Universidad Pontificia Comillas ICADE

INVERSIONES SOSTENIBLES

Autor: Carmen López del Hierro Cabrera

Director: Lourdes Fernández Rodríguez

MADRID | junio 2024

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Carmen López del Hierro Cabrera, estudiante de ADE + Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado “Inversiones sostenibles", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación.

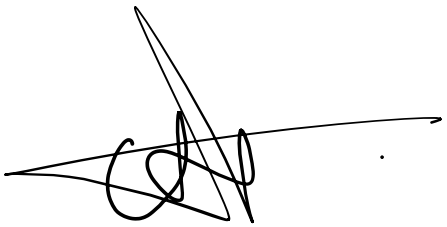
1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
3. **Interpretador de código:** Para realizar análisis de datos preliminares.
4. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
5. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
6. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
7. **Generador de problemas de ejemplo:** Para ilustrar conceptos y técnicas.
8. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
9. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han

dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 26/05/2024

Firma: Carmen López del Hierro Cabrera



RESUMEN

En el presente Trabajo de Fin de Grado se explora el desarrollo y la aplicación de un modelo analítico basado en algoritmos de aprendizaje automático para identificar y evaluar empresas en función de su compromiso con la sostenibilidad. Utilizando una amplia variedad de indicadores financieros y de sostenibilidad ambiental, social y de gobernanza, se implementan técnicas de *clustering* para segmentar empresas y determinar patrones entre su rendimiento y su responsabilidad corporativa. Los resultados indican una correlación positiva entre altas puntuaciones de sostenibilidad y un rendimiento financiero superior, ofreciendo a los inversores una herramienta robusta para la toma de decisiones de inversión responsable. El estudio también discute las limitaciones del modelo actual, incluyendo los desafíos relacionados con la disponibilidad y calidad de los datos, y propone direcciones futuras para la investigación en la integración de prácticas de sostenibilidad en la evaluación corporativa. El enfoque realizado no solo beneficia a los inversores al proporcionar información valiosa sobre la viabilidad financiera y ética de sus inversiones, sino que también promueve un impacto positivo en la sociedad y el medio ambiente.

PALABRAS CLAVE

Responsabilidad Social Corporativa; Inversión responsable; Economía sostenible; Modelo analítico; Aprendizaje automático; *Clustering*; Criterios ESG; Rendimiento financiero

ABSTRACT

In this Bachelor's Thesis, the development and application of an analytical model based on machine learning algorithms to identify and evaluate companies based on their commitment to sustainability is explored. Utilizing a wide range of financial and environmental, social, and governance sustainability indicators, *clustering* techniques are implemented to segment companies and determine patterns between their performance and corporate responsibility. The results indicate a positive correlation between high sustainability scores and superior financial performance, providing investors with a robust tool for making responsible investment decisions. The study also discusses the limitations of the current model, including challenges related to data availability and quality, and proposes future research directions for integrating sustainability practices into corporate evaluation. This approach not only benefits investors by providing valuable information about the financial and ethical viability of their investments but also promotes a positive impact on society and the environment.

KEYWORDS

Corporate Social Responsibility; Responsible Investment; Sustainable Economy; Analytical Model; Machine Learning; Clustering; ESG Criteria; Financial Performance

Índice

Índice de figuras

Índice de tablas

1. Introducción

1.1. Justificación

La realización de este Trabajo de Fin de Grado surge de la necesidad de desarrollar un modelo analítico que aborde la identificación y evaluación de empresas que sigan unos criterios de sostenibilidad, para así, poder invertir en ellas, con el objetivo de fomentar la inversión responsable y contribuir a un desarrollo económico más sostenible.

En el contexto actual, donde la sostenibilidad y la responsabilidad social corporativa están adquiriendo una importancia cada vez más relevante, existe una demanda creciente de herramientas y enfoques que permitan a los inversores tomar decisiones informadas que estén alineadas con consideraciones ambientales, sociales y de gobernanza (ESG por sus siglas en inglés). Estos criterios ESG están siendo considerados cada vez más como indicadores clave de desempeño y riesgo para las empresas, además de un factor clave en el rendimiento empresarial y la innovación estratégica (Porter & Kramer, 2006).

El uso de algoritmos de aprendizaje automático proporciona una ventaja significativa en este sentido, ya que permiten analizar grandes volúmenes de datos y detectar patrones complejos que podrían pasar desapercibidos si se utilizan enfoques tradicionales. Al considerar una amplia gama de factores financieros y sostenibles, estos algoritmos ofrecen una herramienta poderosa para clasificar a las empresas en distintos niveles de sostenibilidad.

Esta clasificación no solo facilita a los inversores la identificación de oportunidades de inversión que se alineen con sus valores y objetivos, sino que también les ayuda a mitigar los riesgos asociados con las inversiones no sostenibles ya que las empresas que adoptan políticas de sostenibilidad consistentes tienden a mostrar un mejor desempeño operativo y a largo plazo en comparación con aquellas que no lo hacen (Eccles, Ioannou, & Serafeim, 2014).

En resumen, el desarrollo de este modelo analítico representa un paso importante hacia la promoción de prácticas de inversión más responsables y la construcción de un futuro económico más sostenible. Su aplicación no solo beneficia a los inversores, sino que también promueve un impacto positivo más amplio en la sociedad y el medio ambiente.

1.2. Objetivos de la investigación

Los objetivos de este Trabajo de Fin de Grado (TFG) son los siguientes:

- Crear un modelo analítico que permita identificar empresas atractivas para la inversión. Estas empresas no solo serán evaluadas por su rendimiento financiero, sino también por su compromiso con la promoción de una economía más sostenible, mediante el cumplimiento de los criterios ESG.
- Facilitar a los inversores la oportunidad de invertir en dichas empresas, impulsando así la inversión responsable y contribuyendo al desarrollo de un mundo sostenible.

1.3. Metodología y estructura

Para lograr los objetivos previamente establecidos se realizará una primera revisión exhaustiva de la literatura existente sobre inversiones sostenibles, a través de diversas fuentes como artículos académicos, informes técnicos y científicos procedentes de plataformas como Google Académico, y sitios web de entidades corporativas, como el BBVA, el Santander o Repsol. También, se llevará a cabo la recopilación de datos financieros y de sostenibilidad de empresas, mediante plataformas *online* que recogen conjuntos de datos públicos, como Kaggle, para su posterior procesamiento. A continuación, se desarrollará un modelo cuantitativo predictivo, en la plataforma R Studio, utilizando algoritmos de aprendizaje automático para clasificar a las empresas en distintos niveles de sostenibilidad. La metodología concluye con la discusión de los resultados y las conclusiones.

2. Inversiones sostenibles y criterios ESG

2.1. Qué son las Inversiones sostenibles

El concepto de inversión sostenible ha existido durante más tiempo del que comúnmente se piensa. Sin embargo, su popularidad ha experimentado un notable aumento en los últimos tiempos. En la actualidad, es crucial para cualquier empresa generar oportunidades de inversión con un impacto positivo, que se rigen por unos criterios ESG.

Aunque los fondos de inversión sostenible superaron los 4.000 en diciembre de 2020, experimentando un significativo crecimiento debido a la pandemia del Covid-19, el origen de la inversión sostenible se remonta a la década de 1970, cuando se establecieron los primeros fondos de inversión sostenible en respuesta al creciente interés en un comportamiento corporativo responsable. El accidente que sufrió el barco Exxon Valdez al encallar en la Bahía del príncipe Guillermo en Alaska, en 1980, que provocó un derrame de 41 millones de litros de crudo en el mar, impulsó la formación de la Coalición para Economías Medioambientales Responsables (CERES por sus siglas en inglés), lo que llevó el concepto de sostenibilidad a las instituciones. A partir de ese momento, han sucedido numerosos acuerdos entre organismos internacionales: en 1990, se firmó el Protocolo de Kioto para reducir las emisiones de dióxido de carbono; en 2000, se estableció el Pacto Mundial de las Naciones Unidas (ONU) para integrar cuestiones ambientales, sociales y de buen gobierno corporativo en el mundo empresarial; y en 2015, 193 países aprobaron los 17 Objetivos de Desarrollo Sostenible (ODS), un plan de acción para lograr un futuro más próspero y sostenible para todos (Santander Asset Management).

A día de hoy, el concepto de inversión sostenible se resume en invertir en activos financieros que tengan en consideración criterios ambientales, sociales y de gobernanza (ESG), además de los criterios financieros tradicionales. Esta estrategia de inversión no solo busca obtener beneficios financieros, sino también generar un impacto positivo en el medio ambiente y la sociedad. De esta manera, al considerar factores ambientales, sociales y de gobernanza en la toma de decisiones de inversión, los inversores pueden contribuir a un futuro más sostenible mientras buscan obtener retornos financieros atractivos (BBVA, 2022).

Existen algunos puntos clave a tener en cuenta sobre las inversiones sostenibles. En primer lugar, cabe destacar que los criterios ESG abarcan a su vez aspectos ambientales (*Enviormental*), como el cambio climático y la eficiencia energética, sociales (*Social*), los derechos humanos, la igualdad y la diversidad, y de gobernanza (*Governance*), la transparencia y la ética empresarial (Repsol, 2023).

En segundo lugar, es importante saber que existen diversos enfoques para las inversiones sostenibles, desde inversiones básicas ESG hasta las inversiones generadoras de impacto.

Además, se ha demostrado que las inversiones sostenibles pueden ser igualmente rentables e incluso superar a las inversiones tradicionales, y que pueden ayudar también a identificar y gestionar riesgos a largo plazo. Esto es un aspecto clave que permite el futuro desarrollo de este tipo de inversiones (Santander Asset Management).

