	38

Cabe destacar que durante la detección de *outliers* se menciona que una gran cantidad de estos provenían de la industria de la biotecnología, esto podría explicarse por mera probabilidad, pues se trata de una de las industrias más

frecuentes. Esto presenta una buena oportunidad para ver la distribución de errores según la industria, análogo a lo realizado con los sectores.

Errores

Se observa en Tabla 6 que la industria de la biotecnología no presenta una distribución muy diferente a la del resto de industrias, por lo que la hipótesis de que se existen un mayor número de *outliers* en esta industria pierde fuerza. Sin embargo, se debe recordar que el número de datos faltantes es tan solo una aproximación de la calidad de los datos, y puede darse el caso de que no le falten datos a la empresa, sin embargo, estos no sean válidos.

Tabla 6 Descripción estadística de los errores por industrias después de la reconstrucción. Datos de las 20 industrias más comunes en el dataset.

	count	mean	std	min	25%	50%	75%	max
Industry								
Software—Application	93.0	0.15	0.03	0.08	0.12	0.15	0.17	0.26
Biotechnology	87.0	0.17	0.05	0.07	0.14	0.16	0.20	0.31
Software—Infrastructure	74.0	0.13	0.03	0.06	0.12	0.13	0.15	0.24
Oil & Gas E&P	60.0	0.18	0.04	0.10	0.16	0.19	0.20	0.28
Medical Devices	59.0	0.15	0.03	0.09	0.13	0.14	0.16	0.24
Specialty Industrial Machinery	57.0	0.14	0.04	0.08	0.12	0.14	0.16	0.23
Telecom Services	52.0	0.13	0.03	0.08	0.10	0.13	0.15	0.23
Communication Equipment	50.0	0.14	0.04	0.06	0.12	0.14	0.17	0.24
Specialty Chemicals	50.0	0.14	0.04	0.07	0.11	0.13	0.17	0.22
Information Technology Services	48.0	0.14	0.04	0.05	0.11	0.13	0.17	0.21
Packaged Foods	48.0	0.15	0.05	0.09	0.12	0.14	0.17	0.29
Asset Management	46.0	0.19	0.07	0.08	0.14	0.16	0.20	0.34
Semiconductors	45.0	0.13	0.03	0.08	0.11	0.13	0.15	0.23
Aerospace & Defense	44.0	0.16	0.04	0.09	0.13	0.16	0.19	0.24
Utilities—Regulated Electric	42.0	0.14	0.03	0.09	0.11	0.14	0.16	0.21
Oil & Gas Equipment & Services	41.0	0.15	0.04	0.09	0.12	0.15	0.16	0.25
Drug Manufacturers—Specialty & Generic	40.0	0.14	0.04	0.06	0.11	0.14	0.16	0.23
Restaurants	39.0	0.14	0.03	0.08	0.12	0.13	0.16	0.23
Diagnostics & Research	38.0	0.13	0.03	0.07	0.11	0.13	0.16	0.22
Internet Content & Information	36.0	0.12	0.03	0.08	0.10	0.11	0.13	0.18

Rentabilidad

En la Tabla 7 se observa la descripción estadística del crecimiento anual compuesto de los precios de cotización de las 20 industrias más rentables según su mediana.

Las industrias más rentables según la mediana han sido “Operaciones de Infraestructura”, “REIT—Industrial”, “Conglomerados Financieros” y “Datos financieros y bolsas de valores”. Como ya se ha mencionado, para algunas industrias se tienen muy pocas empresas, esto se aprecia en el hecho que las industrias mencionadas cuentan con solo 1, 8, 2 y 10 empresas respectivamente. Lo pequeña que es la muestra pone en tela de juicio las conclusiones, por lo que se recomienda un mayor estudio a futuro.

Si se juzga por la media, las industrias más rentables han sido “Operaciones de Infraestructura”, “Fabricantes de automóviles”, “Aeropuertos y Servicios Aéreos” y “Aerolíneas”.

Aunque no se observan todas en la tabla, las industrias con menor desviación típica han sido “Pastelería”, “Aluminio”, “Madera y producción de madera”, “Utilidades: productores de energía independientes” y “Bebidas—Cerveceros”.

De igual manera, las industrias con el percentil 25 más bajo han sido “Servicios de educación y capacitación”, “Equipos y servicios de petróleo y gas” y “Servicios de crédito”.

Debido al tamaño de la muestra y a que no se observa una industria claramente ganadora en todos los aspectos, es más difícil sacar conclusiones sobre qué industrias son mejores para invertir. Los inversores que busquen altas rentabilidades sin importar el riesgo estarán más atraídos la industria de las infraestructuras y los fabricantes de coches, mientras que lo más conservadores se verán más complacidos con la pastelería o la cervecería.

Principales conclusiones:

- No se dispone de un conjunto de datos adecuado para estudiar las empresas por industrias.
- No parece haber gran diferencia en la distribución de errores por industria.

Tabla 7 Descripción estadística de la variable "CAGR with divs" por industrias. Top 20 industrias según rentabilidad mediana.

Industry	count	mean	std	min	25%	50%	75%	max
Infrastructure Operations	1.0	0.27	NaN	0.27	0.27	0.27	0.27	0.27
REIT—Industrial	8.0	0.20	0.09	0.08	0.17	0.19	0.24	0.34
Financial Conglomerates	2.0	0.18	0.31	-0.03	0.07	0.18	0.29	0.40
Financial Data & Stock Exchanges	10.0	0.18	0.04	0.09	0.17	0.18	0.20	0.23
Auto Manufacturers	9.0	0.27	0.27	0.06	0.08	0.17	0.30	0.84
Utilities—Renewable	9.0	0.10	0.16	-0.27	0.10	0.17	0.19	0.25
Airports & Air Services	5.0	0.22	0.20	0.08	0.12	0.17	0.18	0.58
Diagnostics & Research	38.0	0.17	0.22	-0.40	0.03	0.16	0.29	0.74
Healthcare Plans	8.0	0.17	0.07	0.11	0.12	0.16	0.19	0.32
Tobacco	6.0	0.18	0.08	0.11	0.14	0.16	0.20	0.32
Insurance—Life	4.0	0.16	0.07	0.09	0.10	0.16	0.22	0.24
Lumber & Wood Production	2.0	0.15	0.03	0.13	0.14	0.15	0.16	0.17
Retail Apparel & Specialty	1.0	0.15	NaN	0.15	0.15	0.15	0.15	0.15
Health Information Services	20.0	0.12	0.19	-0.29	0.07	0.15	0.19	0.58
REIT—Healthcare Facilities	9.0	0.15	0.06	0.05	0.12	0.15	0.17	0.27
REIT—Residential	9.0	0.20	0.11	0.13	0.13	0.14	0.20	0.40
Recreational Vehicles	10.0	0.14	0.04	0.09	0.11	0.14	0.17	0.21
Residential Construction	16.0	0.14	0.09	0.00	0.09	0.14	0.18	0.32
Utilities—Regulated Gas	13.0	0.14	0.04	0.09	0.11	0.14	0.15	0.25
Asset Management	46.0	0.13	0.13	-0.23	0.06	0.14	0.21	0.46

4.6.3 Análisis por países

Composición

La distribución de las acciones por países es bastante desigual, pues se observa que el 80.1% (2148) de las acciones del dataset pertenecen a Estados Unidos, seguido de un 4.1% (110) pertenecientes a Canadá y un 3.7% (99) pertenecientes a China. Esto coincide con lo esperado al obtener los datos de una página que ofrece mayoritariamente información sobre acciones estadounidenses y después de tener que eliminar algunas acciones que presentaban errores en la conversión de moneda. Esto limita mucho las conclusiones que se pueden extraer para diferentes países, pues esencialmente el dataset es de empresas de EEUU.

Errores

Entre los países estudiados no se observan diferencias significativas en la distribución de errores, más allá de los problemas ya mencionados derivados de la conversión de moneda.

Rentabilidad

Debido a la desigual distribución de las empresas según su país y a la falta de una muestra significativa, se analizan únicamente los tres principales países ya mencionados.

La descripción estadística del retorno se puede ver en la Tabla 8, donde se aprecian retornos similares y positivos para Canadá y Estados Unidos frente a un retorno negativo para China. Las causas de esto pueden ser múltiples, entre ellas la situación geográfica, los modelos de gobiernos o el tipo de cultura.

Cabe mencionar que durante el último periodo que se recogieron datos, el mercado parecía experimentar un miedo generalizado hacia las acciones chinas, debido a los miedos regulatorios y la posibilidad de que se eliminen de los principales mercados financieros estadounidenses. Sin embargo, siempre existen miedos en el mercado y es difícil saber cuál es el estado "natural" de las acciones; pero esta coyuntura podría explicar la diferencia en rentabilidades. Sin embargo, debido a la complejidad socioeconómica del asunto es difícil discernir una causa principal, por lo que queda para un estudio posterior y más profundo.

En cualquier caso, los datos parecen indicar que en el pasado la elección geográfica a la hora de invertir ha sido relevante y tanto Estados Unidos como Canadá parecen haber sido para los inversores mejor elección que China.

Tabla 8 Descripción estadística del retorno anual compuesto en China, Canadá y Estados Unidos.

	count	mean	std	min	25%	50%	75%	max
Country								
CN	99.0	-0.04	0.29	-0.70	-0.21	-0.06	0.14	0.93
CA	110.0	0.08	0.17	-0.61	0.01	0.09	0.17	0.68
US	2148.0	0.09	0.17	-0.63	0.02	0.10	0.15	2.97

Principales conclusiones:

- Parece que el país es un factor relevante a la hora de invertir.
- Estados Unidos y Canadá han sido en el pasado mejores países para los inversores que China.
- A pesar de lo mencionado, se requeriría de un dataset más balanceado para que las conclusiones expuestas fueran confirmadas.

4.6.4 Análisis por edad

Composición

Para el estudio de las acciones según su edad, se discretiza la variable "Years since IPO", catalogando las acciones como:

- "Joven" si Years since IPO $\in [3, 5]$[23]
- "Intermedia" si Years since IPO $\in (5, 10]$
- "Adulta" si Years since IPO $\in (10, 20]$
- "Madura" si Years since IPO $\in (20, 40]$
- "Anciana" si Years since IPO $\in (40, \infty)$

Los resultados de la distribución por edad se aprecian en la Figura 14, donde se puede ver que la gran mayoría de empresas son consideradas maduras. Cabe destacar que la variable "Years since IPO" no contiene información con mucha precisión y ha sido modificada para servir como proxy a la edad de una empresa, sin embargo, hay que tener en cuenta que las empresas existen mucho antes de salir a bolsa.

Cabe destacar el bajo número de empresas jóvenes, esto se debe a una menor franja de edad, teniendo sólo tres años válidos para ser consideradas jóvenes frente intervalos de 5, 10 o 20 años. Esta discretización es arbitraria y sería interesante el estudio de otras distribuciones en el futuro.

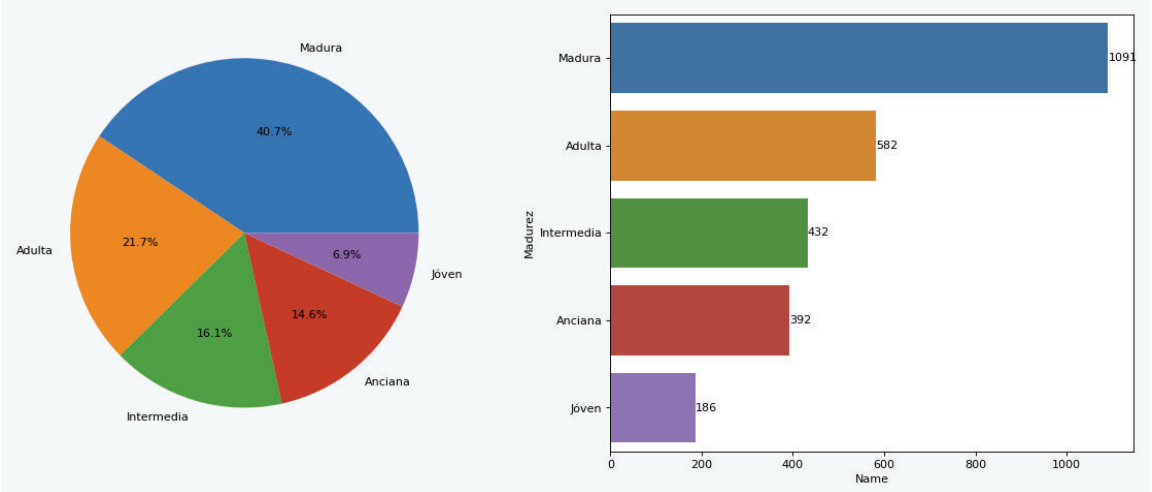

Figura 14: Composición del dataset según su madurez.

Errores

Al comprobar la distribución de errores según la edad de la empresa, no parece que exista ninguna diferencia significativa.

[23] El mínimo de edad en nuestro dataset es de 3 años debido a la decisión de eliminar aquellas empresas con menos años durante la limpieza de datos.

Rentabilidad

Estudiando la rentabilidad anual compuesta según su grado de madurez se plantea un problema interesante. Dado que el crecimiento anual compuesto está ajustado a la variable tiempo, que este sea alto requiere de mantener retornos elevado a lo largo del tiempo.

Mantener retornos extremadamente altos a lo largo del tiempo es complicado. Asumiendo que la rentabilidad de la empresa en bolsa en el largo plazo depende del desempeño del negocio subyacente, existen fuerzas económicas que provocan una regresión a la media.

Si una empresa cosecha grandes rentabilidades en un mercado libre, esto atrae a competidores intentando obtener ganancias similares; lo que reduce la rentabilidad para la empresa originaria. Por otro lado, buenos desempeños empresariales suelen ir acompañadas de un gran crecimiento, por lo que a medida que se vuelve más difícil crecer (esto puede deberse a motivos como saturación del mercado) también se vuelve más difícil mantener el ritmo de generación de ingresos.

Por otro lado, en la economía capitalista existe un sesgo de supervivencia. Solo se ven hoy las empresas que en el pasado fueron exitosas y sobrevivieron. Debido a que los datos recolectados solo incluyen empresas que siguen existiendo, aquellas empresas que en el pasado tuvieron rendimientos extremadamente malos y que, probablemente acabaron quebrando, no aparecen en nuestro dataset.

Estos dos sesgos, provocan que los retornos en el largo plazo sufran menos variaciones (menor desviación típica en la Tabla 9). Este fenómeno puede visualizarse de manera intuitiva en la Figura 15.

Respecto al rendimiento (basándonos en la mediana), entre las categorías de “Intermedia”, “Madura” y “Adulta” (esto es el intervalo de los 5 a 40 años) no parece haber diferencias significativas. Sin embargo, los extremos sí presentan comportamientos diferentes, teniendo las empresas jóvenes un desempeño peor y las empresas ancianas un desempeño mejor. Cabe entonces estudiar los beneficios de invertir en negocios ancianos con modelos de negocio ya probados. Posiblemente este hecho guarde relación con el efecto Lindy [52], expuesto por el filósofo, matemático y economista Nassim Nicholas Taleb.

El efecto Lindy sostiene que cuanto más tiempo sobrevive una tecnología o un elemento no perecedero , más probable es que sobreviva en el futuro. Esta teoría propondría entonces que es más probable que se siga escuchando a Beethoven dentro de 100 años, que a un músico de esta década. Aplicado al campo de la inversión, se podría interpretar como que es más probable que dentro de 100 años exista una empresa como General Electric (fundada en 1892) que Zoom Video Communications (fundada en 2011).

Durante el estudio de las categorías anteriores se ha tratado el riesgo, sin embargo, eso no es posible en esta categoría. Se podría decir que las empresas más seguras para ambas definiciones de riesgo son las ancianas, pero como ya se ha comentado, la muestra está sesgada a su favor, por lo que debe estudiarse de manera apartada en un futuro.

Tabla 9 Descripción estadística del crecimiento anual compuesto según grado de madurez.