También cabe destacar que, en los últimos años, ha habido un notable incremento en el mercado de inversiones sostenibles, con un aumento en la compra de este tipo de activos de 12.462 billones de dólares en 2020 a nivel mundial, en comparación con el año 2016 (Ver Figura 1). Lo que demuestra el creciente interés de los inversores en incorporar consideraciones ESG en sus carteras, y lo que permite el futuro desarrollo de más productos financieros sostenibles (Claver, 2023).

Figura 1: Activos de inversión sostenible globales, 2016-2020 (en miles de millones de dólares estadounidenses)

REGION	2016	2018	2020
Europe*	12,040	14,075	12,017
United States	8,723	11,995	17,081
Canada	1,086	1,699	2,423
Australasia*	516	734	906
Japan	474	2,180	2,874
Total (USD billions)	**22,839**	**30,683**	**35,301**

Fuente: *How asset managers are working on sustainable investing*. ROBECO, 2023

Por último, como consecuencia de dicho crecimiento, la regulación y los estándares relacionados con las inversiones sostenibles han aumentado en los últimos años, como los Principios para la Inversión Responsable (PRI por sus siglas en inglés) de las Naciones Unidas, que se definen como un conjunto de seis principios que guían a las instituciones

financieras en la incorporación de consideraciones ESG en sus prácticas de inversión. Estos principios incluyen la integración de asuntos ESG en el análisis de procesos de toma de decisiones y el ejercicio de la propiedad activa, la promoción de la divulgación adecuada de dichos asuntos por parte de las entidades en las que invierten, el impulso a la aceptación e implementación de los principios en el sector de las inversiones, el trabajo colaborativo para mejorar la aplicación de los principios y la presentación de informes periódicos sobre las actividades y el progreso en la aplicación de los principios (United Nations, 2021).

2.2. Enfoques y estrategias de inversión sostenible

Las empresas tienen distintos enfoques para llevar a cabo sus inversiones sostenibles. Al analizar estos enfoques en función de las distintas estrategias de inversión y de sus objetivos para apoyar activamente la transición hacia una economía más sostenible, se pueden clasificar las inversiones en cuatro categorías:

1) Inversiones básicas ESG
2) Inversiones avanzadas ESG
3) Inversiones alineadas con el impacto
4) Inversiones generadoras de impacto

Estas categorías no solo reflejan el nivel de ambición de las inversiones para contribuir a una economía más sostenible a largo plazo, sino que también valoran su impacto real y los resultados positivos que tienen (Busch, Pruessner, Oulton, Palinska, & Garrault, Methodology for Eurosif Market Studies on Sustainability-related, 2024).

- Las inversiones básicas ESG se enfocan en integrar los factores ambientales, sociales y de gobernanza en su proceso de inversión. Estas inversiones suelen priorizar retornos a largo plazo, y se centran en mitigar riesgos y en excluir empresas que no cumplen con ciertos estándares éticos o de sostenibilidad. Utilizan criterios de criba negativos o positivos para seleccionar o descartar empresas en función de su comportamiento, como, por ejemplo; descartar aquellas que se dediquen a la producción de energía fósil o que en su gestión no respeten los derechos humanos.
- Las inversiones avanzadas ESG van un paso más allá de la simple integración de factores ESG en la valoración de las inversiones. Se centran en analizar de manera más exhaustiva los riesgos y oportunidades asociados a estos factores y en cómo

impactan en la rentabilidad a largo plazo. Estas inversiones pueden influir de manera indirecta en la transición hacia una economía más sostenible. Además de utilizar criterios de criba, realizan mediciones del desempeño ESG de las empresas o activos en los que invierten para tomar decisiones informadas.

- Las inversiones alineadas con el impacto tienen como objetivo principal contribuir a generar cambios positivos en la sociedad y el medio ambiente. Para ello, seleccionan empresas o activos que tienen un impacto social o ambiental positivo y buscan evidenciar y medir este impacto real a través de métricas específicas. Estas inversiones suelen utilizar criterios de criba positivos y negativos para alinear su cartera con sus objetivos de impacto.
- Las inversiones generadoras de impacto van un paso más allá al no solo buscar obtener buenos retornos financieros, sino también contribuir activamente a generar un impacto positivo en la sociedad y el medio ambiente. Estas inversiones utilizan una combinación de criba y de gestión activa para influir en el comportamiento de las empresas en las que invierten y mejorar su impacto. Además, proporcionan evidencias claras y transparentes de cómo contribuyen a mejorar los aspectos sociales o ambientales de sus inversiones.

2.3. Importancia de la sostenibilidad en las decisiones de inversión

La importancia de la sostenibilidad en las decisiones de inversión se puede encontrar en numerosos factores fundamentales.

En primer lugar, la integración de criterios ambientales, sociales y de gobernanza en las estrategias de inversión ayuda a mitigar riesgos financieros y a identificar oportunidades de crecimiento a largo plazo. Estudios existentes señalan que el 90% de los líderes empresariales consideran la sostenibilidad como importante, lo que refleja una conciencia creciente dentro de las personas con más influencia de las empresas sobre el impacto que estas generan en la rentabilidad y la reputación de los negocios (Rafi, 2021).

Además, la adopción de prácticas sostenibles por parte de las empresas, como el uso de materiales sostenibles y la mejora de la eficiencia energética, demuestran un compromiso con el medio ambiente, lo cual puede aumentar el atractivo para los inversores. El 67% de las empresas ya están utilizando estos enfoques (Deloitte, 2022).

Desde la perspectiva de los inversores institucionales, un estudio de IMD revela que el 74% de ellos consideran más probable dejar de invertir en empresas con un desempeño deficiente en sostenibilidad (Haanaes & Olynec, 2022). Por tanto, considerar los criterios ESG al evaluar el potencial de inversión de una empresa y cómo esta gestiona sus riesgos y oportunidades relacionados con la sostenibilidad, ya es un factor importantísimo a la hora de tomar las decisiones de inversión. Como consecuencia, las empresas que adoptan prácticas sostenibles pueden atraer un mayor interés por parte de estos inversores y acceder a capital adicional.

Finalmente, los cambios en las preferencias de los consumidores también están impulsando la importancia de la sostenibilidad en las decisiones de inversión. El 88% de los consumidores, se sentirán más leales hacia una empresa que apoye cuestiones sociales o medioambientales. Esta conexión directa entre la sostenibilidad y la lealtad del cliente resalta la importancia de la sostenibilidad no solo desde una perspectiva financiera, sino también desde una perspectiva de marca y reputación (Butler, 2018).

En resumen, la sostenibilidad se ha convertido en un factor de vital importancia en las decisiones de inversión por su capacidad para generar retornos financieros sostenibles, mitigar riesgos y responder a las expectativas de los inversores y consumidores.

2.4. Criterios para clasificar una inversión como sostenible

Como se ha mencionado previamente, una inversión se considera sostenible cuando se destina a activos que cumplen con criterios ambientales, sociales y de gobernanza. En esta sección, se profundizará un poco más en la metodología que utiliza ESG Enterprise[1], para clasificar a las empresas de esta manera (ESG Enterprise).

En primer lugar, se calculan más de 50 medidas ESG de una empresa, basadas en su información pública, que se dividen en 10 categorías para facilitar el proceso de puntuación ESG. Estas categorías abarcan los tres pilares ESG fundamentales: el ambiental, social y de gobierno corporativo, y son las siguientes: el capital natural, la innovación y el cambio climático, en la parte ambiental, en lo social, el capital social, el

[1] - ESG Enterprise es una de las empresas de análisis de datos, herramientas y software ESG de más rápido crecimiento con más de 250.000 evaluaciones de organizaciones públicas, privadas, municipales y sin fines de lucro en 100 países

capital humano, el empoderamiento y las necesidades básicas, y, por último, en la parte de gobierno corporativo, se encuentran el liderazgo, la gestión y el modelo de negocio.

La puntuación resultante refleja el desempeño integral de la empresa, teniendo en cuenta su eficiencia y su compromiso ESG. Dicha puntuación se normaliza en una escala de 0 a 1.000 y se les asigna una calificación de letra de D a AAA, con el objetivo de obtener una medida más clara y comparable con el resto de las empresas.

Por otra parte, además de evaluar los pilares ESG, la metodología también considera las controversias de fuentes mediáticas, como la prensa o las noticias televisivas. Dichas controversias se examinan cuidadosamente, ya que pueden influir significativamente en la percepción del desempeño ESG de una empresa, especialmente si es negativa debido, por ejemplo, a un escándalo. Las controversias que sean significativas se integran en la puntuación general para equilibrar y complementar la evaluación. La puntuación final se ajusta tomando el promedio de la puntuación ESG general y la de la empresa durante el año financiero. Esto garantiza una evaluación más justa y precisa del desempeño de la empresa en este tipo de situaciones.

3. Modelo analítico

3.1. Introducción

Problema a resolver

Nos encontramos en un entorno en el cuál la sostenibilidad y la responsabilidad empresarial están ganando cada vez más relevancia dado los numerosos beneficios asociados a sus prácticas. Es por ello, por lo que los inversores necesitan herramientas que les permitan tomar decisiones informadas y alinear estas con criterios sostenibles. Mediante la aplicación de técnicas de *clustering*, se puede optimizar el tiempo empleado en la valoración de cada empresa usando la segmentación. Con un modelo de aprendizaje no supervisado realizado en R Studio, se espera encontrar claros subgrupos naturales que puedan ayudar a los inversores a decidir de manera más fundamentada, rápida y con menor margen de error en qué empresas invertir que sean tanto buenas en sus rendimientos financieros, como en su seguimiento de las prácticas ESG.

Base de datos

Para realizar este modelo analítico era necesario obtener una base de datos que tuviera datos tanto financieros como de los rendimientos ESG de distintas empresas. Para conseguirla, se ha tenido que llevar a cabo la unificación de dos *datasets* distintos, ambos procedentes de la plataforma Kaggle, ya que no había ninguno que ya contuviera toda la información que se requería.

La primera base de datos encontrada contaba con 21 columnas (variables) y 700 filas (empresas). Las variables recogían tanto información básica de las compañías: "ticker", "company name", "currency", "exchange", "industry", "logo URL" y "website URL", como información sobre su comportamiento en temas de ESG: "environment_score", "environment_grade", environment_level, "social_score", "social_grade", "social_level", "governance_score", "governance_grade", "governance_level", "total_score, "total_grade", "total_level", "last processing date of the ESG data" y "CIK".

Por otro lado, la segunda base de datos tenía únicamente un total de 14 variables: "Ticker", "Name", "Sector", "Price", "Price/Earnings", "Dividend Yield", "Earnings/Share", "52 Week Low", "52 Week High", "Market Cap", "EBITDA", "Price/Sales", "Price/Book" y "SEC Filings", y 505 empresas.

Una vez que se han descargado ambos ficheros de la web en formato Excel, era necesario unificarlos de tal forma que la base de datos final contuviera únicamente aquellas empresas que estaban en ambos *datasets* ya que eran de las que se iba a tener toda la información necesaria para realizar el análisis.

Para realizar la unificación, se siguieron los siguientes pasos:

1. Descarga de ambos conjuntos de datos en formato Excel.
2. Unión en un único archivo Excel, utilizando dos pestañas distintas.
3. Empleo de la fórmula XLOOKUP para combinar todas las variables en una sola pestaña, utilizando la variable "Ticker" de las compañías como enlace de unión.
4. Eliminación de las compañías que no estaban presentes en ambos conjuntos de datos.

En la base de datos unificada, se realizó también la eliminación de algunas variables que contenían información irrelevante para el análisis o redundante, ya que esto podía provocar una multicolinealidad demasiado alta, lo cual tiene efectos negativos en los modelos de Machine Learning, tales como la dificultad para interpretar sus coeficientes.

La base de datos final resultante será la se utilizará para realizar el análisis y constará de 10 variables, descritas a continuación (Ver Tabla 1), y 365 empresas.

Tabla 1: Análisis de las variables

Variable	Valores	Descripción
Ticker	Ej: DIS	Código abreviado utilizado para identificar las acciones de una empresa en un mercado bursátil
Total_score	0 – 2k	Desempeño ESG integral de la compañía. A mayor puntuación, mejores niveles ESG tiene la compañía
Price	2,8 – 1,8k	Precio de negociación de las acciones de la empresa
Price/Earnings	(252) - 520	Métrica que indica la relación entre el precio de las acciones y las ganancias por acción
Dividend Yield	0 – 12,7	Rendimiento de dividendos, que representa la relación entre el pago de dividendos anual y el precio de las acciones
Earnings/Share	(28,0) – 44,1	Métrica de la rentabilidad de una empresa, calculada como las ganancias divididas por el número de acciones en circulación
Market Cap	2,6b – 810,0b	Representa el producto del precio actual de las acciones y el número total de acciones en circulación
EBITDA	(5,1)b – 79,0b	Ganancias antes de intereses, impuestos y amortización. Es una medida del rendimiento operativo de una empresa
Price/Share	0 – 20,1	Compara el precio de las acciones con los ingresos por acción de la empresa
Price/Book	0 – 1,3k	Compara el precio de las acciones con su valor contable por acción, un indicador del valor relativo de las acciones

Fuente: Elaboración propia

Pre-procesamiento de los datos

Una vez que se había terminado de preparar la base de datos final con una dimensión de 365 x 10, era importante realizar una limpieza de los datos antes de aplicar cualquier técnica de *clustering*. La limpieza de datos implica identificar y corregir errores (como NAs), eliminar *outliers*, y estandarizar o normalizar los datos de aquellas variables que sean numéricas.

Se analizó, por tanto, si alguno de los registros contenía *missing values* para eliminarlos y evitar así problemas en el posterior cálculo de distancias necesario para realizar el modelo. A continuación, se quitaron los *outliers* de aquellas variables en donde era necesario para que sus valores extraordinarios no perjudicasen al análisis. Y finalmente, se estandarizaron los datos numéricos utilizando la función "Scale", para poder normalizar las variables, de tal manera que todas tuvieran una media de 0 y una desviación típica de 1, asegurando así que todas las variables contribuían de manera equitativa al posterior cálculo de distancias. El conjunto de datos resultante, llamado "datanorm", quedaba ya preparado para poder proceder con el análisis.

Análisis exploratorio de las variables

Se calcularon los resúmenes estadísticos de las variables (Ver Figura 2), para poder comprobar, si el mínimo, la media y el máximo de cada una de ellas tenía sentido. Como previamente se había llevado a cabo la eliminación de *outliers*, no se encontró nada que hubiera que corregir y se podía comenzar con la aplicación de los métodos de *clustering*.

Figura 2: Resumen estadístico de las variables

Ticker	Total_score	Price	Price/Earnings	Dividend Yield	Earnings/Share	Market Cap
Length:316	Min. : 600.0	Min. : 10.06	Min. :-59.46	Min. :0.0000	Min. :-6.890	Min. :4.654e+09
Class :character	1st Qu.: 945.8	1st Qu.: 52.51	1st Qu.: 16.03	1st Qu.:0.9184	1st Qu.: 1.650	1st Qu.:1.328e+10
Mode :character	Median :1114.0	Median : 75.56	Median : 19.81	Median :1.8301	Median : 2.995	Median :2.192e+10
	Mean :1051.3	Mean : 97.07	Mean : 21.09	Mean :1.8893	Mean : 3.906	Mean :3.522e+10
	3rd Qu.:1170.2	3rd Qu.:115.03	3rd Qu.: 24.87	3rd Qu.:2.7802	3rd Qu.: 5.255	3rd Qu.:3.901e+10
	Max. :1536.0	Max. :601.00	Max. : 79.93	Max. :6.7844	Max. :23.710	Max. :3.866e+11

EBITDA	Price/Sales	Price/Book
Min. :-206000000	Min. : 0.1532	Min. : 0.000
1st Qu.: 802535500	1st Qu.: 1.6221	1st Qu.: 2.098
Median :1628500000	Median : 2.8332	Median : 3.515
Mean :2289676962	Mean : 3.8399	Mean : 5.637
3rd Qu.:2951304500	3rd Qu.: 4.6297	3rd Qu.: 5.920
Max. :9281000000	Max. :16.8134	Max. :73.840

Fuente: Datos obtenidos a través del análisis realizado con R Studio

3.2. Análisis de Componentes Principales

El método del Componentes Principales (PCA por sus siglas en inglés) es una técnica de reducción de dimensionalidad que se utiliza para transformar un conjunto de variables con alta probabilidad de correlación en un menor conjunto de variables no correlacionadas llamadas componentes principales. Estos componentes principales son una combinación lineal de las variables originales y están ordenados de tal manera que el primer componente principal captura la mayor varianza explicada posible en los datos, el segundo componente principal captura la mayor varianza posible en los datos que quedan tras extraer el primer componente, y así sucesivamente. Por este motivo, la mayoría de la varianza explicada suele estar entre los dos primeros.

Para qué sirve un análisis PCA

1. Para reducir la dimensionalidad: PCA reduce la cantidad de variables en el conjunto de datos, lo que simplifica los modelos sin perder una cantidad significativa de información.
2. Para descorrelacionar variables: Las nuevas variables (componentes principales) no están correlaciones, lo que puede mejorar el rendimiento de los algoritmos de aprendizaje automático que asumen independencia entre variables.
3. Para captura la máxima varianza posible: PCA maximiza la varianza capturada en los primeros componentes principales, lo que permite tener la mayor parte de la información en menos variables.

Por qué se utiliza PCA en este trabajo

Resulta muy útil utilizar el Análisis de Componentes Principales (PCA) para poder reducir las 10 variables iniciales de la base de datos a unos pocos componentes principales, que también capturen la mayoría de la información relevante, por los siguientes motivos:

- Al reducir el número de variables, los resultados del análisis también son más fáciles de interpretar y visualizar, especialmente si los componentes principales capturan patrones significativos en los datos.
- También, porque, aunque se haya tratado de eliminar variables para evitar un exceso de correlación, algunas de las que se han seleccionado relacionadas con

los ámbitos financieros de la empresa pueden estar altamente correlacionadas. PCA va a terminar de eliminar esta correlación restante para mejorar el rendimiento de algoritmos de *clustering* que asumen independencia entre variables.

- Además, PCA puede ayudar a identificar patrones subyacentes en los datos, como agrupamientos naturales de empresas con perfiles similares en términos de sostenibilidad y desempeño financiero y a detectar posibles relaciones entre las diferentes variables y las empresas que resultarán muy útiles para resolver el problema planteado.

Cuando esta técnica sea aplicada, en el capítulo 4 se podrán ver los resultados obtenidos y las salidas proporcionadas por la plataforma R Studio.

3.3. Análisis de los clústeres

El siguiente paso es la utilización de distintos modelos para crear y analizar clústeres con el fin de resolver el problema planteado y cumplir con los objetivos propuestos. Los métodos elegidos son:

- El *Clustering* Jerárquico, ya permite observar cómo los clústeres se dividen y fusionan a diferentes niveles de similitud, proporcionando una comprensión más profunda de las relaciones dentro del conjunto de datos.
- El K-Means, pues es un método sencillo y fácil de implementar que puede manejar grandes volúmenes de datos.

Ambas técnicas van a agrupan a las empresas que tengan características tanto financieras como de sostenibilidad similares, de tal manera, que se facilite la identificación de patrones en los datos que relacionen ambos aspectos y que sean difíciles de encontrar a simple vista.