Madurez	count	mean	std	min	25%	50%	75%	max
Anciana	391.0	0.11	0.06	-0.21	0.08	0.11	0.14	0.36
Intermedia	432.0	0.09	0.22	-0.60	-0.03	0.09	0.22	0.95
Madura	1090.0	0.09	0.14	-0.56	0.03	0.09	0.14	2.97
Adulta	581.0	0.07	0.16	-0.54	-0.02	0.08	0.16	0.83
Jóven	186.0	0.07	0.34	-0.70	-0.15	0.06	0.24	1.52

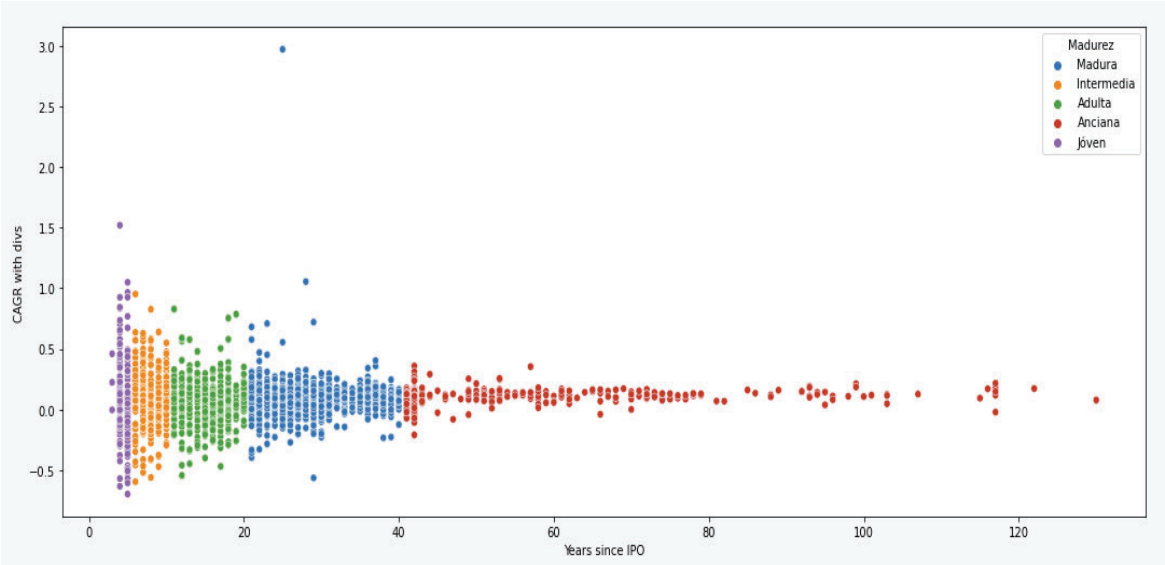

Figura 15: Gráfico de dispersión de los años desde su salida a bolsa frente a la rentabilidad anual compuesta. Puntos coloreados por categoría de madurez.

Principales conclusiones:

- Los retornos de las empresas en el largo plazo parecen sufrir regresión a la media.
- Las empresas ancianas podrían ser la opción más segura a la hora de invertir.
- Es necesario un mayor estudio y el uso de un dataset que incluya empresas que quebraron en el pasado, para evitar el sesgo de supervivencia.

4.6.5 Análisis por tamaño

Composición

Para el estudio de las acciones según su tamaño, se discretiza la variable "Market Cap", catalogando las acciones como:

- "Micro" si Market Cap $\in (0, 300\ millones]$
- "Small" si Market Cap $\in (300\ millones\ , 2000\ millones]$
- "Medium" si Market Cap $\in (2000\ millones, 10\ 000\ millones]$
- "Large" si Market Cap $\in (10\ 000\ millones,\ 200\ 000\ millones]$
- "Mega" si Market Cap $\in (200\ 000\ millones, \infty)$

Los resultados de la distribución se aprecian en Figura 16. Se observa que el dataset se encuentra bastante balanceado en este aspecto, a excepción de la falta de empresas de más de 200 000 millones. La discretización en este caso no es arbitraria y se basa en el consenso habitual de los tamaño de empresas según su capitalización de mercado.

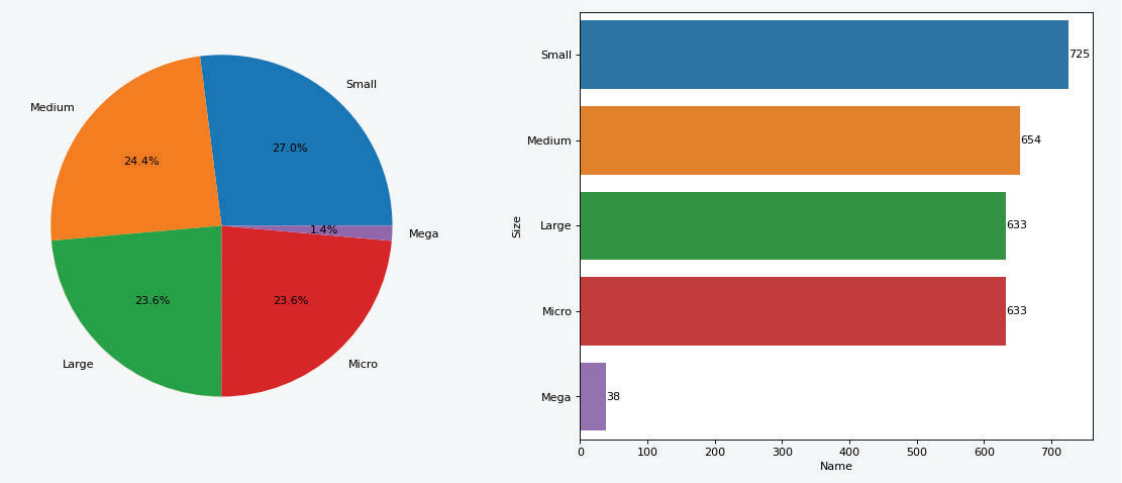

Figura 16: Distribución de las acciones según su tamaño.

Errores

Comprobando la distribución de errores según el tamaño de la empresa en la Tabla 10, se observa que tanto la media, como la mediana, como la desviación típica de los errores después de la reconstrucción decrece ligeramente conforme aumenta el tamaño de las empresas. Esta relación inversa se muestra en la Figura 17.

Tabla 10 Distribución de los errores según el tamaño de las empresas después de realizar la reconstrucción.

Size	count	mean	std	min	25%	50%	75%	max
Micro	625.0	0.17	0.04	0.07	0.14	0.17	0.20	0.34
Small	710.0	0.15	0.04	0.05	0.12	0.15	0.18	0.47
Medium	645.0	0.14	0.04	0.05	0.12	0.14	0.16	0.31
Large	629.0	0.13	0.03	0.06	0.11	0.13	0.15	0.23
Mega	38.0	0.12	0.03	0.07	0.11	0.12	0.14	0.18

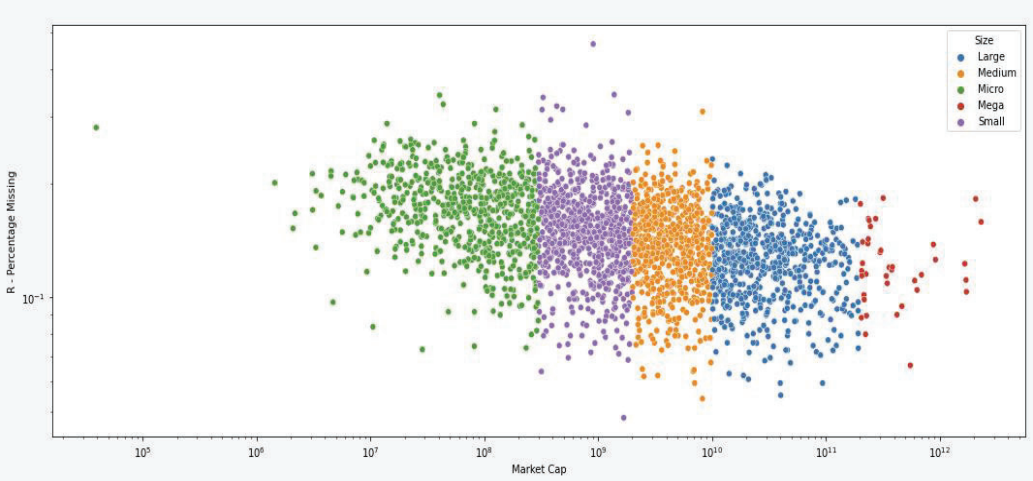

Figura 17: Gráfica de dispersión entre las variables "Market Cap" y "R - Percentage Missing". Ejes logarítmicos. Puntos coloreados según su categoría de tamaño.

Esto refuerza la hipótesis expuesta anteriormente de que la calidad de los datos viene afectada por el tamaño de las empresas. Puede ser sensato entonces utilizar las acciones de mayor capitalización a la hora de extraer conclusiones (teniendo en cuenta el sesgo que esto supone) para evitar datos de mala calidad.

Rentabilidad

Al igual que durante el estudio del tiempo de vida de las empresas, es complicado extraer conclusiones satisfactorias basadas en los últimos datos en lo referente al tamaño. Aquellas empresas que tuvieron un rendimiento satisfactorio en el pasado son más grandes en la actualidad, creando una falsa impresión de que, a mayor tamaño, mayor rentabilidad.

Es decir, para el estudio efectivo de la rentabilidad según el tamaño, habría de usarse el tamaño que tenían las empresas hace años. Esto es lo que se realizará en la siguiente categoría, tomando como punto de partida el año 2003.

Principales conclusiones:

- Parece que la calidad de los datos en el dataset está relacionada con el tamaño de las empresas.

4.6.6 Análisis por tamaño en 2003

El estudio del tamaño en 2003 de las empresas es especialmente relevante cuando se quiere relacionar el tamaño con la rentabilidad, pues no es posible realizar el estudio en base al tamaño actual. Se ha escogido el año 2003 puesto que fechas anteriores presentan una menor cantidad de datos y, en fechas cercanas como el año 2001 y 2002, el mercado se encontraba deshinchándose masivamente debido a la burbuja tecnológica del año 2000.

Composición

Para el estudio de la composición según el tamaño en 2003 se discretizará esta variable según las mismas reglas usadas en la discretización del tamaño actual, ya expuesto en el apartado anterior.

En la Figura 18 se observa la distribución por tamaño de las empresas estudiadas en el año 2003. Llama la atención que la distribución es fundamentalmente diferente a la actual, preponderando las empresas pequeñas y medianas, sin ninguna empresa "Mega". Los motivos que explican este fenómeno son principalmente:

1. Un crecimiento con el tiempo del tamaño de las empresas. Empresas tan grandes como Amazon, Tesla o Apple son fenómenos relativamente nuevos, pues ni tan siquiera las empresas más grandes del pasado llegaron a esas capitalizaciones[24].
2. El sesgo de supervivencia del que ya se ha hablado. Fueron muchas las empresas enormes que quebraron durante la burbuja del año 2000, sobre las cuales no se tienen datos.

Por supuesto, que sería posible una diferente distribución por otro motivo no mencionado aquí, por lo que sería interesante estudiarlo a futuro. Sin embargo, aunque la distribución sea diferente, se tiene una muestra relevante para realizar el estudio de rentabilidad.

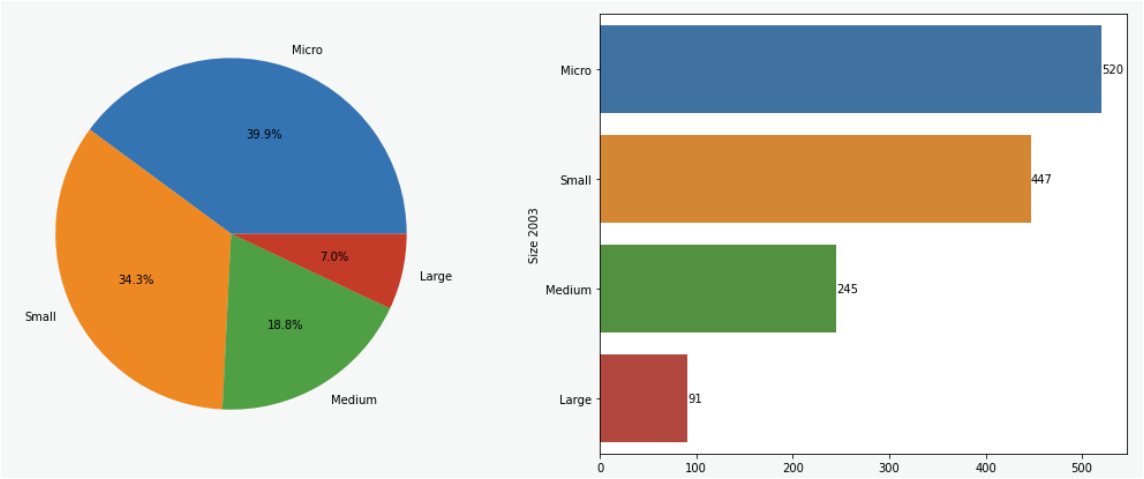

Figura 18: Distribución de las acciones según su tamaño en 2003.

[24] Cabe destacar que esto es cierto en términos nominales, sin embargo, ajustando a la inflación se podría encontrar que esto no es cierto en términos reales. De cualquier manera, es una explicación válida para el fenómeno que observamos en la gráfica.

Errores

Antes de proceder al estudio de la rentabilidad cabe destacar la distribución de errores. Se vuelve a encontrar una correlación negativa entre el tamaño de la empresa y el número de errores, sin embargo, esto era de esperar pues también existe una correlación positiva entre su tamaño en 2003 y su tamaño actual (como se ve en la Figura 19).

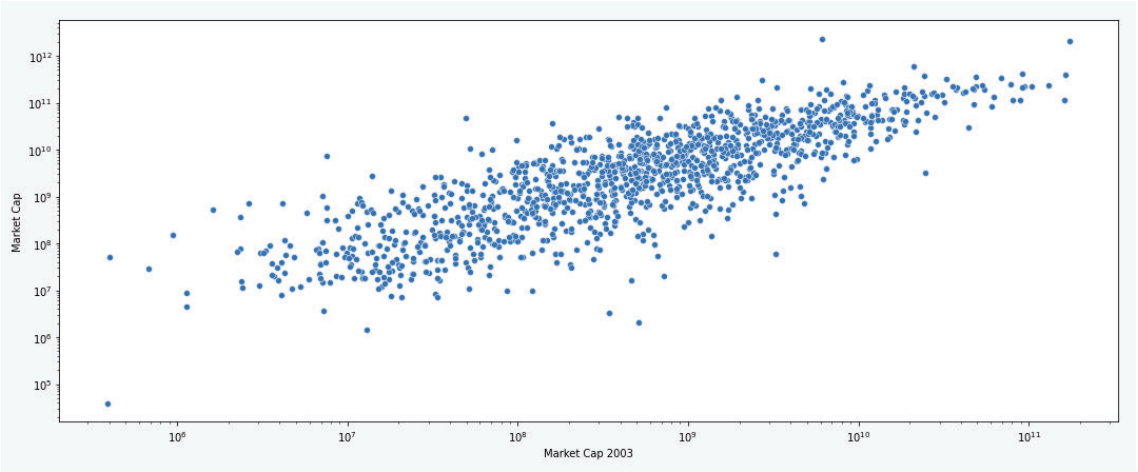

Figura 19 : Diagrama de dispersión entre la capitalización de mercado en 2003 y en la actualidad.

Rentabilidad

Analizando la rentabilidad por su tamaño en 2003 en la Tabla 11 se observa un hecho que contraviene lo establecido. Actualmente se piensa que las acciones de baja capitalización tienen mejor desempeño que las acciones de gran capitalización en el largo plazo, sin embargo, en el periodo estudiado se ha dado el caso contrario. Se observa que las acciones más rentables han sido las más grandes, mientras que las menos rentables han sido las más pequeñas. Nótese que el grupo "NaN", representa las acciones de las cuales no se tienen datos en 2003 y se ha mantenido su grupo como referencia.

En términos de riesgo las acciones de gran capitalización también parecen mejor opción, pues han tenido menor desviación típica y mínimos menos pronunciados.

La explicación para las conclusiones obtenidas es el hecho de que se está estudiando un solo periodo. Existen periodos donde las empresas de gran capitalización lo hacen mejor que aquellas de pequeña capitalización, pero también existen periodos donde se cumple lo opuesto [53]. Por este motivo encontrar un periodo concreto para un conjunto de empresas concreto no presenta una contraposición a los hechos ya recabados. Sin embargo, sería interesante contrastar los hechos del periodo actual con otros trabajos estudiando el mismo periodo.