El *Clustering* Jerárquico, por un lado, va a organizar a las empresas en una estructura jerárquica, creando clústeres que se van uniendo. Además, en este método se va a construir un dendrograma, que es un árbol donde cada nodo representa un clúster. Al principio de su construcción, cada empresa empieza en su propio clúster, y a partir de ahí se van a ir fusionando los más similares en pasos sucesivos. Una ventaja importante de

este método es que no requiere especificar el número de clústeres de antemano y ofrece una visualización clara de cómo se forman y se relacionan.

Por otro lado, el método K-Means va a dividir el conjunto de datos en un número específico de clústeres (k). Cada empresa se asignará al clúster con el centroide más cercano, que es el punto medio o el centro de un clúster. Luego, se recalcularán los centroides basándose en las empresas asignadas y se repitirá este proceso hasta que estos consiguan estabilizarse. K-Means es rápido y funciona bien con grandes cantidades de datos, pero su resultado puede variar dependiendo de cómo se elijan los centroides iniciales.

En ambos casos, para crear la agrupación, se va a buscar la solución óptima, es decir, aquella qué minimice la varianza que hay entre los clústeres.

Para complementar el análisis y comprobar la calidad de los clústeres formados por un algoritmo como K-Means para distintos valores de k, se va a implementar la técnica *Average Silhouette*. La *silhouette* de una observación básicamente mide qué tan similar es a su propio clúster en comparación con otros clústeres, y su valor para cada punto puede variar entre -1 y 1.

- *Silhouette* cercana a 1: La observación está bien agrupada y es similar a otras en su clúster.
- *Silhouette* cercana a 0: La observación está en el límite entre dos clústeres.
- *Silhouette* negativa: La observación podría estar mal asignada a su clúster y es más similar a un clúster diferente.

El *Average Silhouette*, por tanto, no va a ser más que el promedio de las *silhouette values* de todas las observaciones en el conjunto de datos. Un promedio cercano a 1 indicará una buena asignación de clústeres, mientras que un valor bajo o negativo sugerirá que los estos pueden no ser adecuados.

4. Resultados del modelo

4.1. Interpretación de los resultados

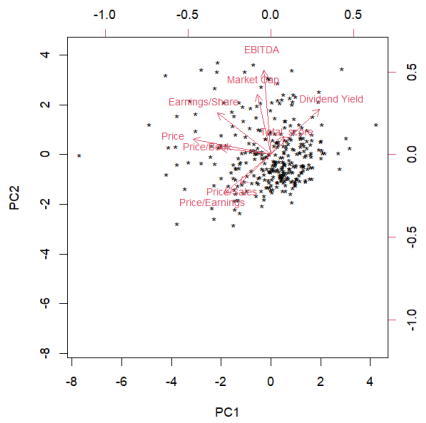
En este punto del trabajo, se va a aplicar al conjunto de datos toda la teoría explicada en la sección anterior.

Aplicación del método de análisis de componentes principales (PCA)

Se aplica el método PCA a la base de datos a través de la plataforma de R Studio como análisis exploratorio para ver si se puede reducir su dimensión. El *output* que se obtiene de dicho análisis no es el esperado, pues para conseguir una varianza explicada significativa de los datos (84%) se necesitan hasta 6 componentes principales. Esto no se ha considerado una reducción relevante del número de variables del *dataset* original y, por tanto, no se va a utilizar en el resto del trabajo más que para las representaciones gráficas en dos dimensiones con los dos primeros componentes principales.

Se representan los dichos dos primeros componentes principales (PC1 y PC2), que entre ellos explican el 39% de la varianza explicada de la base de datos, en un *plot* (Ver Figura 3). Cada punto del gráfico representa a una empresa, y su posición en el espacio la determinan dichas componentes. Las flechas rojas, por otro lado, hacen referencia a las distintas variables y muestran cuál es su contribución a PC1 y PC2. También se extrae una tabla que contiene los valores de dichas componentes principales para interpretarlas a continuación (Ver Figura 3).

Figura 3: Método PCA



	PC1	PC2
Total_score	0.09920007	0.13585237
Price	-0.59052831	0.11476978
Price/Earnings	-0.35477794	-0.29370315
Dividend Yield	0.36932170	0.34266051
Earnings/Share	-0.40707309	0.31492896
Market Cap	-0.10290121	0.45490103
EBITDA	-0.05231001	0.63791643
Price/Sales	-0.23122601	-0.22158355
Price/Book	-0.38297622	0.05042247

Fuente: Datos obtenidos a través del análisis realizado con R Studio

Interpretación de los resultados obtenidos

De este análisis exploratorio se puede observar, en primer lugar, que la variable "Total Score", la cual refleja la métrica sobre la sostenibilidad y responsabilidad empresarial, tiene las siguientes cargas en los componentes principales:

- PC1: 0,0992
- PC2: 0,1359

Estas cargas sugieren que el "Total Score" contribuye de manera moderada tanto a PC1 como a PC2. Aunque no es la variable dominante en ninguno de los componentes, su presencia en ambos indica que los factores de sostenibilidad tienen una influencia perceptible junto con variables financieras más tradicionales.

Por otro lado, se puede apreciar que el componente principal PC1 da un peso similar a variables como "Price", "Earnings/Share", "Price/Book", "Price/Earnings" y "Price/Sales", y menos peso a otras variables como “Market Cap” o “EBITDA”. Esto sugiere que PC1 está capturando principalmente la información sobre la valoración financiera de la empresa.

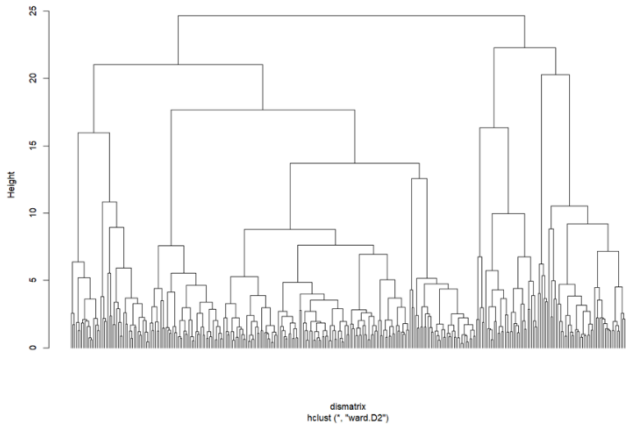
PC2, en cambio, tiene altas cargas en variables como "EBITDA", "Market Cap", "Dividend Yield" y "Earnings/Share", lo cual indica que esta componente mide la rentabilidad y el tamaño de la empresa.

Aplicación del método de *Clustering* Jerárquico

Como se ha mencionado anteriormente, para emplear el método jerárquico de *clustering*, no es necesario predefinir un número de clústeres a formar, sino que, en su lugar, se construye un dendrograma que es una estructura jerárquica basada en una matriz de distancias. Dicha matriz de distancias se calcula con la distancia euclidiana, dado que en este modelo todas las variables empleadas son numéricas y han sido estandarizadas previamente.

Por otro lado, es necesario determinar la proximidad entre los distintos clústeres, lo que se conoce como criterio de enlace, y en este caso, se ha optado por el método de Ward, ya que minimiza la pérdida de información al combinar clústeres. El dendrograma resultante se muestra a continuación (Ver Figura 4).

Figura 4: Dendrograma de clústeres



Fuente: Datos obtenidos a través del análisis realizado con R Studio

Después de obtener el dendrograma, hay que determinar dónde realizar el corte para obtener el número óptimo de clústeres. Al observar el dendrograma, parece que cortar en 2 o en 3 podrían ser las mejores opciones, pues a partir de esas divisiones, los clústeres comienzan a juntarse a alturas más altas, lo que indica una mayor distancia entre ellos, y, por tanto, que los datos de los clústeres que se formarían van a ser más similares entre sí, y más distintos a los de los otros.

- Opción 1: Se corta el dendrograma para obtener 3 clústeres

Se corta el dendrograma para que queden 3 clústeres y se añade una columna a los datos con el clúster de cada registro. Se obtiene que las observaciones de cada clúster son: 231, 50 y 35 respectivamente. Se hace un resumen estadístico para analizar su media y después se añade una columna con la proporción de observaciones que hay dentro de cada clúster (Ver Tabla 2).

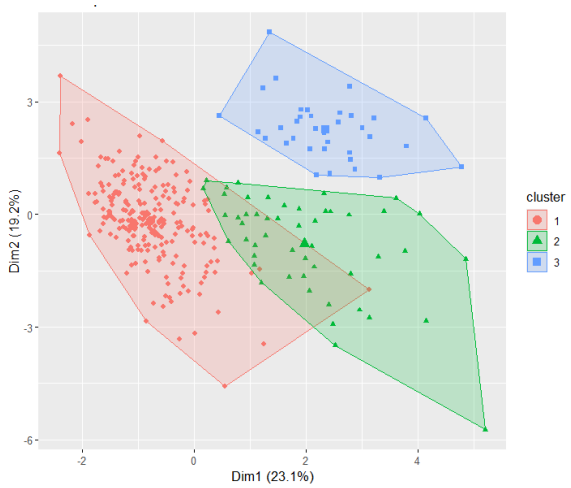
Tabla 2: Perfil de los clústeres (k=3) con *Clustering* Jerárquico

Metric	Clúster 1	Clúster 2	Clúster 3
Total Score	1.038	1.104	1.066
Price	76,2	202	84,0
Price/Earnings	21,7	20,6	17,5
Dividend Yield	1,88	1,22	2,88
Earnings/Share	2,69	9,52	3,93
Market Cap	2.470.885.387	3.843.658.252	10.353.254.455
EBITDA	1.705.368.100	2.451.641.480	5.914.994.714
Price/Sales	4,05	3,42	3,07
Price/Book	4,77	10,3	4,61
Proportion	0,73	0,16	0,11

Fuente: Datos obtenidos a través del análisis realizado con R Studio

A continuación, se utiliza PCA para visualizar las observaciones en dos dimensiones con el paquete Factor Extra y la función “fviz_cluster” (Ver Figura 5).

Figura 5: Clúster *plot* (k=3)



Fuente: Datos obtenidos a través del análisis realizado con R Studio

La Figura 5 muestra el análisis de clústeres realizado proyectado en dos dimensiones. Cada punto en el gráfico representa una observación, y los puntos están coloreados y marcados según el clúster al que pertenecen.

Los Clúster 1 (rojo) y 3 (azul) están claramente separados, indicando que las observaciones dentro de cada uno son similares entre sí y distintas de las de otros. El Clúster 2 (verde), por el contrario, muestra cierta superposición con el Clúster 1 (rojo), lo que sugiere que algunas observaciones en estos clústeres no están claramente diferenciadas.

Por otra parte, el Clúster 1 (rojo) tiene una alta densidad de puntos, lo que indica que es el clúster más grande en términos de número de observaciones. El Clúster 3 (azul) tiene una menor densidad, mostrando que es un clúster más pequeño.

Las dos dimensiones, Dim1 y Dim2, representan los componentes principales que explican la mayor variabilidad en los datos, con Dim1 explicando el 23,1% de la variabilidad y Dim2 explicando el 19,2%. Esto indica que estas dos dimensiones capturan una cantidad significativa de la información en los datos originales.

- Opción 2: Se corta el dendrograma para obtener 2 clústeres

Se corta el dendrograma para que queden 2 clústeres y de la misma forma que antes se añade una columna a los datos con el clúster de cada registro. Se obtiene que las observaciones de cada clúster son: 231 y 85 respectivamente. También se hace un resumen estadístico para analizar su media y después, se añade una columna con la proporción de observaciones que hay dentro de cada clúster (Ver Tabla 3) .

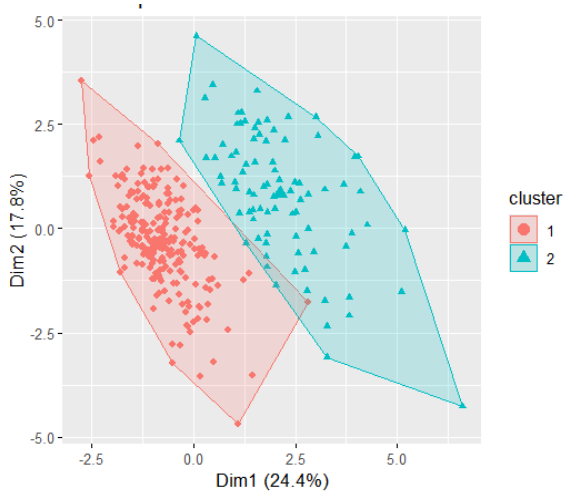
Tabla 3: Perfil de los clústeres (k=2) con *Clustering* Jerárquico

Metric	Clúster 1	Clúster 2
Total Score	1.038	1.088
Price	76,2	154
Price/Earnings	21,7	19,3
Dividend Yield	1,88	1,91
Earnings/Share	2,69	7,22
Market Cap	2.417.088.387	6.524.815.512
EBITDA	1.705.368.100	3.877.622.224
Price/Sales	4,05	3,27
Price/Book	4,77	7,98
Proportion	0,73	0,27

Fuente: Datos obtenidos a través del análisis realizado con R Studio

A continuación, siguiendo también el mismo proceso que en la opción anterior, se utiliza PCA para visualizar las observaciones en dos dimensiones con el paquete Factor Extra y la función "fviz_cluster" (Ver Figura 6).

Figura 6: Clúster *plot* (k=2)



Fuente: Datos obtenidos a través del análisis realizado con R Studio

Los clústeres 1 (rojo) y 2 (azul) están claramente separados, indicando que las observaciones dentro de cada clúster son más similares entre sí y distintas de las observaciones en el otro. Aunque hay una ligera superposición entre ambos clústeres, la separación general es bastante clara, lo que sugiere una buena diferenciación entre los dos grupos.

Por otro lado, ambos clústeres tienen una densidad relativamente alta de puntos, sugiriendo que los dos son significativos en términos de número de observaciones. Además, la distribución de los puntos en ambos es compacta, lo cual es un indicativo positivo de la cohesión interna de cada clúster.

Igual que en la opción anterior, las dos dimensiones, Dim1 y Dim2, representan los componentes principales que explican la mayor variabilidad en los datos, con Dim1 explicando el 24,4% de la variabilidad y Dim2 explicando el 17,8%, siendo esta una cantidad significativa de la información en los datos originales.

Se concluye de las observaciones realizadas sobre los gráficos de ambas opciones que tiene más sentido utilizar únicamente 2 clústeres, es decir cortar el dendograma en k=2, ya que la visualización de los clústeres muestra una separación entre los grupos más clara, demostrando que las observaciones dentro de cada clúster son más similares entre sí y más diferentes a las de los otros. Además, al observar cómo de compactos están los puntos dentro de cada clúster, vemos también que los clústeres de k=2 tienen una cohesión interna más alta. Esto significa que las observaciones dentro de cada uno son bastante similares, lo cual es un indicativo positivo de un buen *clustering*. Por tanto, se va a llevar a cabo la interpretación de los resultados de únicamente la opción 2 (k=2).

Interpretación de los resultados obtenidos en la opción 2

De la Tabla 3 (ver página 19), se pueden obtener las siguientes conclusiones sobre los clústeres formados.

- El Clúster 1 presenta una puntuación total media de sostenibilidad de 1.038, la cual es inferior al del Clúster 2, que tiene un promedio de 1.088, lo que significa que el Clúster 2 podría estar compuesto por empresas que tienen prácticas de sostenibilidad ligeramente más robustas y mejor implementadas. En términos de valor de mercado, el Clúster 1 tiene un “Market Cap” mucho más elevado

comparado con el Clúster 2, lo que indica que el primer grupo incluye empresas de mayor tamaño.

- Al analizar otras métricas financieras, se observa que el Grupo 1 tiene un rendimiento en términos de "Earnings/Share" y "Price/Earnings" (2,69 y 21,7 respectivamente) considerablemente más bajo que el Clúster 2 (7,22 y 154 respectivamente). Esto puede sugerir que, aunque el Clúster 1 incluye empresas más grandes y establecidas, el Clúster 2 está compuesto por empresas con una mayor rentabilidad por acción, lo que es muy atractivo para los inversores interesados en rendimientos a corto plazo.
- La variable "EBITDA" sigue un patrón similar, en el Grupo 1 muestra un valor significativamente mayor que el Clúster 2, lo cual está en consonancia con la mayor capitalización de mercado del Clúster 1. Sin embargo, las métricas de "Price/Sales" y "Price/Book" son más altas en el Clúster 2, lo que sugiere que el mercado podría estar valorando más positivamente cada dólar de ventas o cada dólar de activos en empresas del Clúster 2, lo que podría ser debido a percepciones de un mayor potencial de crecimiento, una gestión más eficiente o prácticas de sostenibilidad más robustas.

Este análisis muestra, por tanto, una posible correlación entre un alto compromiso con la sostenibilidad y un mejor rendimiento financiero, particularmente observado en las empresas del Clúster 2, como por ejemplo Costco o Broadcom. Estas compañías, que presentan puntuaciones más altas de sostenibilidad, también exhiben indicadores de su *performance* financiera superiores. Sin embargo, las empresas del Clúster 1, como CME Group o Marriott International al ser más grandes y estables, podrían disponer de más recursos para invertir en prácticas sostenibles. Además, dichas empresas tienden a mostrar rendimientos más consistentes y menos volátiles, lo que podría explicar sus ratios de Price/Earnings y Earnings/Share relativamente más bajos en comparación con las empresas del Clúster 2, que, siendo más pequeñas, muestran mayor volatilidad en su rendimiento financiero.

Aplicación del método K-Means

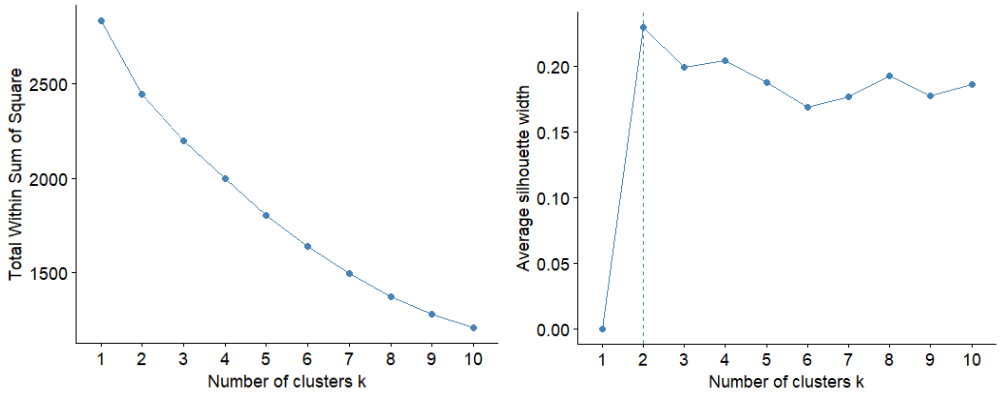
Es necesario especificar el número de clústeres a formar con este tipo de algoritmo para que las observaciones se asignen a cada uno de ellos. El objetivo es determinar el número óptimo de clústeres que minimice la dispersión dentro de cada uno, buscando formar los

grupos más heterogéneos posibles. Para determinar dicho número en el conjunto de datos, se pueden utilizan distintos métodos de evaluación diferentes como, por ejemplo, el método del codo, que ayuda a identificar el valor a partir del cual disminuye significativamente la mejora en la compactación de los clústeres, siendo ese el punto óptimo de k (número de clústeres) que hay que elegir, y el Método *Silhouette*, que mide la cohesión y la separación de los clústeres, eligiendo aquel valor de k que haga los clústeres más homogéneos entre ellos y más separados entre sí.