Tabla 11 Descripción estadística del crecimiento anual compuesto según el tamaño en 2003.

	count	mean	std	min	25%	50%	75%	max
Size 2003								
Large	91.0	0.13	0.05	0.02	0.10	0.12	0.15	0.27
Medium	245.0	0.12	0.07	-0.17	0.08	0.12	0.16	0.47
Small	446.0	0.10	0.07	-0.28	0.07	0.11	0.15	0.29
NaN	1378.0	0.08	0.23	-0.70	-0.04	0.08	0.18	2.97
Micro	520.0	0.06	0.09	-0.40	0.01	0.07	0.12	0.36

Principales conclusiones:

- La distribución de las empresas según su tamaño parece haber cambiado sustancialmente con el tiempo, aunque se requiere un mayor estudio.
- Parece que el periodo estudiado coincide con un periodo de en el que las empresas de mayor capitalización tuvieron mejores rendimientos que aquellas de menor capitalización.

4.6.7 Análisis por rentabilidad histórica

Composición

Con el objetivo de comprender qué factores son aquellos que distinguen a la empresas "ganadoras" de las "perdedoras", es interesante estudias el conjunto de datos según su rentabilidad. Para esto, se discretiza la variable "CAGR with divs", catalogando las acciones según su rentabilidad. Como ya se ha explicado lo "buena" o "mala" que es la rentabilidad depende enteramente del coste de oportunidad, es decir, del rendimiento relativo del resto de acciones ya activos. Por este motivo, se separan las acciones según los percentiles de su rentabilidad contando dividendos.

- Rentabilidad "Muy Mala" si es inferior o igual al percentil 20.
- Rentabilidad "Mala" si igual o inferior al percentil 40, pero superior al 20.
- Rentabilidad "Mediocre" si igual o inferior al percentil 60, pero superior al 40.
- Rentabilidad "Buena" si igual o inferior al percentil 80, pero superior al 60.
- Rentabilidad "Muy Buena" si rentabilidad superior al percentil 80.

Debido a la división que se ha realizado la distribución es del 20% para cada grupo de rentabilidad.

Errores

Siguiendo el procedimiento de las categorías anteriores se comprueba la distribución de errores en cada categoría, sin encontrar nuevas diferencias significativas entre las clases de rentabilidad.

Rentabilidad

Debido a la naturaleza de la categoría no tiene sentido realizar un estudio de la rentabilidad en este apartado.

Factores Fundamentales

La principal diferencia entre las acciones con muy buena rentabilidad y las acciones con muy mala rentabilidad es la ausencia de correlaciones fuertes en estas segundas, hecho que se aprecia en la Figura 20. Esto podría significar que los precios de estas acciones han disminuido por circunstancias ajenas a los fundamentales, como pueden ser factores psicológicos de mercado o características cualitativas. Dado que los últimos datos son temporalmente cercanos a la pandemia por Covid-19, es posible que sus precios estén deprimidos por las circunstancias excepcionales.

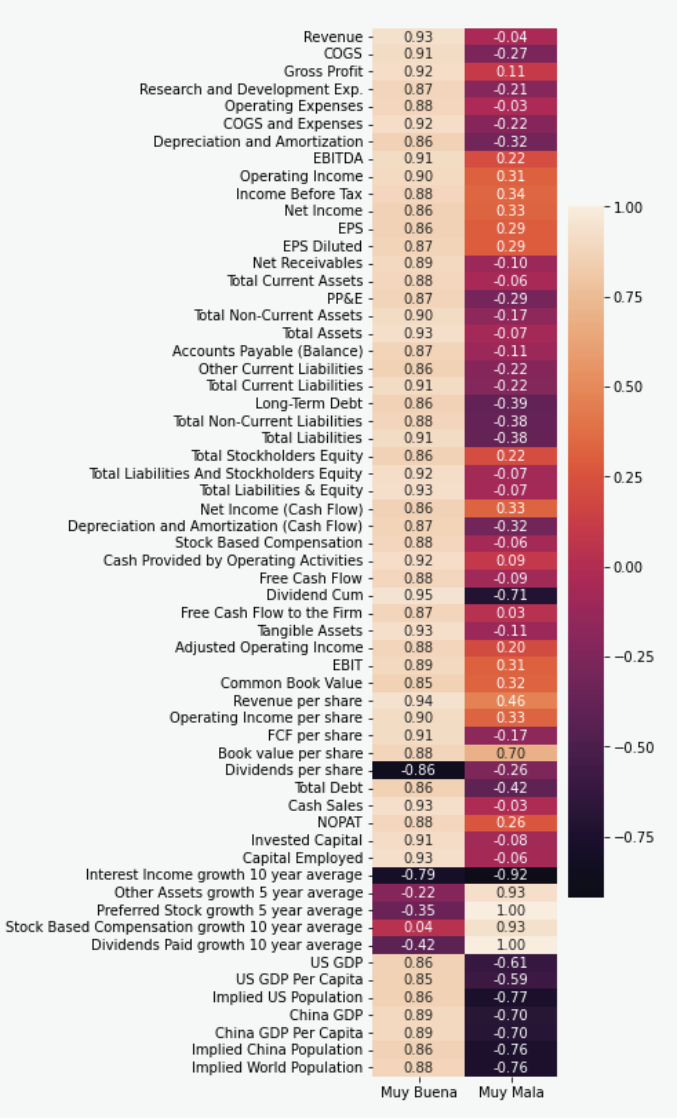

Figura 20 Correlaciones de los factures fundamentales significativos con el precio según rentabilidad

Para el grupo con buenas rentabilidades las altas correlaciones vienen dadas con los elementos ya vistos durante el análisis por sectores: ingresos, activos, dividendos y elementos macroeconómicos principalmente. Por el contrario, las correlaciones que presenta el grupo de mala rentabilidad son más interesantes. Aquellas variables con correlaciones fuertes son:

- **Variables macroeconómicas** (con correlaciones un poco más débiles). Esto podría estar ocasionado por ser acciones a los que el entorno macroeconómico no ha favorecido, sin embargo, al comparar las distribuciones por industrias entre ambos grupos de rentabilidad no se observan diferencias tan significativas en composición. Podría darse el caso, que justamente estas acciones hubieran sido afectadas negativamente dentro de sus respectivas industrias, pero no parece probable. Parece entonces que esta relación es casualidad, y que los factores macroeconómicos no perjudicaron significativamente a estas acciones.
- **Dividendos**. El pago de dividendos parece tener una relación inversa con estas empresas, cuanto más dividendos se pagan más sube el precio, lo cual no tiene sentido a nivel económico. La explicación más probable entonces es que esta relación es accidental, o aparece por un motivo desconocido.
- **Crecimiento de la compensación en acciones y crecimiento del número de acciones.** Se agrupan ambos factores, ya que un crecimiento en la compensación en acciones irremediablemente lleva a un crecimiento en el número de acciones, pues estas son de nueva creación. Al aumentar el número de acciones disminuye el valor por acción, por lo que es lógico que disminuya el precio por acción. En este casi sí es probable que la relación sea causal.
- **Crecimientos de las partidas "*Other Assets*" y "*Other liabilities*".** No queda clara la relación de estas con el decremento del precio. Podría darse que las empresas que se encuentran en peor posición financiera utilicen estas partidas para realizar maniobras contables y aparentar una mejor salud, sin embargo, es solo una teoría. Es necesaria una mayor investigación y mayores conocimientos de contabilidad para desentrañar el origen de la correlación (suponiendo que no sea casualidad).

Principales conclusiones:

- Las correlaciones del grupo de empresas con muy buen rendimiento se asemejan a las ya observadas en el análisis por sectores.
- Las correlaciones fuertes en el grupo con rendimiento muy malo son escasas y, de las pocas existentes varias parecen ser casualidad o no tener explicación aparente.
- La principal relación obtenida del estudio del grupo de muy mala rentabilidad es: a mayor crecimiento de la compensación basada en acciones, menor rentabilidad en el largo plazo.

4.6.8 Análisis por porcentaje de las acciones que poseen los *insiders*[25]

Composición

Para el estudio, se discretiza la variable "Insider Percentage" en tres categorías dependiendo de la cantidad de acciones que posea la directiva. Esta nueva categoría se guarda en la columna "Insider Ownership" y se calcula de la siguiente manera:

- Insider ownership "Bajo" si $\text{Insider Percentage} \in [0, 0.33]$
- Insider ownership "Medio" si $\text{Insider Percentage} \in (0.33, 0.66]$
- Insider ownership "Alto" si $\text{Insider Percentage} \in (0.66, 1]$

La distribución de las acciones en estas categorías se aprecia en la Figura 21. Se puede ver que el hecho de que un alto o medio porcentaje de las acciones las posean los *insiders* es algo poco usual. Comúnmente se piensa que esto es algo positivo, pues alinea los intereses de los directivos con el de los accionistas, teóricamente dando lugar a mejores rendimientos en bolsa.

Cabe destacar, que usualmente también se considera significativo la variación de este porcentaje, se cree que el hecho de que los *insiders* compren es buena señal, mientras que el hecho de que vendan puede ser algo malo. Debido a que solo poseemos los datos para el último periodo, no se puede estudiar este fenómeno.

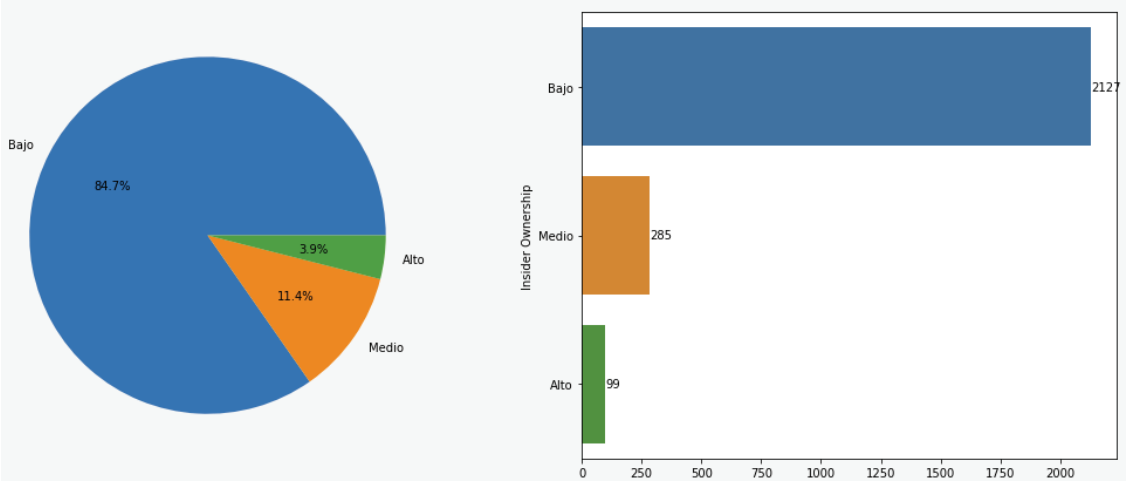

Figura 21: Distribución de las acciones según porcentaje de la empresa que poseen los "Insiders".

Errores

Respecto a la cantidad de datos faltantes, aquellas empresas con mayor *insider ownership* tienen un mayor número de datos faltantes, sin embargo, esto parece ser solo coincidencia. Las empresas con mayor *insider ownership* son aquellas más pequeñas, existiendo una correlación negativa entre ambas variables como se puede ver en la Figura 22. Por lo que la mayor tasa de errores

[25] Por "Insiders" se hace referencia a personas relacionadas con la empresa como pueden ser los directivos u otras personas importantes relacionadas con la misma.

viene dada por la menor capitalización de mercado como ya se mostraba en la Figura 17.

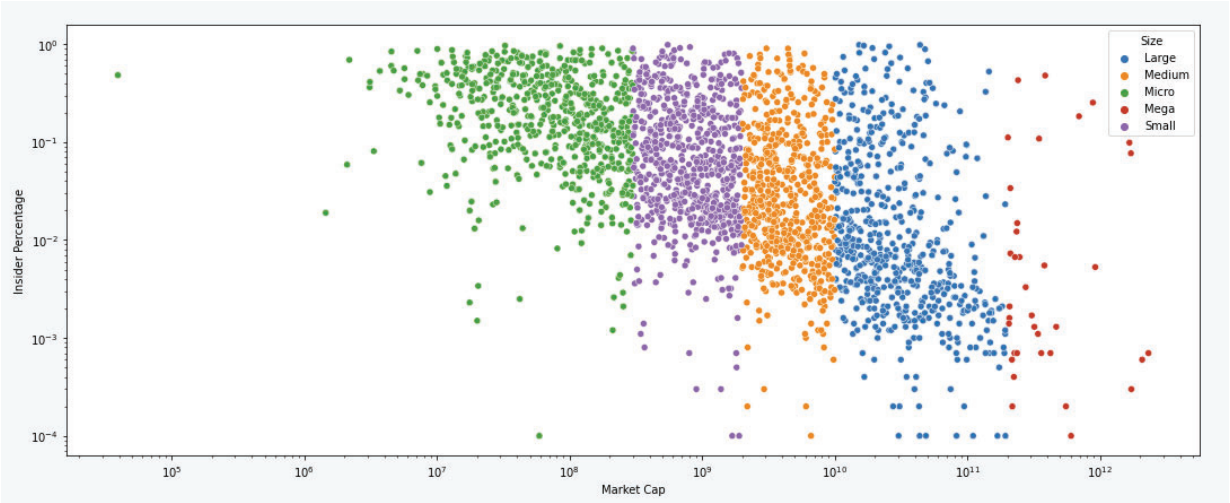

Figura 22: Gráfico de dispersión entre las variables "Insider Percentage" y "Market Cap". Ejes logarítmicos. Coloreados según categoría de tamaño.

Rentabilidad

Debido a que solo se disponen de datos sobre la propiedad de los insiders en la actualidad, no es posible realizar un análisis de como esta afecta al desempeño en bolsa. Para poder realizar este análisis debería tenerse los datos de propiedad de hace años, al igual que se ha realizado con el tamaño.

Por otra parte, analizando según los datos actuales (asumiendo que la propiedad no haya cambiado), no se obtienen relaciones concluyentes.

Principales conclusiones:

- Parece que existe una correlación inversa entre el porcentaje de la empresa que poseen los insiders y el tamaño de esta.

4.6.9 Análisis por porcentaje de las acciones que poseen los inversores institucionales[26]

Composición

Se discretiza la variable "Institution Percentage" de la misma manera que se discretizó la variable "Insider Percentage". Se obtiene así la composición que se observa en la Figura 23 cuya distribución sorprende al ser la norma que la propiedad institucional sea alta.

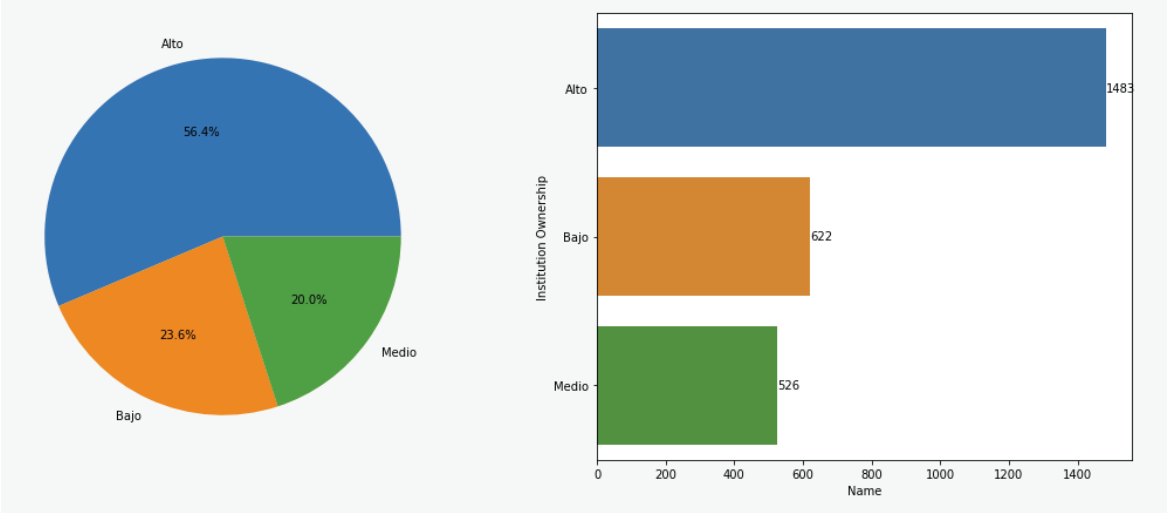

Figura 23: Distribución de las acciones según porcentaje de propiedad institucional.

Errores

En este caso se observa un caso parecido a lo ocurrido en el estudio del *insider ownership*. Parece que a menor *insider ownership* aumentan ligeramente los errores en los datos, pero de nuevo esta relación es falsa, y el causante es la disminución del tamaño de las empresas.

Existe regulación sobre las condiciones en las que los inversores institucionales pueden invertir el dinero y en qué empresas pueden hacerlo, con el objetivo de evitar manipulaciones en el mercado. Estos requisitos impiden o dificultan a los inversores institucional comprar grandes cantidades de acciones de baja capitalización. Este hecho se observa en la Figura 24, donde puede apreciarse un rápido decremento (pues el eje Y es logarítmico) de la posesión de empresas consideradas "Micro".

[26] Por "inversores institucionales" se hace referencia a inversores profesionales pertenecientes a entidades que invierten el dinero de sus clientes. Normalmente estas entidades son fondos de inversión, bancos o aseguradoras.

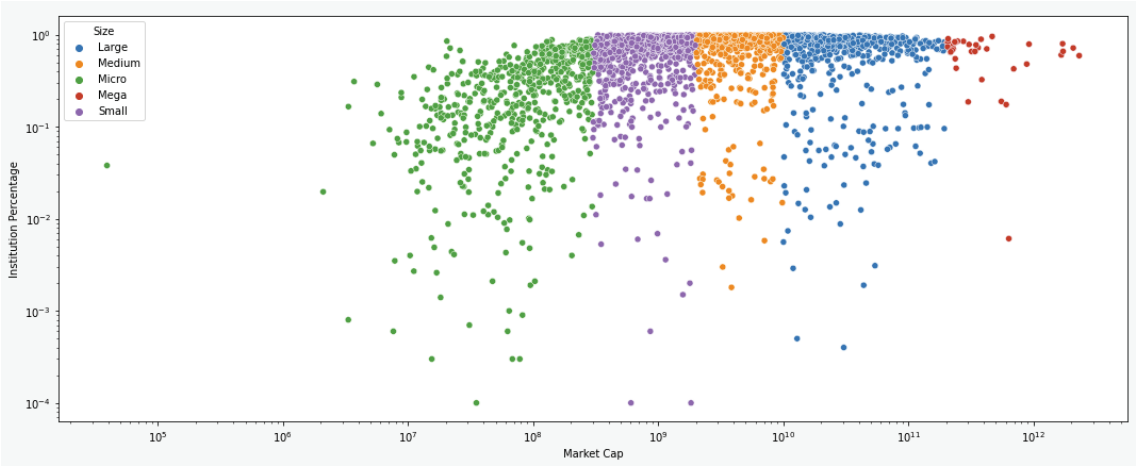

Figura 24: Gráfico de dispersión de las variables "Market Cap" e "Institution Percentage". Ejes logarítmicos. Puntos coloreados según su categoría de tamaño.

Rentabilidad

El estudio de la rentabilidad en este caso sufre el mismo problema que el apartado anterior, no se tienen datos sobre la propiedad de las instituciones hace años. Sin embargo, asumiendo la misma proporción de propiedad de instituciones todos los años, sí se observa una correlación positiva entre el porcentaje de la compañía que tienen los inversores institucionales y el desempeño en bolsa. Esta relación se piensa que viene dada por la ya explicada relación entre el tamaño de la empresa actual y el porcentaje de propiedad institucional. Esto invalida la posibilidad de sacar conclusiones definitivas por falta de datos, pues está claro que las empresas que hoy son más grandes son porque tuvieron buenos rendimientos en el pasado.

Principales conclusiones:

- El porcentaje de propiedad de los inversores institucionales cae rápidamente para las empresas “micro”.

4.7 Modelo predictivo

El objetivo del modelo predictivo es dual. Por un lado, se quiere encontrar una manear de utilizar los datos fundamentales para predecir el desempeño de las acciones a largo plazo. Por otro lado, se quiere averiguar qué factores fundamentales son los que más aportan al desempeño.

La variable objetivo que se quiere predecir es “Close”, es decir, el precio de cierre. Para ello se pretende hacer uso de una selección del resto de variables, con la intención de evitar redundancias en información o variables sin utilidad. La inversión en valor o “Value Investing” es una corriente minoritaria de la inversión que se centra en el análisis de los negocios para la predicción del precio de las acciones, sin atender a los patrones en el precio. Quizás es por este hecho que en la literatura actual es difícil encontrar trabajos sobre la predicción del precio de las acciones basados en los datos fundamentales del

negocio, pues normalmente se hace uso de indicadores técnicos. Los trabajos encontrados con el mismo enfoque son limitados, teniendo habitualmente no más de 10 variables predictoras, un conjunto de menos de 500 empresas, un enfoque cortoplacista de meses o días y un conjunto de datos financieros no mayor a 5 años.

El autor de este trabajo no tiene constancia de un modelo públicamente expuesto en la literatura de una magnitud cercana a la actual, pues se cuenta con más de 2000 empresas para fines predictivos, más de 30 años de datos financieros, un enfoque largoplacista y más de 200 variables financieras (sin contar ratios derivados como el crecimiento).

Para la predicción del precio se utilizará un modelo regresión con XGBoost, uno de los más recientes modelos basados en árboles. El uso de métodos basados en árboles para la predicción de los precios de cotización es una práctica relativamente nueva con resultados prometedores [54]. La evolución del uso de árboles en esta área a lo largo de los año se ilustra en la Figura 25.

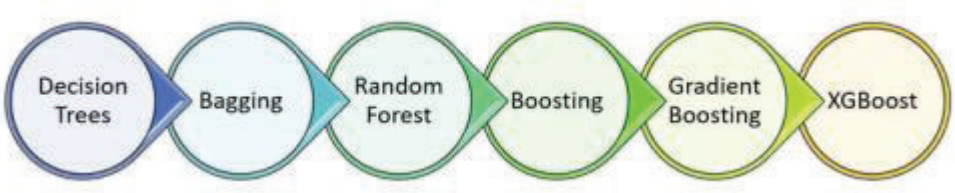

Figura 25: Evolución del uso de modelos de árboles [54]

El método de Gradient Boosting es parecido a AdaBoost pues agrega predictores secuencialmente a un modelo conjunto ("ensemble model"), cada uno de ellos corrigiendo su desempeño pasado. En contraste con AdaBoost, el Gradient Boosting ajusta un nuevo predictor a los errores residuales (hechos por el predictor anterior) usando el descenso de gradiente para encontrar fallas en las predicciones del aprendiz anterior. En general, el modelo final es capaz de emplear el modelo base para disminuir los errores con el tiempo.

XGBoost es un método conjunto ("ensemble method") basado en árboles (como Gradient Boosting) y el método aplica el principio de impulso para los aprendices débiles. Además, XGBoost cuenta con mejoras en la velocidad y el rendimiento.

4.7.1. Selección de datos

Se debe seleccionar qué variables serán aquellas que se utilicen para el entrenamiento del modelo, pero antes, se debe eliminar todas aquellas que puedan filtrar información sobre la variable objetivo (el precio). Una vez eliminadas estas variables, se introducen algunas otras al modelo, específicamente pensadas para valoración de empresas.

Lista de variables principales introducidas[27]:

[27] También se incluyen algunas variables derivadas de estas, como la media de los últimos 3,5 o 10 años.

- PER[28]
- Intrinsic Value
- Intrinsic Value FCF
- Price Moving Average
- Price to Book
- IV Price to Book
- Price to Free Cash Flow
- IV Price to Free Cash Flow
- Grahams Number
- Price to Graham Number
- IV Price to Graham Number
- PEG
- IV PEG
- Last Year Price

Una vez se han introducido las nuevas columnas enfocadas a la valoración, se utilizará la librería "featurewiz" para la eliminación automática de las variables superfluas o con alta correlación bajo el enfoque "Mínima Redundancia Máxima Relevancia" (MRMR). La librería utiliza el método "SULOV" y "XGBoost Recursivo" para reducir las características en su conjunto de datos a las mejores características para el modelo. La reducción de variables es necesaria puesto que el dataset en este punto cuenta con más de 1000 variables.

SULOV significa "Searching for Uncorrelated List of Variables" (Búsqueda de lista de variables no correlacionadas). El funcionamiento del algoritmo es el siguiente:

1. Encontrar todos los pares de variables altamente correlacionadas que excedan un umbral de correlación.
2. Encontrar su puntaje MIS (puntaje de información mutua) para la variable objetivo. MIS es un método de puntuación no paramétrico. Por lo tanto, es adecuado para todo tipo de variables y objetivos.
3. Coger cada par de variables correlacionadas y eliminar la que tenga el puntaje MIS más bajo.
4. Lo que queda son los que tienen los puntajes de información más altos y la menor correlación entre sí.

Una vez que SULOV ha seleccionado variables que tienen puntajes altos de información mutua con la menor correlación entre ellas, se usa XGBoost para encontrar repetidamente las mejores características entre las variables restantes. El método "XGBoost Recursivo" comienza dividiendo los datos en conjuntos de entrenamiento y conjuntos de validación. Posteriormente encuentra las variables más relevantes y repite este proceso en el siguiente conjunto. Tras repetir esto 5 veces se obtienen 5 conjuntos con las variables más relevantes, por lo que se combinan los conjuntos y se eliminan duplicados.

Tras la ejecución de esta parte, el número de variables se ha reducido hasta tener alrededor de 100 variables; todas ellas sin alta correlación ni información mutua. Sin embargo, este número es demasiado elevado y utilizando los datos obtenidos de XGBoost sobre importancia de variables, se puede reducir considerablemente ese número. La reducción de variables trae consigo un

[28] En todas las métricas donde se requiera el precio se utiliza el precio del año anterior.

empeoramiento del modelo, a cambio de mayor simplicidad y capacidad de generalización para el modelo. Por este motivo, se verán aumentar las métricas de error de los modelos.

En la Figura 26 se observan las variables más importantes para el modelo predictivo, con el objetivo de ver su importancia relativa. Cabe destacar que no es estrictamente necesario el uso de las variables escogidas. Las variables altamente correlacionadas descartadas durante la ejecución de SULOV, podrían sustituir a sus representantes en el dataset. Por ejemplo, la principal variable "*IV Grahams Number*" se encuentra altamente correlacionada con "*Book value per share*". Ambas variables aportan información sobre los activos de la empresa, aunque de una manera ligeramente diferente. Cambiando una variable por la otra el modelo sigue recibiendo el mismo tipo de información, sin embargo, una variable será más efectiva para el modelo que la otra, y quizás, una variable sea más fácil de entender para el usuario humano que la otra. En el caso presente, no se sustituyen variables, pero se tienen en cuenta las alternativas para una mejor explicabilidad del modelo.

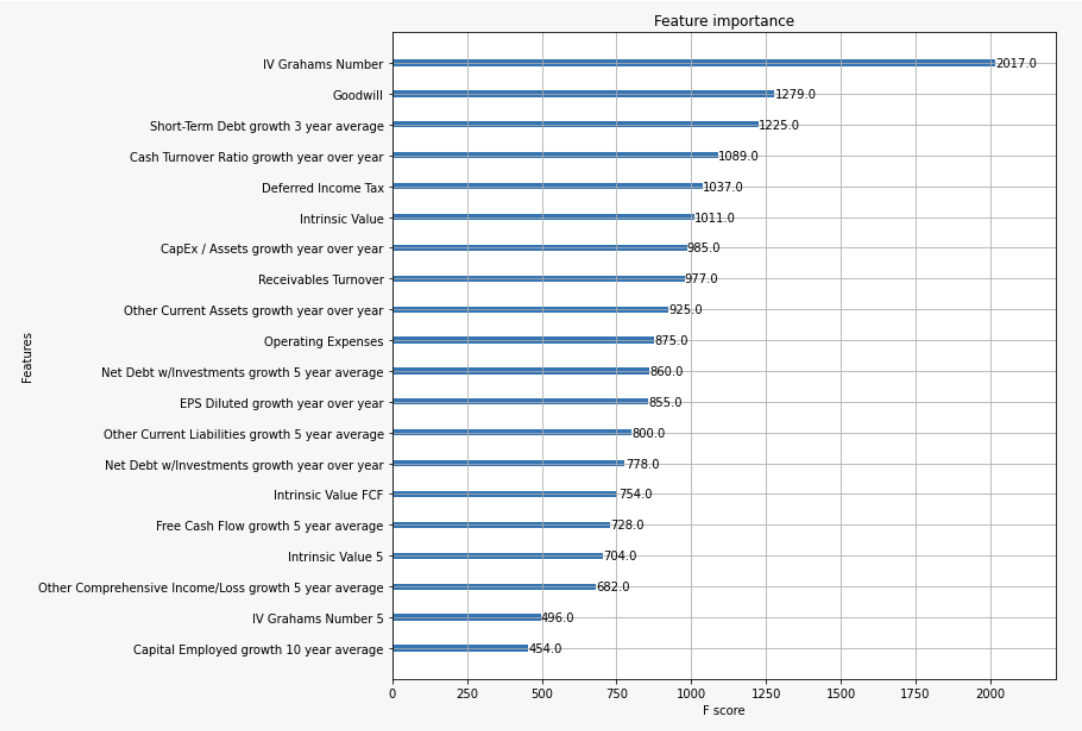

Figura 26 Importancia de las principales variables del modelo

La importancia de estas variables puede cambiar según el orden de permutación, por lo que también es relevante el estudio de su rango de variabilidad, tal y como se muestra en la Figura 27; donde se observa que, a pesar de la variabilidad, el orden de importancia queda claro.

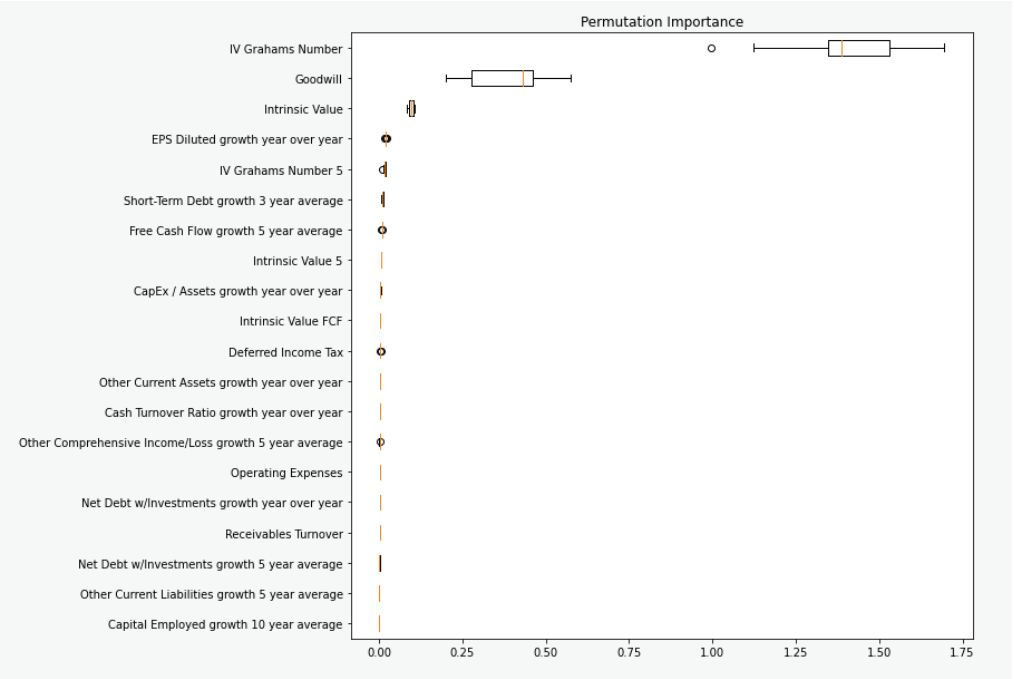

Figura 27 Importancia de las variables según la permutación

Teniendo en cuenta ambos factores y tras varias iteraciones con diferentes números de variables, se elige el conjunto de las 20 variables mostradas en la Figura 26.