El primero que se va a ejecutar con R Studio es el método del codo, el cual implica calcular la suma de las distancias al cuadrado entre los clústeres para diferentes valores de k y encontrar el punto en el que la tasa de cambio de esta suma disminuya significativamente. En este caso, el gráfico extraído sugiere que el número óptimo de clústeres está entre 2 y 3 (Ver Figura 7), pues es dónde la pendiente deja de ser tan ascendente, y pasa a ser más plana, significando que la tasa de cambio ya no disminuye de una manera considerable.

El segundo método ejecutado es de *Silhouette*. Con este método se va a buscar el valor de k que maximice el coeficiente promedio de la silueta, lo que indicará que los clústeres son más homogéneos y están bien separados entre sí. En este análisis se observa que el coeficiente de la silueta es más alto para k = 2 (Ver Figura 7).

Figura 7: Método del codo y método *Silhouette*



Fuente: Datos obtenidos a través del análisis realizado con R Studio

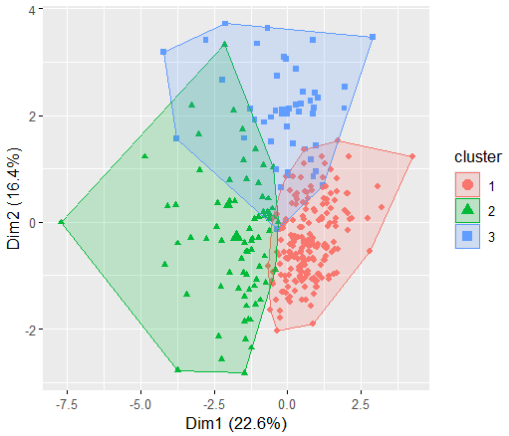
A continuación, se va a realizar el análisis con K-Means para 2 y 3 clústeres (k=2 y k=3 respetivamente) y posteriormente se analizarán únicamente los resultados de aquel que funcione mejor, aunque por el análisis que se acaba de realizar para ver el número óptimo de clústeres, ya se prevé que va a ser el que agrupa únicamente en 2.

- K-Means con 3 clústeres *(k=3)*

Se realizan 50 asignaciones aleatorias para clasificar a las observaciones en los 3 clústeres. Como se ha mencionado anteriormente, con este enfoque se busca encontrar mínimos locales y obtener así agrupaciones más consistentes. Si se examina el tamaño de los clústeres, se puede observar que el Clúster 1 contiene la mayor cantidad de observaciones, con un total de 194.

Además, se emplean las dos primeras variables resultantes del Análisis de Componentes Principales para examinar la estructura de los clústeres en un espacio bidimensional (Ver Figura 8).

Figura 8: Clúster *plot* de 3 clústeres



Fuente: Datos obtenidos a través del análisis realizado con R Studio

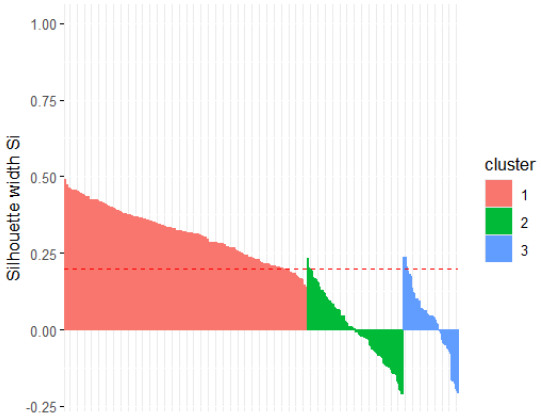
De la Figura 8, se deduce que los clústeres están relativamente bien separados, pero que existe una cierta superposición, especialmente entre los clústeres 1 (rojo) y 3 (azul) y entre los clústeres 3 (azul) y 2 (verde) que indica que algunas observaciones no están tan claramente diferenciadas entre estos clústeres, y por tanto que pueden estar mal asignadas.

Para valorar el problema que crea la superposición, se hace una evaluación más profunda utilizando el método *Average Silhouette*. El *Average Silhouette*, como se ha explicado en el capítulo anterior, evalúa la calidad de los clústeres, lo que permite verificar si las observaciones están adecuadamente asignadas en cada uno (Ver Figura 9).

Se identifica que los clústeres 2 y 3 tienen un conjunto de observaciones con siluetas negativas, indicando una asignación incorrecta. Además, la silueta media global calculada

para todos los clústeres es de 0,20, lo que sugiere una calidad de agrupamiento moderada-baja.

Figura 9: *Silhouette plot* de 3 clústeres

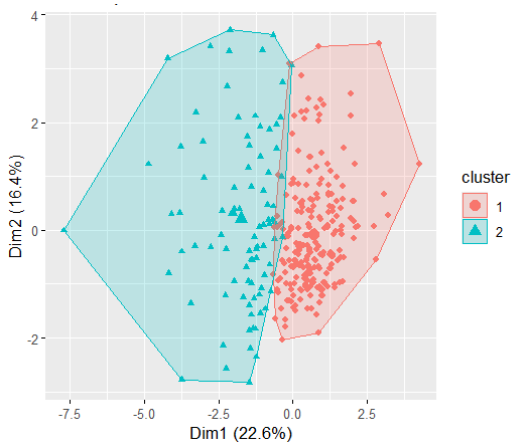


Fuente: Datos obtenidos a través del análisis realizado con R Studio

- K-Means con 2 clústeres *(k=2)*

A continuación, se realiza el mismo proceso, pero asignando las observaciones a 2 clústeres. El *output* de R Studio que se obtiene es el siguiente (Ver Figura 10).

Figura 10: Clúster *plot* de 2 clústeres

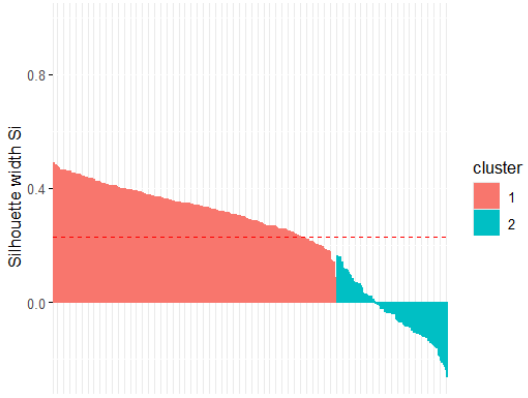


Fuente: Datos obtenidos a través del análisis realizado con R Studio

Ambos clústeres tienen una densidad bastante alta de puntos, lo que indica que cada uno tiene un número significativo de observaciones. Además, la distribución de los puntos dentro de cada clúster es razonablemente compacta, sobre todo en el Clúster 1 (rojo), indicando que las observaciones dentro de cada uno son bastante similares entre sí.

En comparación con el gráfico k=3, este gráfico muestra una mejor separación y una menor superposición, lo que sugiere que k=2 es una mejor opción para representar la estructura de los datos. Se realiza el análisis del *Average Silhouette* para asegurar de que las conclusiones son ciertas (Ver Figura 11).

Figura 11: *Silhouette plot* de 2 clústeres



Fuente: Datos obtenidos a través del análisis realizado con R Studio

Se observa que el Clúster 2 tiene un gran número de observaciones con siluetas negativas, indicando que su asignación ha sido incorrecta. Sin embargo, aun así, la silueta media global para ambos clústeres formados es de 0,23, lo que sugiere una calidad de agrupamiento ligeramente mayor que en el caso anterior.

Por otro lado, se saca el perfil de los clústeres de tal manera que se pueda ver el valor de las variables en cada uno para que se analice a continuación (Ver Tabla 4).

Tabla 4: Perfil de los clústeres (k=2) con K-Means

Metric	Clúster 1	Clúster 2
Total Score	195	213
Price	30,1	86,5
Price/Earnings	9,08	12,9
Dividend Yield	1,32	1,10
Earnings/Share	2,61	4,85
Market Cap	3.569.632.173	5.156.422.856
EBITDA	2.005.288.577	2.530.002.335
Price/Sales	2,79	4,03
Price/Book	3,13	13,2

Fuente: Datos obtenidos a través del análisis realizado con R Studio

Se van a analizan únicamente los subgrupos obtenidos con k=2, pues los resultados son ligeramente mejores que para k=3, al haber confirmado que sus clústeres tienen una mayor calidad al ser su silueta media global ligeramente superior.

Interpretación de los resultados obtenidos en la opción 2

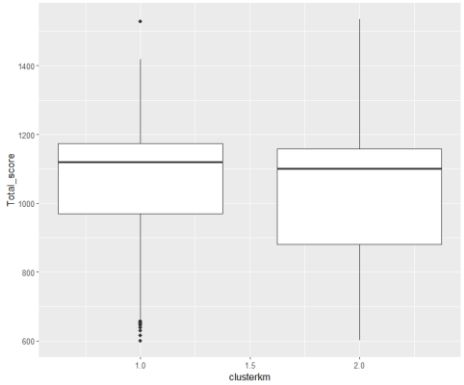
De la Tabla 4 (ver página 25), se pueden obtener las siguientes conclusiones sobre los clústeres formados.