Algunas relaciones que pueden ser útiles para el entendimiento de las variables son:

- "*IV Grahams Number*" relacionado con el valor en libros (activos) de la empresa
- "*Intrinsic Value*" relacionado con el crecimiento (o decrecimeinto) de las acciones
- "*Short-Term Debt growth 3 year average*" relacionado con multiples ratios de salud financiera
- "*Receivables Turnover*" relacionado con el crecimiento en gastos de capital por acción
- "*Capital Employed growth 10 year average*" relacionado con el crecimiento del ROCE

4.7.2. Elaboración de los modelos

Haciendo uso de XGBoost se quiere crear un regresor que permita predecir el precio de las acciones a futuro. Para este objetivo se han tomado diferentes perspectivas y se han realizado 4 tipos de acercamientos representados en la Tabla 12.

El primer método consiste en la creación de un dataset donde se reflejan los datos fundamentales agregados de cada empresa, ya sea utilizando la media o la mediana. De esta manera se entrenará al modelo con los datos fundamentales medios y los precios medios, eliminando en gran medida los vaivenes psicológicos del mercado. A pesar de estar entrenado con los datos agregados,

se utilizará para hacer estimaciones en base a datos de un solo año. Cabe destacar que tanto este primer método, como el segundo han sido entrenados con el precio medio del propio año[29].

Este enfoque agrega los datos haciendo la media de todos los años, sin embargo, si se intenta predecir información del 2005, en esa media usada para los entrenamientos hay datos posteriores a esa fecha. En un principio no debería haber problema debido a dos principales motivos:

I. Los datos pertenecen a otras empresas.
II. El enfoque del modelo es atemporal. El modelo no genera una relación temporal, tan solo observa las variables y genera una predicción, como si de una foto se tratara. Lo único que une al modelo con sus datos pasados son las variables de crecimiento.

Aunque en principio no debería suponer un problema, pues a ojos del modelo una entrada del 2005 es igual a una del 2020, se ha optado por el desarrollo de otro enfoque sin este problema.

El segundo enfoque pretende solucionar la incertidumbre sobre si de verdad existe filtrado de información. Al igual que el primer enfoque el dataset se crea con los datos agregados de cada empresa utilizando la media, sin embargo, solo se toman datos hasta el año que se va a predecir. Es decir, si se va a realizar una predicción para el año 2005, se crea el dataset con las medias de los años anteriores a 2005. Posteriormente el dataset se divide en el conjunto de entrenamiento y de pruebas, se entrena y se realiza la predicción. Esto elimina por completo la posibilidad de filtrar información sobre el futuro.

El tercer enfoque es análogo al primero, pero con una diferencia fundamental, esta vez el precio objetivo es el del año siguiente. De esta manera el modelo entrena para predecir los precios a futuro (al año siguiente más concretamente). Sin embargo, al ser análogo al primer método, tiene los posibles mismos problemas de filtración de información.

Por este motivo el cuarto enfoque es análogo al segundo, pero de nuevo, intentando predecir los precios futuros al igual que el tercero.

Tabla 12 Modelos de regresión para el precio

	Precio medio de este año	Precio medio futuro
Con potencial filtrado de información	Modelo 1	Modelo 3
Sin potencial filtrado de información	Modelo 2	Modelo 4

En la preparación de los modelos se ha utilizado búsqueda aleatoria en hiperparámetros para el ajuste de estos. Utilizando una “rejilla” o “red” que contenga diferentes hiperparámetros en las filas y un rango de posibles valores

[29] A primera vista puede que no parezca útil un modelo que predice el precio medio del año que ya ha pasado. Esto sí tiene utilidades que se discutirán más adelante.

en las columnas, se prueban diferentes combinaciones de estos hasta encontrar una distribución óptima.

¿Por qué hacer un modelo que predice el precio medio del año ya pasado?

Existen dos principales motivos para esto. El primer motivo es el carácter didáctico, ya que se puede llegar a comprender con este modelo los factores fundamentales que han contribuido al precio de este año. El segundo motivo viene relacionado con un una de las estrategias de inversión que se van a poner a prueba.

Gracias al R^2 se puede saber qué porcentaje de variabilidad es explicado por el modelo. El modelo propuesto solo contiene variables fundamentales y de negocio, por lo que es muy probable que las variables faltantes sean psicológicas o de una índole más abstracta. El hecho de que el modelo requiera de una entrada de datos fundamentales durante un año hace que sólo pueda utilizarse o con suposiciones o una vez acabado el año para la empresa. Sin embargo, aunque el año acabe al día siguiente la empresa sigue cotizando en bolsa y ya se disponen los últimos fundamentales para calcular el precio medio del año que acaba de pasar. El valor de las empresas no suele variar demasiado en periodos de 6 meses, por lo que la estimación del precio medio del año ya acabado puede ser una buena guía para la compra o venta de acciones. No solo eso, si no que el modelo propuesto se ve menos afectado por las tendencias psicológicas que adopte el mercado, por lo que puede servir como un indicador "más racional" del valor de las acciones.

4.7.3. Evaluación de los modelos

Para la evaluación de los modelos se tendrán en cuenta dos principales tipos de métricas: las métricas tradicionales para modelos de regresión (R^2, MAE, RMSE) y una métrica relativa a la inversión: el crecimiento anual compuesto de un portfolio que hubiera utilizado ese modelo. Esta segunda métrica no es inmediata y requiere de la construcción de una pequeña herramienta que permita similar los rendimientos de la estrategia de inversión en el pasado.

En el caso presente se ha desarrollado una función que utilizando datos históricos simula el rendimiento del portfolio año a año y devuelve su crecimiento anual compuesto desde el inicio. La estrategia de inversión tiene dos principales parámetros: el número de acciones que intenta comprarse a cada periodo y el número de años hasta el siguiente rebalanceo. El primer parámetro indica cuántas acciones se quieren tener en el portfolio, mientras que el segundo es el número de año que se esperarán tras comprar una acción para vender, y repetir el proceso de compra. Este proceso consiste en, para un determinado año, hacer las predicciones con el modelo correspondiente y anotarlas. Una vez anotadas se calcula la diferencia entre las predicciones (los precios esperados) y los precios reales a final de año. Se ordenan las acciones según su diferencia entre el precio real y el esperado, cogiendo las N (según el número de acciones a comprar en cada periodo) primeras de la lista ordenada, para coger las acciones más infravaloradas. Entonces estas acciones se añaden

al portfolio simulando una compra[30] y pasado el periodo de espera, se consulta su precio y se simula su venta, dejando como diferencia la plusvalía (o minusvalía) de la operación. Para todas las evaluaciones se ha utilizado validación cruzada para garantizar un resultado fiable, aunque nada garantiza que el modelo funcionará igual una vez desplegado. Para las métricas de evaluación (R^2 y crecimiento anual compuesto del portfolio asociado) se ha cogido la media de múltiples iteraciones para garantizar consistencia en los resultados.

Resultados

Comenzando con los modelos 1 y 3, se observa en la Tabla 13 que en ambos modelos consiguen un R^2 similar, mientras que en las otras métricas el modelo 1 es ligeramente superior. Era de esperar pues el modelo 3 es el que realiza predicciones a futuro, por lo que se espera que su desempeño sea peor.

Tabla 13 Métricas de evaluación de los modelos 1 y 3 [20 variables]

	Model 1	Model 3
R_Square	0.693536	0.695795
Negative MAE	-12.541368	-17.765981
Negative RMSE	-60.992109	-62.369036

Dado que los modelos 2 y 4 utilizan datasets diferentes para la predicción de cada año (simulando que cada año se incorpora la nueva información), se evaluarán por años. Si se está evaluando el año 2003 el dataset de entrenamiento estará formado por la media de los datos fundamentales hasta el 2002.

La evaluación de estos modelos se puede observar en la Tabla 14, donde se observa que las predicciones son peores los primeros años, aunque pasan a niveles aceptables durante la segunda mitad. Como ya se ha comentado el modelo no contempla factores psicológicos entre sus variables, por lo que es casi neutral[31] respecto a estos. Esta mejora en los años tardíos puede deberse a que, para entonces los datasets son mucho más grandes.

Cabe destacar lo ya observado en los modelos 1 y 3, y es que el modelo 2 es superior en todas las métricas al modelo 4 (sobre todo en periodos de crisis como 2008); hecho que se explica por la dificultad de predecir los precios futuros.

[30] Más concretamente, se simula que se disponen de $1000 para la compra de acciones y se reparten de manera equitativa entre las acciones a comprar. De esta manera se evita que el precio de cotización sobre pondere unas acciones.

[31] "Casi neutral" debido a que los precios dados para su entrenamiento pueden estar afectados por los ciclos eufórico-depresivos del mercado, aunque este efecto se haya reducido mediante el uso de medias y medianas para agregarlos.

Tabla 14 Evaluación del modelo 2 y del modelo 4

	Model 2: R_Square	Model 4: R_Square	Model 2: MAE	Model 4: MAE	Model 2: RMSE	Model 4: RMSE
2001	-1.4206	-16.7311	-25.2374	-29.3748	-165.4004	-249.3465
2002	-4.8625	-5.9616	-21.1391	-20.0617	-190.0951	-173.8216
2003	-0.5893	-1.7916	-18.4879	-20.0057	-133.7568	-148.0934
2004	-1.5062	-1.9695	-19.9263	-20.8868	-133.1962	-148.934
2005	-47.6755	-85.1917	-54.2359	-44.402	-499.2643	-500.4276
2006	0.2121	0.251	-29.2732	-30.3767	-213.559	-284.9794
2007	0.3297	0.0745	-23.9546	-28.6091	-226.4819	-279.0354
2008	-0.6108	-49.3862	-34.4015	-49.0399	-265.0578	-484.6644
2009	0.2478	-0.0489	-30.1871	-26.2402	-228.1994	-203.2727
2010	0.2284	0.1387	-28.62	-29.6476	-214.1392	-230.8871
2011	0.3272	0.4986	-24.3674	-20.6745	-192.7344	-159.2181
2012	0.1503	0.293	-24.9857	-19.4828	-179.5882	-143.2802
2013	0.6046	0.5686	-20.4043	-18.4231	-114.9272	-112.8476
2014	0.6036	0.6498	-17.9878	-16.2389	-103.1092	-90.7134
2015	0.312	-0.3616	-18.2654	-17.6734	-116.3096	-124.731
2016	0.7218	0.5897	-15.1814	-16.1605	-79.8479	-88.021
2017	0.7149	0.724	-13.4788	-13.7608	-69.1729	-60.8638
2018	0.6581	0.6298	-12.9817	-14.0704	-68.7667	-69.883
2019	0.6709	0.591	-13.6951	-14.5944	-66.3717	-68.456
2020	0.6723	0.3564	-12.6092	-14.6578	-65.6374	-77.0975

El modelo de 20 variables obtiene métricas sustancialmente peores en la evaluación que el modelo con 100 variables. Como referencia, el modelo 1 tenía un R^2 cercano a 0.9 y el modelo 3 uno cercano a 0.86. Como ya se ha expuesto, se espera que los modelos de 20 variables tengan menor precisión. La intención es que a pesar de que se vea un decremento significativo en métricas como el R^2 , la métrica más importante (la rentabilidad del portfolio al hacer backtesting) se mantenga similar. Esto es posible dado que no se necesita saber el precio exacto de las acciones para invertir, solo tener una idea de si está barata o cara, algo que el modelo podría ser capaz de conseguir con tan solo 20 variables.

Rentabilidad histórica de los modelos

El segundo método de evaluación de los modelos es el backtesting[32]. El procedimiento es el ya explicado anteriormente, haciendo predicciones para cada año y comprando acciones en base a ello.

Para esta evaluación se requieren de dos parámetros, el primero es el número de acciones que se intentan comprar en cada periodo. Si este número es de 5 se intentarán comprar las 5 acciones más baratas, si solo hubiera 3 baratas se compran solo 3. El segundo parámetro es el número de años que se mantienen las acciones compradas y el número de años que se tarda en hacer el rebalanceo. En el caso base se compran acciones, se mantienen un año y se venden; realizando otra vez el mismo proceso. Si el número de años a mantener

[32] Por "*backtesting*" se hace referencia al uso del modelo en años anteriores para crear carteras de inversión, simulando su uso en un entorno real.

las acciones es de 3, se mantendrán las acciones durante 3 años y se realizarán compras cada 3 años.

Además, con el objetivo de detectar sesgos en el modelo, se realizará una predicción para "posiciones largas", es decir, intentando detectar empresas baratas para comprarlas y ganar dinero; y una predicción para "posiciones cortas", es decir, intentando encontrar acciones caras.

Posiciones largas

En la Tabla 15 se observa el crecimiento anual compuesto de los diferentes modelos, según el número de acciones que se intentan comprar y el periodo de mantenimiento de la compra.

Tabla 15 Rentabilidad histórica de los modelos según número de acciones. Portfolio Long.

Numbers of stocks	1	6	11	16	21	26	Mean	Std	Sharpe
Long Model									
Model 1	0.29	0.63	0.62	0.41	0.36	0.45	0.46	0.13	3.54
Model 1: Hold 3 Years	0.24	0.34	0.21	0.22	0.26	0.28	0.26	0.04	6.50
Model 1: Hold 5 Years	0.34	0.49	0.23	0.32	0.24	0.24	0.31	0.09	3.44
Model 2	0.33	0.48	0.56	0.45	0.47	0.41	0.45	0.07	6.43
Model 2: Hold 3 Years	0.28	0.24	0.23	0.36	0.24	0.21	0.26	0.05	5.20
Model 2: Hold 5 Years	0.31	0.4	0.26	0.18	0.2	0.27	0.27	0.07	3.86
Model 3	0.72	0.69	0.54	0.45	0.46	0.49	0.56	0.11	5.09
Model 3: Hold 3 Years	0.29	0.41	0.19	0.17	0.25	0.18	0.25	0.08	3.12
Model 3: Hold 5 Years	0.13	0.25	0.28	0.32	0.29	0.19	0.24	0.06	4.00
Model 4	0.32	0.65	0.56	0.43	0.44	0.42	0.47	0.11	4.27
Model 4: Hold 3 Years	0.14	0.16	0.26	0.29	0.3	0.25	0.23	0.06	3.83
Model 4: Hold 5 Years	0.06	0.17	0.19	0.18	0.24	0.2	0.17	0.06	2.83

Lo primero que llama la atención es la aparición del ratio Sharpe. Este ratio fue desarrollado por el premio Nobel William F. Sharpe y se utiliza ayudar a comprender el rendimiento de una inversión en comparación con su riesgo , por lo que cuanto más elevado, mejor. Juzgando por el ratio Sharpe el modelo 2 es superior, pues ofrece rentabilidades elevadas con la menor volatilidad.

Otro hallazgo importante es que la rentabilidad de los modelos, al mantener las posiciones durante más tiempo, se ve disminuida. Esto se explica por las condiciones cambiantes del mercado y la empresa, mientras que el modelo no ve actualizado durante ese periodo.

Otro factor por destacar es el hecho de que conforme se aumenta el número de acciones que se compran, disminuye ligeramente la rentabilidad. Puesto que se cogen las acciones ordenadas según su infravaloración, cada acción adicional que se incorpora está "menos barata", disminuyendo el retorno medio del portfolio. La teoría de portfolios moderna argumenta que esta rentabilidad perdida, se pierde a cambio de obtener un portfolio más seguro, es decir, menos volátil.

Esto se comprueba en la Tabla 16, donde se han separado los modelos en dos grupos. El grupo 1 son las rentabilidades cogiendo 1, 6 y 11 acciones; mientras que el grupo 2 son las rentabilidades cogiendo 16, 21 y 26 acciones. Es decir, se trata de una división para poner a prueba si carteras más concentradas son más volátiles. Esta afirmación se sostiene en base a los datos, pues se aprecia como los portfolios más concentrados son más rentables a cambio de tener mayor desviación típica. Desde el punto de vista de la teoría de portfolios moderna, los portfolios con los ratios Sharpe más elevados son los diversificados, y, por tanto, los recomendables.

Tabla 16 Media, desviación típica y ratio Sharpe según grupo de acciones

Numbers of stocks	Mean 1	Std 1	Mean 2	Std 2	Sharpe 1	Sharpe 2
Long Model						
Model 1	0.51	0.19	0.41	0.05	2.68	8.20
Model 1: Hold 3 Years	0.26	0.07	0.25	0.03	3.71	8.33
Model 1: Hold 5 Years	0.35	0.13	0.27	0.05	2.69	5.40
Model 2	0.46	0.12	0.44	0.03	3.83	14.67
Model 2: Hold 3 Years	0.25	0.03	0.27	0.08	8.33	3.38
Model 2: Hold 5 Years	0.32	0.07	0.22	0.05	4.57	4.40
Model 3	0.65	0.10	0.47	0.02	6.50	23.50
Model 3: Hold 3 Years	0.30	0.11	0.20	0.04	2.73	5.00
Model 3: Hold 5 Years	0.22	0.08	0.27	0.07	2.75	3.86
Model 4	0.51	0.17	0.43	0.01	3.00	43.00
Model 4: Hold 3 Years	0.19	0.06	0.28	0.03	3.17	9.33
Model 4: Hold 5 Years	0.14	0.07	0.21	0.03	2.00	7.00

Examinando la Tabla 15 con cuidado se observa algo contraintuitivo, los modelos 3 y 4 tienen una rentabilidad similar a los modelos 1 y 2. Los modelos 3 y 4 son aquellos que intentan realizar predicciones a futuro, además de que ya se visto que su R^2 es sistemáticamente inferior al de los modelos 1 y 2. No solo eso, las rentabilidades son similares a las del portfolio realizado por los modelos de 100 variables, confirmando la teoría de que no es necesario la exactitud que estos ofrecen. Como ya se ha comentado antes, puesto que se cogen las acciones más baratas, el modelo tan solo tiene que ser bueno indicando cuales son las más baratas; sin necesidad de estimar el precio correctamente.

Comparación con el SP500 y el universo de acciones de referencia

Ya se ha comentado que lo importante de las inversiones no es el retorno absoluto, sino relativo. Para saber si los portfolio de acciones generados por estos modelos hubiera sido un buen negocio debe compararse frente al retorno de mercado. Más importantemente, deben compararse frente al retorno del universo de empresas estudiadas. Si nuestro dataset de manera natural genera rendimientos del 50% anual, conseguir esa rentabilidad con el modelo implica que el modelo no aporta nada.

Para el periodo estudiado el crecimiento anual compuesto del SP500 fue del 5%, muy por debajo de cualquiera de los modelos generados, mientras que el del universo de empresas fue del 6%.

El interés compuesto se magnifica conforme pasa el tiempo, creciendo de manera exponencial. En la Figura 28 se observa el crecimiento que tendría cada portfolio durante los siguientes 10 años si mantuvieran su crecimiento anual compuesto medio. Una diferencia de pocos puntos porcentuales se magnifica en el largo plazo, dando lugar a diferencias de decenas de miles de dólares.

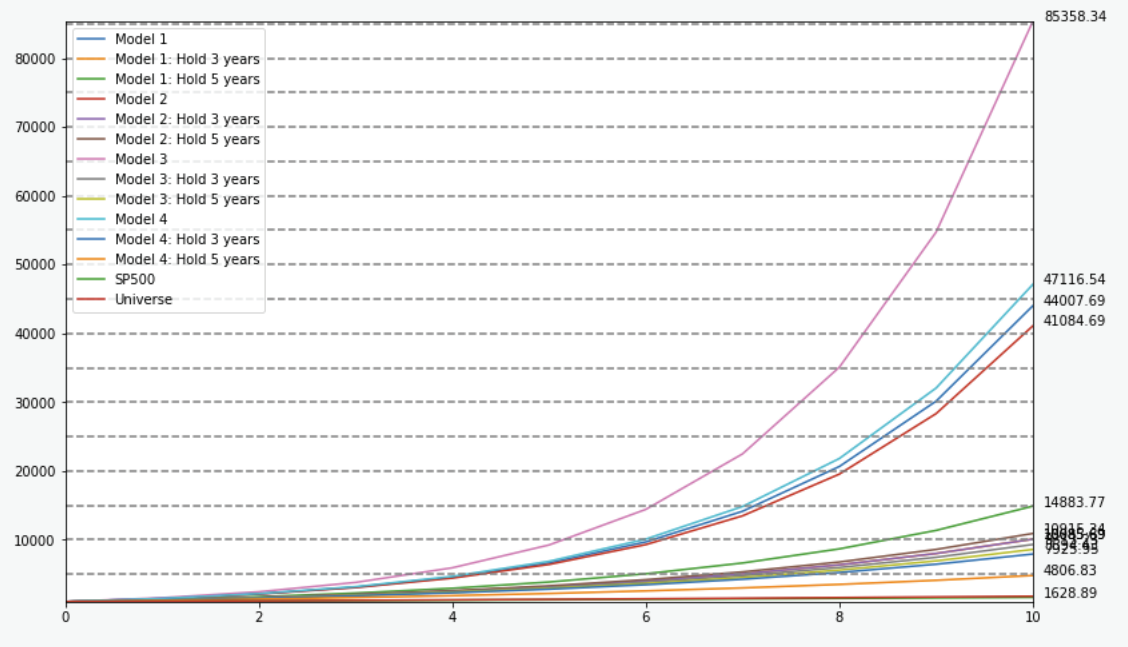

Figura 28 Comparativa del desempeño de los diferentes modelos según su CAGR medio [Long Portfolio]

Posiciones cortas

Ya se ha comentado que también se va a analizar el portfolio generado por los modelos al intentar comprar acciones caras. El objetivo es comprobar si los modelos son capaces de detectar correctamente tanto las acciones baratas como las caras. Si esto fuera cierto, podrían aplicarse estrategias long-short de portfolio, para obtener una cartera de acciones neutra respecto de los mercados alcistas y bajistas. Por ende, se busca que la rentabilidad de los portfolios generados sea lo más baja posible. En la Figura 29 se aprecia las rentabilidades de los portfolios generados, esta vez con resultados no tan satisfactorios.

Algunos modelos son capaces de generar portfolios que lo hacen peor que el mercado, sin embargo, también se generan otros que lo hacen mucho mejor. Parece que los modelos tienen un sesgo alcista, son capaces de detectar la infravaloración de empresas, pero no su sobrevaloración. Cabe destacar que el modelo completo de 100 parámetros (a pesar de no ser perfecto), obtenía mejores resultados detectando cuando las acciones estaban infravaloradas.

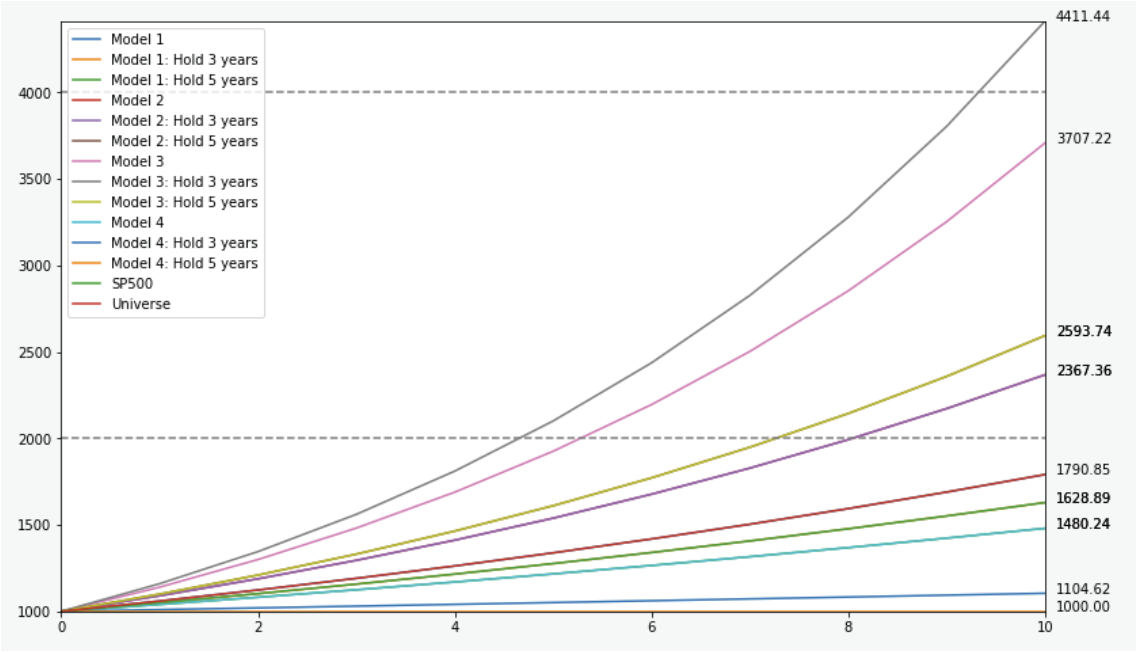

Figura 29 Comparativa del desempeño de los diferentes modelos según su CAGR medio [Short Portfolio]

Dado que todos los portfolio generados por los modelos hubieran sido capaces de batir al SP500 y a su universo de empresas durante el periodo estudiado, se considera que el desarrollo de los modelos para detección de acciones infravaloradas ha sido un éxito. Estos modelos, sin embargo, tienen dificultades para juzgar que acciones están sobrevaloradas, y no son fiables para estrategias *long-short*.

4.8. Despliegue de la aplicación

Poor último, se ha desarrollado una herramienta con el objetivo de ayudar a los inversores particulares a tomar mejores decisiones de inversión. Esta aplicación web reutiliza el código utilizado durante la recolección y el procesamiento de datos, para recolectar los datos en tiempo real y reconstruir estos en el momento.

En la aplicación se pueden encontrar resúmenes de los principales documentos contables de la empresa, expuesta de manera sencilla y visual para que sea accesible a todo el mundo. Además, se han incorporado métricas clave para la correcta predicción del precio a futuro, en base a lo hallado durante el desarrollo de este proyecto.

A continuación, se proporciona la lista comentada brevemente de las funciones que ofrece la herramienta.

Historial de precios, información básica y resumen del negocio

Se muestra en la Figura 30. Permite consultar el precio histórico, así como información básica relativa a la empresa. En la parte superior derecha se encuentra situada la barra de búsqueda para las diferentes acciones.

Alguna de la información que ofrece es:

- Precio actual
- Sector
- Industria
- País
- Capitalización de mercado
- Máximo y mínimo de las últimas 52 semanas
- Resumen del negocio
- Resumen del rendimiento pasado de las acciones
- Porcentaje de acciones en manos de *insiders*
- Porcentaje de acciones en manos de inversores institucionales
- Años desde su salida a bolsa

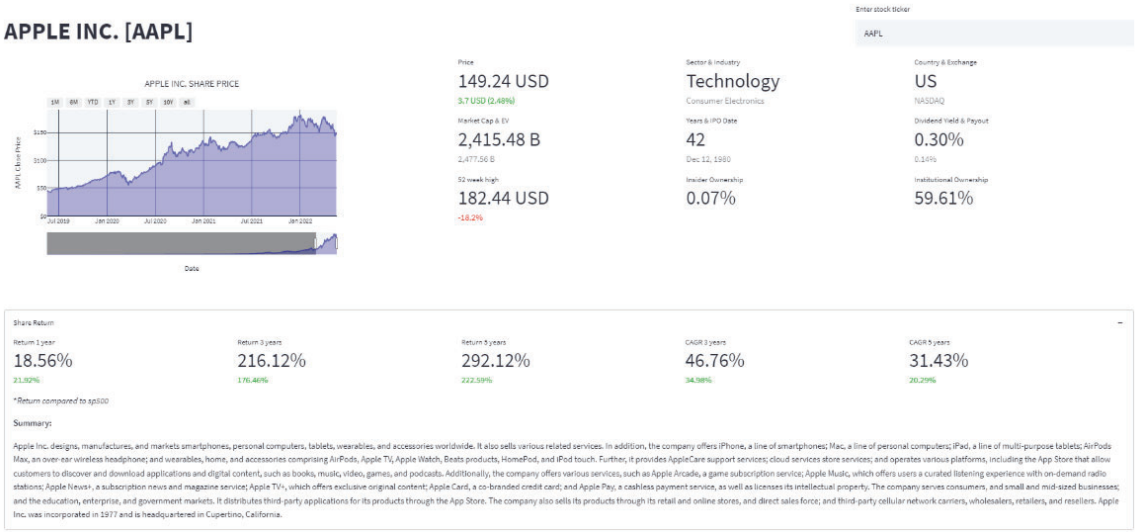

Figura 30 Aplicación web : Historial de precios, información básica y resumen del negocio

Resumen de la cuenta de pérdidas y ganancias

Se muestra en la Figura 31. En esta sección se resumen los ingresos de la empresa y su desempeño durante los últimos 5 años, representados con un diagrama de barras interactivo. Durante el estudio se ha podido comprobar que el crecimiento de los beneficios operativos juega un rol importante en el desempeño de las acciones, así como el crecimiento en el número de acciones que también se muestra.

Alguna de la información que ofrece es:

- Ingresos por acción
- Beneficios por acción
- Flujo de caja libre por acción
- Dividendos por acción
- Crecimiento del número de acciones
- Evolución de los ingresos, beneficio bruto, EBITDA, beneficio operativo, etc…

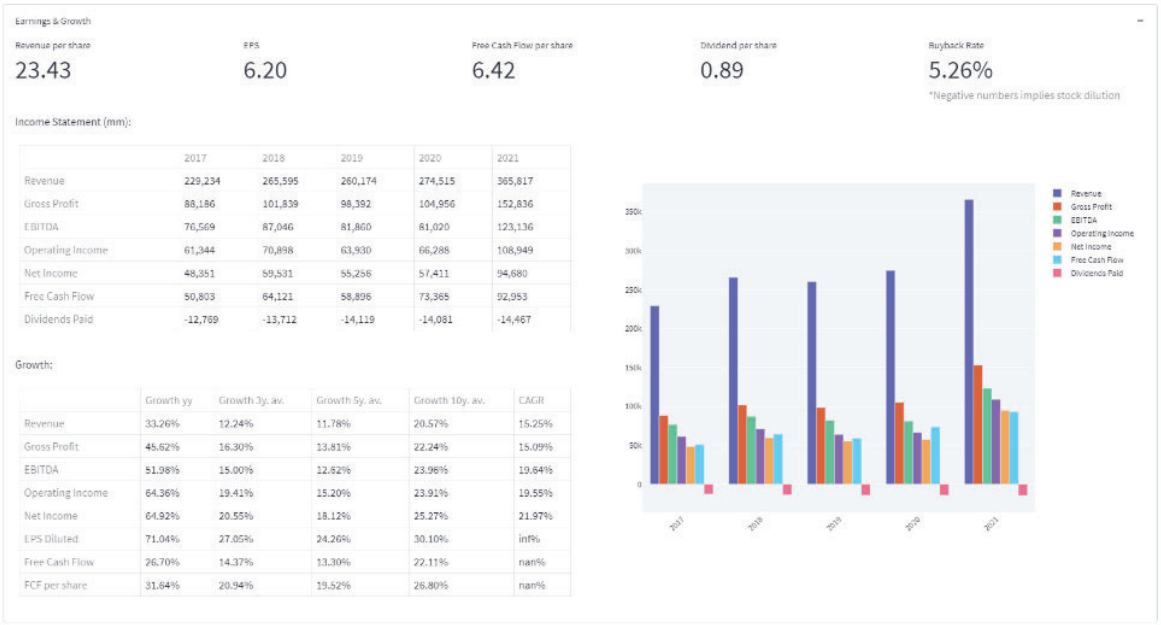

Figura 31 Aplicación web : Resumen de la cuenta de pérdidas y ganancias

Resumen de los márgenes de rentabilidad y los principales ratios financieros

Se muestra en la Figura 32. En esta sección se muestra un resumen de la evolución de los márgenes de la empresa durante los últimos 5 años. También se muestra la evolución de los principales ratios de rentabilidad de la empresa, que han demostrado ser relevantes a la hora de estimar los precios a futuro.

Alguna de la información que ofrece es:

- Crecimiento y valor del ROE durante los últimos 5 años
- Crecimiento y valor del ROIC durante los últimos 5 años
- Crecimiento y valor del ROA durante los últimos 5 años
- Márgenes de beneficio bruto, beneficio operativo, EBITDA, beneficio neto, etc…

Figura 32 Aplicación web: Resumen de los márgenes y la rentabilidad del negocio

Resumen de los ratios de salud financiera

Se muestra en la Figura 33. La salud financiera es vital en la inversión en bolsa, una empresa que quiebra ve su valor de cotización disminuir significativamente; pudiendo llegar a no valer nada.