- El Clúster 1 muestra una puntuación total media de sostenibilidad de 195, mientras que el Clúster 2 tiene una ligeramente superior de 213. Esta diferencia indica que el Clúster 2 está más comprometido con la sostenibilidad. En términos de rendimiento financiero, el Clúster 1 tiene un ratio Price/Earnings considerablemente menor (30,1) comparado con el del Clúster 2 (86,5), lo que indica una valoración más alta en el mercado para el segundo grupo, posiblemente debido a expectativas futuras de crecimiento o a una percepción de menor riesgo asociado a sus prácticas sostenibles.
- En lo que respecta a "Earnings/Share", el Clúster 2 muestra un valor más alto (4,85) en comparación con el Clúster 1 (1,32), reforzando la idea de que las empresas con mejores prácticas de sostenibilidad generan mayores beneficios por acción, lo cual es un indicador clave de rentabilidad. Sin embargo, es importante resaltar que el Grupo 2 también tiene una mayor capitalización de mercado, indicada por un "Market Cap" de aproximadamente 5,15 billones de dólares, frente a los 3,56 billones de dólares del Grupo 1, lo que sugiere que las empresas de este grupo no solo son más grandes, sino que también son posiblemente más influyentes en sus respectivos sectores.
- El análisis de las ratios Price/Sales y Price/Book revela que el Grupo 2 tiene valores ligeramente más altos (4,03 y 13,2 respectivamente) que el Grupo 1 (2,79 y 3,13 respectivamente), lo que podría interpretarse como un mercado dispuesto a pagar más por cada dólar de ventas y por cada dólar de valor en libros, en empresas que demuestran mayor sostenibilidad.

Este análisis proporciona la evidencia de que las empresas con un mayor compromiso hacia la sostenibilidad no solo son valoradas más favorablemente por el mercado, sino que también están logrando un desempeño financiero superior. Esta tendencia también sugiere que las prácticas sostenibles podrían estar contribuyendo al atractivo a largo plazo de estas empresas en el mercado.

Por último, se presenta un resumen estadístico por clúster, para comprobar el valor de la variable que es más importante para el análisis: "Total_score", y se representa en un gráfico *boxplot* (Ver Figura 12). Esta figura proporciona una visión clara de la distribución de dicha variable dentro de cada clúster, mostrando la mediana, los cuartiles y los valores atípicos, permitiendo entender la variabilidad de la puntuación ESG dentro de cada grupo.

Figura 12: *Boxplot*



Fuente: Datos obtenidos a través del análisis realizado con R Studio

La Figura 12 también proporciona *insights* relevantes para inversores y analistas interesados en la sostenibilidad corporativa, ya que las características de distribución de estas cajas varían notablemente, aunque la mediana de las puntuaciones de sostenibilidad sea similar en ambos casos.

En el Clúster 1, se observa una amplia variabilidad en las puntuaciones de sostenibilidad. La extensión de la caja, que va desde aproximadamente 900 hasta casi 1.200, indica que este grupo incluye empresas con un amplio rango de compromisos hacia prácticas sostenibles. Además, la presencia de varios valores atípicos, tanto altos como bajos, sugiere que algunas empresas están significativamente más avanzadas o rezagadas en sus prácticas de sostenibilidad en comparación con la mayoría del grupo. Esta variabilidad puede señalar oportunidades de inversión para aquellos que buscan compañías que podrían estar subvaloradas o que muestran potencial de mejora en sus prácticas de sostenibilidad.

Por otro lado, el Clúster 2 muestra una mayor consistencia en las puntuaciones de sostenibilidad, con menos valores extremos. Esto indica que las empresas en este clúster son generalmente homogéneas en cuanto a su enfoque de sostenibilidad, manteniendo un nivel de práctica más uniforme. Para inversores que prefieren estabilidad y menor riesgo

en sus selecciones de inversión sostenible, el Clúster 2 podría representar una opción más atractiva.

Aunque las medianas de los dos clústeres son comparables, la diferencia en la dispersión y consistencia de las puntuaciones puede tener implicaciones significativas. El Clúster 1, con su mayor variabilidad, ofrece una diversidad que puede ser atractiva para inversores que buscan potencial de crecimiento o están interesados en apoyar empresas que se encuentran en el camino hacia mejores prácticas sostenibles. En contraste, el Clúster 2 podría apelar a aquellos que valoran la consistencia y la previsibilidad en las prácticas de sostenibilidad corporativa, facilitando una inversión más segura y estable.

Este enfoque subraya la importancia de considerar no solo el nivel promedio de sostenibilidad, sino también la variabilidad y consistencia de estas prácticas dentro de los clústeres. Esta perspectiva es crucial para tomar decisiones informadas en el ámbito de la inversión sostenible, donde la consistencia en la adopción de este tipo prácticas puede ser tan importante como el nivel de compromiso.

5. Respuesta al problema

5.1. Cómo se podría aplicar el modelo en la toma de decisiones de inversión

Implementar un modelo de *clustering* en la toma de decisiones de inversión ofrece múltiples ventajas y puede ser particularmente valioso en el ámbito de la inversión sostenible. Este enfoque permite a los inversores y analistas segmentar el mercado de manera eficiente, identificando grupos de empresas con características financieras y de sostenibilidad similares. Al hacerlo, facilita la identificación rápida de oportunidades de inversión que no solo cumplen con criterios financieros, sino también con criterios de responsabilidad social y ambiental.

1. Segmentación del mercado e identificación de oportunidades

Al aplicar técnicas de *clustering*, se forman grupos de empresas que exhiben patrones similares en términos de rendimiento y prácticas de sostenibilidad. Esta segmentación del mercado es crucial para los inversores que buscan empresas que no solo sean rentables, sino que también cumplan con altos estándares de sostenibilidad. Por ejemplo, un clúster puede incluir empresas que, además de tener sólidos fundamentos financieros, muestran un compromiso excepcional con la sostenibilidad ambiental. Identificar estos grupos permite a los inversores dirigir sus recursos hacia compañías que están alineadas con sus valores sin comprometer el rendimiento financiero.

2. Evaluación de riesgo y estrategias de diversificación

El conocimiento profundo que proporciona el *clustering* sobre las características de diferentes grupos de empresas también facilita la evaluación de riesgos y la diversificación de la cartera. Al entender las dinámicas específicas de cada clúster, los inversores pueden planificar estrategias de inversión que equilibren adecuadamente el riesgo y el rendimiento. Por ejemplo, invertir en empresas de varios clústeres puede ayudar a mitigar los riesgos asociados con sectores o prácticas específicas, asegurando una cartera más estable y menos vulnerable a las fluctuaciones del mercado único.

3. *Benchmarking* y análisis competitivo

Además, el análisis de clústeres permite realizar comparaciones efectivas dentro de grupos homogéneos. Los inversores pueden identificar no solo a las empresas líderes que superan a sus similares en términos de rendimiento financiero y prácticas de sostenibilidad, sino también a aquellas que pueden estar más rezagadas. Esta información permite a los inversores apoyar a aquellas empresas que realmente están marcando la diferencia en sus industrias.

4. Seguimiento y reevaluación continua

Una vez que las inversiones están en marcha, los clústeres también facilitan el seguimiento continuo del rendimiento y las prácticas de las empresas. Los cambios significativos en la clasificación de una empresa dentro de su clúster pueden ser indicativos de una modificación en sus fundamentos, lo que puede requerir una reevaluación de la inversión. Esta capacidad para responder rápidamente a los cambios asegura que las estrategias de inversión permanezcan relevantes y efectivas.

5. Comunicación e información

Para gestores de fondos y asesores financieros, los resultados del *clustering* proveen una base sólida para la comunicación de estrategias de inversión a los clientes. El enfoque basado en datos demuestra un método riguroso y metódico para la selección de inversiones, lo cual es especialmente importante en una era donde los inversores están cada vez más interesados en cómo sus inversiones impactan el mundo.

6. Alineación con regulaciones y estándares globales

Finalmente, el uso de modelos que integran consideraciones de sostenibilidad es crucial para cumplir con las regulaciones globales emergentes sobre divulgación y prácticas sostenibles. Este enfoque no solo satisface los requisitos legales y de conformidad, sino que también promueve la responsabilidad corporativa y el desarrollo sostenible.

El análisis realizado con el método K-Means con k=2 es un ejemplo concreto de cómo esta técnica puede guiar las decisiones de inversión. En este caso, el Grupo 2, que ha demostrado tener mayores puntuaciones de sostenibilidad y mejores fundamentos financieros, representa una opción de inversión atractiva. Los inversores que buscan

alinear sus carteras con sus valores éticos sin comprometer el rendimiento financiero encontrarían en el Grupo 2 a los candidatos idóneos.

6. Conclusiones

6.1. Conclusiones clave del estudio

La investigación llevada a cabo destaca la creciente importancia de incorporar criterios de sostenibilidad en las decisiones de inversión. El desarrollo de un modelo analítico basado en algoritmos de aprendizaje automático ha permitido identificar y evaluar empresas según su compromiso con prácticas sostenibles, promoviendo así la inversión responsable. Este modelo no solo clasifica a las empresas por su rendimiento financiero, sino que también pondera significativamente su compromiso con los criterios ambientales, sociales y de gobernanza.

Las conclusiones clave obtenidas del estudio son la siguientes:

1. Correlación entre sostenibilidad y rendimiento financiero

Las empresas que han demostrado un alto compromiso con la sostenibilidad también han presentado un mejor desempeño financiero. Particularmente en el Grupo 2, identificado a través del método K-Means con k=2, las empresas con puntuaciones más altas de sostenibilidad también mostraban mejores indicadores financieros como en los ratios Earnings/Share y Price/Earnings. Esto confirma que las prácticas sostenibles no solo benefician al medio ambiente y a la sociedad, sino que también pueden mejorar la rentabilidad y la valoración de mercado de una empresa.