Alguna de la información que ofrece es:

- Caja de la empresa
- Inversiones a corto plazo
- Deuda total
- Deuda Neta
- Deuda / Fondos Propios
- Deuda / EBITDA
- Ratio de liquidez
- Test ácido

Figura 33 Aplicación web: Resumen de la posición financiera de la empresa

Resumen de los ratios de valoración

Se muestra en la Figura 34. Los ratios de valoración son las variables que mayor valor aportan a la hora de predecir el precio futuro. Estos ratios dan información sobre lo barata o cara que se encuentra una empresa.

Alguna de la información que ofrece es:

- PER
- PEG
- Price to Book
- Price to FCF
- Price to Sales
- EV to EBITDA
- EV to FCF
- Modelo completo para calcular el valor intrínseco de la empresa mediante un descuento de flujos de caja

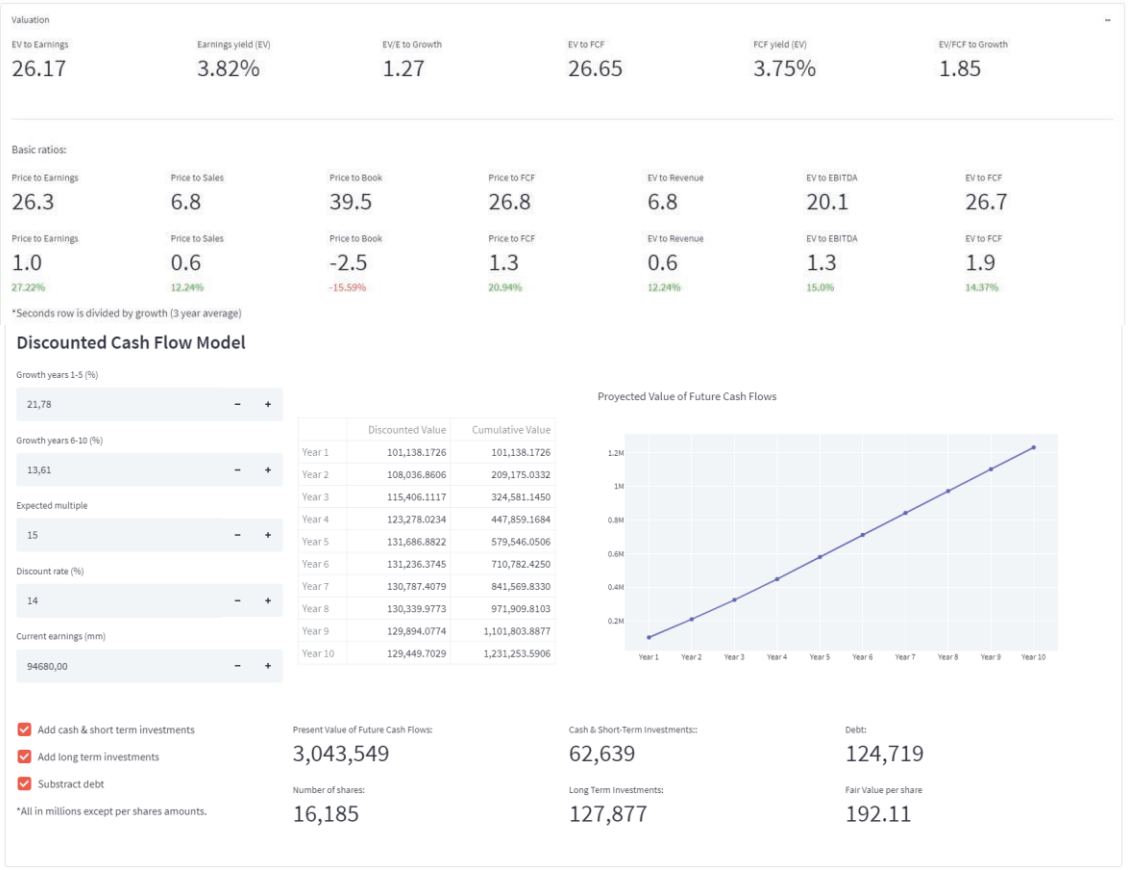

Figura 34 Aplicación web: Resumen de los principales ratios y métodos de valoración

Capítulo 5

Conclusiones

En el presente trabajo se ha buscado encontrar los principales factores fundamentales que contribuyen al desempeño de las acciones en bolsa. Para lograr este objetivo se han construidos datasets sobre datos históricos de miles de empresas, recolectando y limpiando dichos datos. Los datasets creados han sido explorados y analizados, con el objetivo de entender cuáles han sido estos factores que afectaban a los retornos de las acciones. Posteriormente, con una mejor comprensión, se han desarrollado modelos de regresión capaces de predecir exitosamente los precios de las acciones, basándose en datos fundamentales. Por último, se ha elaborado una aplicación web, utilizando una interfaz sencilla y amigable, que pone al alcance de todo el mundo las conclusiones del estudio; mostrando los datos más relevantes a la hora de analizar empresas.

5.1 Conclusiones

Se ha cumplido de manera exitosa el principal objetivo del proyecto, pues se han encontrado hechos comprobables y respaldados por los datos sobre los cuales es posible construir una metodología de inversión.

Algunos de los hechos manifestados durante el estudio son:

- Existe menos competencia de inversores institucionales en el mercado de empresas “micro”.
- Se observa una mayor cantidad de insiders con acciones en las empresas pequeñas.
- Las empresas ancianas resultan menos volátiles que las más jóvenes.
- Estados Unidos y Canadá resultaron ser mejores lugares para invertir que China, durante el periodo estudiado.
- Los sectores intensivos en capital y fuertemente dependientes de sus activos se ven beneficiados por entornos de tipos de interés decrecientes.
- Durante el periodo estudiado los sectores más rentables y seguros han sido “Utilidades”, “Bienes raíces”, “Consumo Cíclico” y “Servicios financieros”, aunque esto podría cambiar para otros periodos.

Respecto a los factores fundamentales que más influencian en los precios de las acciones, gracias al análisis realizado y a la información aportada por los modelos desarrollados, podemos enumerar una serie de factores importantes (omitiendo aquellos altamente correlacionados):

- Los beneficios por acción actuales multiplicados por el PER del año anterior, así como el valor intrínseco calculado por un descuento de flujos de caja[33]

[33] Con los datos de entrada especificados en el anexo.

- Los activos totales de la empresa y su cantidad de caja y equivalentes.
- Los Beneficios y el crecimiento de estos
- El crecimiento (o decrecimiento) en el número de acciones en circulación
- Los dividendos repartidos a los accionistas
- Algunos ratios de salud financiera (como la deuda entre los activos)
- Algunos ratios de rentabilidad (como el ROE y el ROCE) y su crecimiento
- Diversos factores macroeconómicos como los tipos de interés.

Es con estas conclusiones que se da por cumplido el objetivo principal del trabajo.

De manera particular se evalúan los objetivos específicos propuestos con el fin de asegurar el cumplimiento del objetivo principal:

Aplicar metodología CRISP-DM para el estudio de los datos fundamentales de las empresas cotizadas en bolsa.

Se ha realizado con éxito el ciclo completo de la metodología propuesta, desde la investigación para obtener conocimientos de negocio hasta el desarrollo de una herramienta que de utilidad a los datos recolectados durante el estudio. Por este motivo se considera que se ha cumplido con el objetivo propuesto.

Realizar modelos predictivos que ayuden a la construcción de una cartera de inversiones.

Los modelos predictivos se han demostrado de utilidad, dando resultados excepcionales durante el *backtesting* y demostrando su capacidad para batir al mercado de manera consistente durante los últimos 20 años.

Confeccionar una herramienta que haga uso de las conclusiones extraídas para ayudar a un inversor a tomar mejores decisiones.

Se ha elaborado de manera exitosa una aplicación web que pone a disposición de los inversores individuales la información más relevante, según el estudio realizado, para investigar empresas. Todo esto se realiza mediante una interfaz amigable y sin necesidad de tener avanzados conocimientos sobre contabilidad financiera.

Discutir los resultados obtenidos, evaluando el procedimiento seguido, así como las conclusiones obtenidas.

Los resultados obtenidos se han discutido de manera eficaz en su apartado correspondiente, así como el procedimiento seguido y las decisiones tomadas a cada paso. Las conclusiones obtenidas se han expuesto de manera simple y ordenada.

5.2 Propuestas de Mejora

El trabajo se ha realizado de manera satisfactoria, cumpliendo los objetivos de negocio y de data Mining, superando con creces las expectativas iniciales del proyecto. Aun así, existen una gran cantidad de limitaciones para esta investigación, así como una gran variedad de maneras de explorar este tema con mayor rigor.

La primera de las limitaciones es el volumen y la calidad de los datos disponibles. Más y mejores datos afectarían positivamente al desarrollo de los modelos predictivos, ayudando a sacar conclusiones menos sesgadas y más alineadas con la realidad. Otra limitación es el uso de la desviación típica estándar como factor para medir el riesgo, podrían utilizarse métodos más rigurosos para un mejor control del riesgo.

El proyecto ha sido desarrollado por tan solo una persona, por lo que también existen limitaciones en cuanto a los conocimientos, los sesgos, el tiempo y las capacidades de la persona que ha realizado el proyecto. Contando con mejores conocimientos sobre contabilidad o economía es posible que pudiera realizarse una mejor reconstrucción de los datos. Contando con más puntos de vista es posible que pudieran evitarse sesgos del autor. Contando con más tiempo podrían probarse más modelos de aprendizaje automático, así como dedicar más tiempo al procesamiento de datos.

Existen más mecanismos de validación del modelo no presentados durante el proyecto, que podrían utilizarse para, potencialmente, mejorar la generalización del modelo. Adicionalmente podrían explorarse diferentes métodos de selección de variables como entradas para los modelos, potencialmente, generando modelos más sencillos y potentes. También existen una serie de decisiones arbitrarias tomadas a lo largo del proceso, que con mayor rigor y meditación podrían mejorar los resultados del proyecto. La herramienta desarrollada también presenta limitaciones, por lo que podría estudiarse a futuro cuales son los indicadores que deberían estar presentes en la misma, eliminando aquellos que menos aportarán. Por último, podrían mezclarse en futuros proyecto el análisis de sentimientos, el análisis técnico y el análisis fundamental para una predicción más acertada.

Capítulo 6

Análisis de impacto

El presente trabajo ha tenido como principal objetivo el acercamiento de la inversión a los inversores particulares y el estudio de bases sólidas para una mejor inversión en bolsa. Se espera con este trabajo proporcionar fundamentos sencillos sobre la inversión para ayudar a los inversores con menos capital y dinero sin acceso a formación; por este motivo son cuatro los objetivos de desarrollo sostenible en los que este trabajo se enfoca.

La educación de calidad es uno de los bienes más preciados de la sociedad moderna, una buena educación global garantiza el desarrollo del ser humano como especie. Existe una gran cantidad de desinformación en el mundo de las finanzas, con varias escuelas de inversión, cientos de métodos para ganar dinero y muchas capas de complejidad a cada paso. Esto sumado a la necesidad de comprender varias disciplinas como son la psicología, la economía o la contabilidad crea altas barreras de entrada que impide que la persona media se forme de manera adecuada. Además, este sector sufre de manera adicional, pues es habitual la aparición de "gurús de las finanzas" que prometen métodos rápidos para conseguir dinero e intentan engañar a la gente aprovechando su mala situación económica. El enfoque didáctico es una parte íntegra del proyecto y se espera que este ayude a la formación académica de los lectores en cuanto a renta variable se refiere, siempre teniendo en cuenta las limitaciones del trabajo.

El fin de la pobreza, el crecimiento económico y la reducción de la desigualdad económica son tres criterios que pueden ir de la mano acompañados. Comúnmente se piensa que la inversión en bolsa es un instrumento para la gente adinerada que les permite conseguir todavía más dinero; esta percepción es errónea. Cualquiera puede participar de la inversión en bolsa, una disciplina hoy en día enormemente democratizada gracias a la facilidad de acceder a un bróker por internet desde cualquier parte del mundo. Muchas de las desigualdades en cuanto a crecimiento económico se deben a una diferente distribución del capital, por ejemplo, los ricos y los ancianos (más propensos a invertir) se protegen mejor de la inflación que la clase baja o media (más propensa a mantener el ahorro en el banco). Estas desigualdades también se acentúan debido a las altas barreras de entradas ya comentadas, cuando no todo el mundo puede permitirse el pago de universidades privadas especializadas en contabilidad y finanzas. Este trabajo espera democratizar la información sobre la inversión en bolsa compartiendo de manera pública y gratuita todos los avances y a hallazgos. Se espera que cualquier persona con interés en el mundo de la inversión pueda acceder a este documento y utilice la bolsa como un método de construcción de riqueza a largo plazo, reduciendo la pobreza y las desigualdades.

Bibliografía

[1] G. S. A. a. K. P. Valavanis, «Surveying stock market forecasting techniques -Part 1: Conventionalmethods,» *Computation Optimization in Economics and Finance Research Compendium,* pp. 49-104, 2013.

[2] G. S. a. K. P. Valavanis, «Surveying stockmarket forecasting techniques - Part 11: Soft computing methods,» *Expert Systems with Applications,* vol. 36, nº 3, pp. 5932-5941, 2009.

[3] K. L. X. Z. L. S. E. N. a. M. L. Y Hu, «Application of evolutionary computation for rule discovery in stock algorithmic trading: A titerature review,» *Applied Soft Computing,* vol. 36, nº 2015, pp. 534-551.

[4] D. Kahneman, Pensar rápido, pensar despacio, Farrar, Straus and Giroux, 2011.

[5] S. S. P. T. K. K. J. Patel, «Predicting stock and stock price index movement using Trend Deteterministic Data Preparation and machine learning techniques,» *Expert Systems with Apptications,* vol. 42, nº 1, pp. 259-268, 2015.

[6] C. H. F. C. P. E. Chong, «Deep leamingnetwotks for stock market analysis and prediction: methodology, data representations, and case studies,» *Expert Systems with Apptications,* vol. 83, pp. 187-205, 2017.

[7] T. S. Quah, «DJIA stock selection assisted by neural network,» *Expert Systems with Applications,* vol. 35, nº 1, pp. 50-58, 2008.

[8] Y. Huang, L. F. Capretz y D. Ho, «Machine Learning for Stock Prediction Based on Fundamental Analysis,» *Electrical and Computer Engineering Publications,* 2021.

[9] Y. Huang, L. F. Capretz y D. Ho, «Neural Network Models for Stock Selection Based on Fundamental Analysis,» *Electrical and Computer Engineering Publications,* 2019.

[10] Z. L. A Namdari, «Integrating fundamental and technical analysis of stock market through mullti-layer perceptron,» *IEEE Technology and Engineering Management Conference,* pp. 1-6, 2018.

[11] T. Bohn, «Improving Long Term Stock Market Prediction with Text Analysis,» 2017. [En línea]. Available: https//ir.lib.uwo.ca/etdl4497. [Último acceso: 13 5 2022].

[12] B. Graham, El Inversor Inteligente, Harper & Brothers, 1949.

[13] J. M. Keynes, The General theory of employment interest and money, London: McMillan, 1936.

[14] G. A. Akerlof y R. J. Shiller, Animal spirits, Princeton University Press, 2009.

[15] A. Dhaoui, S. Bourouis y M. A. Boyacioglu, The Impact of Investor Psychology on Stock Markets: Evidence from France, Journal of Academic Research in Economics, 2013.

[16] N. N. Taleb, El cisne negro: El impacto de lo altamente improbable, Booket, 2012.

[17] G. Soros, The Alchemy of Finance, Wiley, 1987.

[18] F. Nicholson, Price-Earnings Ratios in Relation to Investment Results, Financial Analysts Journal, 1968.

[19] S. Basu, Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A test of the Efficient Markets Hypothesis, Journal of Finance, 1977.

[20] R. K. L. R. Rosenberg B, Persuasive Evidence of Market Inefficiency, Journal of Portfolio Management, 1985.

[21] J. Greenblatt, El pequeño libro que aún vence al mercado, Deusto, 2016.

[22] J. D. Piotroski, Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers, The University of Chicago Graduate School of Business, 2002.

[23] V. Kotu y B. Deshpande, Data Science: Concepts and Practice, Second Edition, Elsevier, 2018.

[24] O. Maimon y L. Rokach, Data Mining and Knowledge Discovery Handbook, Springer US, 2010.

[25] A. Azevedo y M. F. Santos, «KDD, semma and CRISP-DM: A parallel overview,» ResearchGate, 2008.

[26] S. F. Conservancy, «Selenium,» [En línea]. Available: https://www.selenium.dev/documentation/webdriver/. [Último acceso: 12 2 2022].

[27] L. Richardson, «Crummy,» [En línea]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Último acceso: 16 2 2022].

[28] R. Aroussi, «PyPi,» [En línea]. Available: https://pypi.org/project/yfinance/. [Último acceso: 1 3 2022].

[29] «Python,» [En línea]. Available: https://docs.python.org/3/library/threading.html. [Último acceso: 2022].

[30] «Python,» [En línea]. Available: https://docs.python.org/3/library/multiprocessing.html. [Último acceso: 2022].

[31] The pandas development team, «PyData,» [En línea]. Available: https://pandas.pydata.org/docs/user_guide/index.html. [Último acceso: 2022].

[32] N. Developers, «Numpy,» [En línea]. Available: https://numpy.org/doc/stable/user/index.html#user. [Último acceso: 2022].

[33] The SciPy community, «SciPy,» [En línea]. Available: https://docs.scipy.org/doc/scipy/tutorial/index.html#user-guide. [Último acceso: 2022].

[34] D. D. E. F. M. D. a. t. M. d. t. John Hunter, «Matplotlib,» [En línea]. Available: https://matplotlib.org/stable/users/index. [Último acceso: 2022].

[35] M. Waskom, «PyData,» [En línea]. Available: https://seaborn.pydata.org/api.html. [Último acceso: 2022].

[36] Plotly, «Plotly,» [En línea]. Available: https://dash.plotly.com. [Último acceso: 2022].

[37] scikit-learn developers, «Scikit-learn,» [En línea]. Available: https://scikit-learn.org/stable/modules/classes.html. [Último acceso: 2022].

[38] xgboost developers, «Readthedocs,» [En línea]. Available: https://xgboost.readthedocs.io/en/stable/. [Último acceso: 2022].

[39] AutoViML, «Github,» [En línea]. Available: https://github.com/AutoViML/featurewiz. [Último acceso: 2022].

[40] Streamlit Inc, «Streamlit,» [En línea]. Available: https://docs.streamlit.io. [Último acceso: 2022].

[41] «Investopedia,» 10 Septiembre 2021. [En línea]. Available: https://www.investopedia.com/terms/m/modernportfoliotheory.asp#:~:text=The%20modern%20portfolio%20theory%20(MPT)%20is%20a%20practical%20method%20for,an%20acceptable%20level%20of%20risk.&text=A%20key%20component%20of%20the,low%20risk%20and%20low%20return.. [Último acceso: 2021 4 22].

[42] W. Buffet, «The Superinvestors of Graham and Doddsville,» *Columbia Business School Magazine,* 1984 .

[43] S. Klarman, Margin of Safety: Risk-Averse Value Investing Strategies for the Thoughtful Investor, HarperCollins, 1991.

[44] V. Chernikov, «Roic.ai,» [En línea]. Available: https://roic.ai. [Último acceso: 2 2022].

[45] Yahoo, «Yahoo Finance,» [En línea]. Available: https://finance.yahoo.com. [Último acceso: 1 2022].

[46] «Macrotrends,» [En línea]. Available: https://www.macrotrends.net.

[47] «FreddieMac,» [En línea]. Available: https://www.freddiemac.com/pmms/pmms30.

[48] «ST. Louis FED,» [En línea]. Available: https://fred.stlouisfed.org/series/USSTHPI.

[49] «Stooq,» [En línea]. Available: https://stooq.com/q/d/?s=%5Ecry&c=0&i=y.

[50] «U.S. BUREAU OF LABOR STATISTICS,» [En línea]. Available: https://data.bls.gov/pdq/SurveyOutputServlet.

[51] «Federal Reserve Bank of Atlanta,» [En línea]. Available: https://www.atlantafed.org/chcs/wage-growth-tracker.

[52] N. N. Taleb, Antifrágil: Las cosas que se benefician del desorden, Paidós, 2013.

[53] E. Norland, «CME Group,» 26 Octubre 2020. [En línea]. Available: https://www.cmegroup.com/education/featured-reports/equities-comparing-russell-2000-vs-sandp-500.html. [Último acceso: 1 5 2022].

[54] M. Nabipour, P. Nayyeri, H. Jabani, S. Shamshirband y A. Mosavi, «Deep Learning for Stock Market Prediction,» Preprints, 2020.

Anexo 1: Cálculo de los ratios financieros

1. **Tax Rate** = Income Tax expense (Gain) / Income Before Tax
2. **Deferred Revenue** = Deferred Revenue (Current) + Deferred Revenue (Non-Current)
3. **Net Interest Income** = Interest Income - Interest Expense (Gain)
4. **Free Cash Flow to the Firm** = Free Cash Flow - Interest Expense (Gain) * (1 - Tax Rate)
5. **Tangible Assets** = Total Assets - Goodwill and Intangible Assets
6. **Adjusted Operating Income** = Gross Profit - Selling, General and Administrative Exp. - Depreciation and Amortization
7. **EBIT** = Income Before Tax + Interest Expense (Gain)
8. **EBITDA** = EBIT + Depreciation and Amortization
9. **Operating Cash Flow** = Operating Income + Depreciation and Amortization - Income Tax expense (Gain) + Change in Working Capital
10. **Common Book Value** = Total Stockholders Equity - Preferred Stock
11. **Tangible Book Value** = Total Stockholders Equity - Goodwill and Intangible Assets
12. **Common Tangible Book Value** = Tangible Book Value - Preferred Stock
13. **Free Cash Flow Ratio** = Free Cash Flow / Revenue
14. **Selling, General and Administrative Exp. Ratio** = Selling, General and Administrative Exp. / Revenue
15. **Research and Development Exp. Ratio** = Research and Development Exp. / Revenue
16. **Other Expenses Ratio** = Other Expenses / Revenue
17. **Net Interest Income Ratio** = Net Interest Income / Revenue
18. **Depreciation and Amortization Ratio** = Depreciation and Amortization / Revenue
19. **EBIT Margin** = EBIT / Revenue
20. **Adjusted Operating Margin** = Adjusted Operating Income / Revenue
21. **Operating Cash Flow Margin** = Operating Cash Flow / Revenue
22. **Cost-to-Income Ratio** = Operating Expenses / Operating Income
23. **Operating Expense Ratio** = Operating Expenses / Revenue
24. **Selling, General and Administrative Exp. Ratio** = Selling, General and Administrative Exp. / Gross Profit
25. **Revenue per share** = Revenue / Weighted Average Shares Outstanding
26. **Operating Income per share** = Operating Income / Weighted Average Shares Outstanding
27. **FCF per share** = Free Cash Flow / Weighted Average Shares Outstanding
28. **CAPEX per share** = CAPEX / Weighted Average Shares Outstanding
29. **Book value per share** = Total Stockholders Equity / Weighted Average Shares Outstanding
30. **Total Debt** = Short-Term Debt + Long-Term Debt
31. **Dividends per share** = Dividends Paid / Weighted Average Shares Outstanding
32. **Financial leverage** = Total Debt / Total Stockholders Equity
33. **Cash to Debt Ratio** = Cash and Cash Equivalents / Total Debt
34. **Cash & Investments** = Cash and Short-Term Investments + Investments
35. **Cash & Investments to Debt Ratio** = Cash & Investments / Total Debt
36. **Net Debt** = Total Debt - Cash and Short-Term Investments
37. **Net Debt w/Investments** = Net Debt - Investments

38. **Working Capital** = Total Current Assets - Total Current Liabilities
39. **Current Ratio** = Total Current Assets / Total Current Liabilities
40. **Quick Ratio** = (Cash and Cash Equivalents - Accounts Receivable)/ Total Current Liabilities
41. **Cash to Current Assets** = Cash and Short-Term Investments / Total Current Assets
42. **Cash to Assets** = Cash and Short-Term Investments / Total Assets
43. **Debt to Equity** = Total Current Liabilities / Total Stockholders Equity
44. **Debt to Assets** = Total Current Liabilities / Total Assets
45. **Interest Coverage** = EBIT / Interest Expense (Gain)
46. **Current Liability Coverage Ratio** = (Operating Cash Flow - Dividends Paid)/ Total Current Liabilities
47. **Cash Ratio** = Cash and Short-Term Investments / Total Current Liabilities
48. **Net Working Capital to Assets** = Working Capital / Total Assets
49. **Long Term Debt Ratio** = Total Non-Current Liabilities / Total Assets
50. **Total Debt/ Assets** = Total Debt / Total Assets
51. **Net Debt/ Assets** = Net Debt / Total Assets
52. **Total Debt/ Equity** = Total Debt / Total Stockholders Equity
53. **Total Liabilities/ Equity** = Total Liabilities / Total Stockholders Equity
54. **Net Debt/ Equity** = Net Debt / Total Stockholders Equity
55. **Equity Ratio** = Total Stockholders Equity / Total Assets
56. **Equity Multiplier** = Total Assets / Total Stockholders Equity
57. **Debt/ Tangible Book Value** = Total Liabilities / Tangible Book Value
58. **Net Debt/ EBITDA** = Net Debt / EBITDA
59. **Cash to Debt** = Cash and Short-Term Investments / Total Debt
60. **Cash Flow Coverage Ratio** = Operating Cash Flow / Total Liabilities
61. **Free Cash Flow/ Long Term Debt** = Free Cash Flow / (Long-Term Debt + Capital Lease Obligations)
62. **Debt Leverage Ratio** = Total Liabilities / EBITDA
63. **Interest Expense to Debt Ratio** = Interest Expense (Gain) / Total Debt
64. **Cash Sale**s = Revenue - Accounts Receivable
65. **Cash Revenue Adjustment** = Accounts Receivable / Deferred Revenue
66. **CFO/ Net Income** = Operating Cash Flow / Net Income
67. **Return on Equity** = Net Income / Common Book Value
68. **NOPAT** = Operating Income * (1 - Tax Rate)
69. **Invested Capital** = Total Liabilities And Stockholders Equity - Total Current Liabilities
70. **Return on Invested Capital** = NOPAT / Invested Capital
71. **Capital Employ**ed = Total Assets - Total Current Liabilities
72. **ROCE** = EBIT / Capital Employed
73. **CFROI** = Operating Cash Flow / Capital Employed
74. **Asset Turnover Ratio** = Revenue / Total Assets
75. **Inventory Turnover** = COGS / Inventory (Balance)
76. **Receivables Turnover** = Revenue / Accounts Receivable
77. **Cash Turnover Ratio** = Revenue / Cash and Cash Equivalents
78. **Gross Profitability Ratio** = Gross Profit / Total Assets
79. **Tangible Gross Profitability Ratio** = Gross Profit / Tangible Assets
80. **Cash Return on Invested Capital** = Free Cash Flow to the Firm / Invested Capital
81. **Adjusted Return on Capital Employed** = Adjusted Operating Income / Capital Employed
82. **Cash Return on Capital Employed** = Free Cash Flow to the Firm / Capital Employed
83. **EBIT Return on Assets** = EBIT / Total Assets
84. **EBIT Return on Tangible Assets** = EBIT / Tangible Assets

85. **Return on Assets** = NOPAT / Total Assets
86. **Return on Tangible Equity** = Net Income / Common Tangible Book Value
87. **Cash Return on Equity** = Operating Cash Flow / Total Stockholders Equity
88. **Return on Retained Earnings** = Net Income / Retained Earnings
89. **R&D / Assets** = Research and Development Exp. / Total Assets
90. **R&D / Book** = Research and Development Exp. / Total Stockholders Equity
91. **CapEx / Assets** = -CAPEX / Total Assets
92. **CapEx / Fixed Assets** = -CAPEX / PP&E
93. **Retained Earnings / Total Assets** = Retained Earnings / Total Assets
94. **Inventory / Assets** = Inventory (Balance) / Total Assets
95. **Accounts Receivable / Assets** = Accounts Receivable / Total Assets
96. **Plowback Ratio** = (Net Income - Dividends Paid)/ Net Income
97. **Dividend & Repurchase / FCF** = (Dividends Paid - Common Stock Repurchased - Common Stock Issued) / Free Cash Flow
98. **Dividend & Repurchase / EBITDA** = (Dividends Paid - Common Stock Repurchased - Common Stock Issued) / EBITDA
99. **PER** = Close (Año anterior) / EPS
100. **PER_3** = PER (Media de los últimos 3 años)
101. **PER_5** = PER (Media de los últimos 5 años)
102. **PER_10** = PER (Media de los últimos 10 años)
103. **Intrinsic Value** = EPS * PER
104. **Intrinsic Value 3** = EPS * PER_3
105. **Intrinsic Value 5** = EPS * PER_5
106. **Intrinsic Value 10** = EPS * PER_10
107. **Price to Book** = Close (Año anterior) / Book value per share
108. **Price to Book 3** = Price to Book (Media de los últimos 3 años)
109. **Price to Book 5** = Price to Book (Media de los últimos 5 años)
110. **Price to Book 10** = Price to Book (Media de los últimos 10 años)
111. **IV Price to Book** = Book value per share * Price to Book
112. **IV Price to Book 3** = Book value per share * Price to Book 3
113. **IV Price to Book 5** = Book value per share * Price to Book 5
114. **IV Price to Book 10** = Book value per share * Price to Book 10
115. **Price to Free Cash Flow** = Close (Año anterior) / FCF per share
116. **Price to Free Cash Flow 3** = Price to Free Cash Flow (Media de los últimos 3 años)
117. **Price to Free Cash Flow 5** = Price to Free Cash Flow (Media de los últimos 5 años)
118. **Price to Free Cash Flow 10** = Price to Free Cash Flow (Media de los últimos 10 años)
119. **IV Price to Free Cash Flow** = FCF per share * Price to Free Cash Flow
120. **IV Price to Free Cash Flow 3** = FCF per share * Price to Free Cash Flow 3
121. **IV Price to Free Cash Flow 5** = FCF per share * Price to Free Cash Flow 5
122. **IV Price to Free Cash Flow 10** = FCF per share * Price to Free Cash Flow 10
123. **Grahams Number** = sqrt(22.5*EPS * Common Tangible Book Value)
124. **Grahams Number 3** = Grahams Number (Media de los últimos 3 años)
125. **Grahams Number 5** = Grahams Number (Media de los últimos 5 años)
126. **Grahams Number 10** = Grahams Number (Media de los últimos 10 años)
127. **Price to Grahams Number** = Close (Año anterior) / Grahams Number
128. **Price to Grahams Number 3** = Price to Grahams Number (Media de los últimos 3 años)
129. **Price to Grahams Number 5** = Price to Grahams Number (Media de los últimos 5 años)
130. **Price to Grahams Number 10** = Price to Grahams Number (Media de los últimos 10 años)

131. **IV Grahams Number** = Grahams Number * Price to Grahams Number
132. **IV Grahams Number 3** = Grahams Number * Price to Grahams Number 3
133. **IV Grahams Number 5** = Grahams Number * Price to Grahams Number 5
134. **IV Grahams Number 10** = Grahams Number * Price to Grahams Number 10
135. **PEG** = PER / EPS Diluted growth 3 year average
136. **PEG 3** = PEG (Media de los últimos 3 años)
137. **PEG 5** = PEG (Media de los últimos 5 años)
138. **PEG 10** = PEG (Media de los últimos 10 años)
139. **IV PEG** = EPS Diluted growth 3 year average * PER
140. **IV PEG 3** = EPS Diluted growth 3 year average * PER_3
141. **IV PEG 5** = EPS Diluted growth 3 year average * PER_5
142. **IV PEG 10** = EPS Diluted growth 3 year average * PER_10

Este documento esta firmado por

Firmante	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
Fecha/Hora	Thu Jun 02 15:40:55 CEST 2022
Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
Numero de Serie	561
Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)