2. Mayor estabilidad y menor riesgo

Las empresas dentro de los clústeres de alta puntuación de sostenibilidad tienden a ser más estables y se perciben con menos riesgo. Esto es particularmente atractivo para inversores que buscan estabilidad a largo plazo en sus inversiones.

3. Diversificación estratégica

La segmentación del mercado utilizando técnicas de *clustering* permite a los inversores identificar oportunidades de inversión diversificadas que alinean rentabilidad con responsabilidad. El Grupo 2, identificado a través del método *Clustering* Jerárquico con k=2, por ejemplo, lo forman empresas de distintos sectores e industrias, como Costco (Clubes de Almacenes e hipermercados) o Broadcom (Telecomunicaciones), lo que

facilitaría la creación de una cartera robusta y resiliente que esté mejor equipada para manejar las dinámicas cambiantes del mercado global.

4. Fomentar la inversión responsable es crucial

Invertir en empresas que priorizan la sostenibilidad contribuye directamente a un impacto positivo en el medio ambiente y en la sociedad. Estas inversiones apoyan prácticas empresariales que minimizan el daño ecológico, mejoran las condiciones laborales y fomentan una gobernanza corporativa ética y transparente.

Además, a medida que las regulaciones globales se vuelven más estrictas en términos de sostenibilidad y transparencia, las empresas que ya adoptan estas prácticas están mejor posicionadas para cumplir con estos estándares y satisfacen las crecientes expectativas de los consumidores y *stakeholders* que valoran la responsabilidad corporativa.

Por otro lado, las prácticas sostenibles a menudo impulsan la innovación al requerir que las empresas busquen nuevas tecnologías y procesos para reducir su huella ambiental. Esto puede traducirse en una ventaja competitiva significativa, posicionando a las empresas como líderes en sus respectivos campos.

También la inversión responsable atrae a una base más amplia de inversores, especialmente a aquellos interesados en los beneficios a largo plazo y la sostenibilidad global. Estos inversores son cruciales para el crecimiento y la estabilidad a largo plazo de las empresas.

En conclusión, el modelo desarrollado en este Trabajo de Fin de Grado no solo beneficia a los inversores al proporcionarles una herramienta robusta para la toma de decisiones informadas, sino que también promueve prácticas empresariales que pueden sostener y mejorar la sociedad y el entorno. Al integrar consideraciones ESG en la evaluación de empresas, este modelo pone de manifiesto que la sostenibilidad es intrínsecamente vinculante a la rentabilidad y la responsabilidad, redefiniendo lo que significa invertir con éxito en el mundo contemporáneo.

6.2. Limitaciones y futuras áreas de investigación

En el desarrollo de este Trabajo de Fin de Grado, se han identificado varias limitaciones que podrían influir en la generalización y aplicabilidad de los resultados obtenidos. Una

de estas limitaciones es la posible restricción en la variedad y representatividad de los datos. Dependiendo de la fuente utilizada, como por ejemplo Kaggle en este caso, es posible que los datos no abarquen todas las industrias o regiones de manera exhaustiva, limitando así la capacidad de generalizar los resultados obtenidos a nivel global.

Otra limitación importante es la simplificación de variables. En el proceso de modelización, se han seleccionado y simplificado las variables específicas para el análisis, lo que puede haber dejado fuera factores relevantes que influyen en la sostenibilidad y el rendimiento financiero de las empresas. Esta simplificación puede resultar en que el modelo no capture completamente la complejidad de las interacciones entre prácticas sostenibles y rendimiento financiero.

Por último, los criterios de sostenibilidad ambiental, social y de gobernanza están en constante evolución, lo que también representa un desafío. El modelo desarrollado podría quedar desactualizado si no se adaptan continuamente para reflejar nuevas normativas, expectativas sociales y avances tecnológicos que influyen en las prácticas de sostenibilidad.

Por otro lado, posibles futuras áreas de investigación pueden ser las siguientes:

- Expansión del conjunto de datos: Para futuros trabajos, sería bueno ampliar el conjunto de datos para incluir más empresas, sectores e incluso más regiones geográficas para que el modelo sea más robusto y generalizable.
- Integración de datos en tiempo real: Investigar cómo integrar datos en tiempo real para ayudar a mejorar la precisión del modelo y para permitir que los inversores tomen decisiones basadas en la información más actualizada posible.
- Análisis del impacto a largo plazo: Se podrían realizar estudios longitudinales para entender mejor cómo las prácticas de sostenibilidad impactan el rendimiento financiero a largo plazo de las empresas.
- Exploración de nuevas variables y modelos: Investigar el impacto de nuevas variables ESG y experimentar con diferentes modelos de aprendizaje automático para mejorar la precisión y la interpretabilidad de los resultados.
- Desarrollo de herramientas interactivas: Desarrollar herramientas interactivas basadas en el modelo que permitan a los usuarios explorar diferentes escenarios

de inversión podría ser un paso práctico para aplicar los resultados de la investigación.

Estas áreas además de abordar las limitaciones del estudio actual también abren nuevos caminos para investigaciones futuras que pueden enriquecer aún más el campo de la inversión sostenible y responsable.

7. Bibliografía

BBVA. (2022). *¿Qué es la inversión sostenible y cuáles son sus beneficios?* BBVA

Busch, T., Pruessner, E., Oulton, W., Palinska, A., & Garrault, P. (2024). *Methodology for Eurosif Market Studies on Sustainability-related.* University Hamburg; AIR; Eurosif.

Butler, A. (2018). *Do Customers Really Care About Your Environmental Impact?* Forbes

Claver, A. (2023). *How asset managers are working on sustainable investing.* ROBECO.

Deloitte. (2022). *Deloitte 2022 CxO Sustainability Report.* Deloitte

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The Impact of Corporate Sustainability on Organizational Processes and Performance. Management Science, 60(11), 2835-2857.

ESG Enterprise . (s.f.). *ESG Ratings Methodology.*

Haanaes, K. & Olynec, N. (2022). *Why all businesses should embrace sustainability.* IMD

Porter, M. E., & Kramer, M. R. (2006). Strategy & Society: The Link between Competitive Advantage and Corporate Social Responsibility. Harvard Business Review, 84(12), 78-92.

Rafi, T. (2021). C*orporate Strategies Should Be Focused On Sustainability*. Forbes

Repsol. (2023). *¿Qué son los criterios ESG y por qué son importantes? La clave para el inversor socialmente responsable.* Repsol

Santander Asset Management. (s.f.). *¿Qué es la inversión sostenible y qué son los criterios ASG por los que se rige?* Santander

United Nations. (2021). *Principios para inversión sostenible.* PRI.

8. Anexo

Código R Studio

```
rm(list = ls())

install.packages("ggplot2")
install.packages("factoextra")
install.packages("NbClust")
install.packages("dplyr")
install.packages("reshape2")
install.packages("cluster")
install.packages("dendextend")
install.packages("gplots")

library(gplots)
library(ggplot2)
library(factoextra)
library(NbClust)
library(dplyr)
library(reshape2)
library(cluster)
library(dendextend) # para evitar problemas con la paleta de colores


setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
getwd()


################################
#TECNICAS DE CLUSTERING#
################################
datos <- readxl::read_xlsx("DataBase.xlsx")
head(datos)
str(datos)
summary(datos)

#Todas las variables son numéricas por lo que no hace falta convertirlas

#Vemos si hay algun NA y lo eliminamos
sum(is.na(datos)) # Al ejecutarlo vemos que hay 0 NA por lo que no hace falta omitir nada

#Eliminamos los outliers
datos<-datos[datos$`Price/Earnings`<100 & datos$`Price/Earnings`>-100, ]
datos<-datos[datos$`Price/Sales`<20, ]
datos<-datos[datos$`Price/Book`<100, ]
datos <- datos[datos$EBITDA > -411100000 & datos$EBITDA < 9895000000, ]
```

```
#Estandarizamos las variables numéricas
datosnorm<-scale(datos[,2:10])
View(datosnorm)

#Hacemos un previo análisis exploratorio de las variables
summary(datos)#Resumen estadístico de las variables


#Calculamos la matriz de distancias

dismatrix <- dist(datosnorm, method = "euclidean")# Calculamos la matriz de distancias
dismatrix

###ANALISIS PCA###
pcs<-prcomp(datosnorm)
biplot(pcs, scale=0, cex = c(1, 0.8), xlabs=rep("*", nrow(datosnorm))) #Vemos un biplot para ver si podemos observar un número de clusters conveniente
pcs$rotation
var_explicada <- pcs$sdev^2/sum(pcs$sdev^2)
var_explicada_acum <- cumsum(var_explicada)
var_explicada_acum
###CLUSTERING JERARQUICO####
#En primer lugar utilizamos la matriz de distancias previamente calculada. El método utilizado es el de distancia euclídea ya que todas las variables son numéricas y estás estandarizadas

hc <- hclust(dismatrix, method = "ward.D2") #Utilizamos el criterio de enlace ward
plot(hc,hang = -1,labels=FALSE)

#Una vez obtenido el dendrograma vamos a decidir por donde se debe cortar para ver el número de clusters conveniente
#A simple vista, un número óptimo podría ser 2 o 3, ya que a partir de este, los clusters ya se empiezan a juntar a alturas más altas, lo cual significa que estarán más alejados los unos de los otros

heatmap.2(x = datosnorm, scale = "none",
          distfun = function(x){dist(x, method = "euclidean")},
          hclustfun = function(x){hclust(x, method = "ward.D2")},
          density.info = "none",
          trace = "none",
          col = bluered(256),
          cexCol=0.8)
#Este gráfico  muestra tanto un dendograma de las variables, como de las observaciones
#mostrando también la correlación entre las variables y las observaciones

####3 CLUSTERS####
cluster <-  cutree(hc, k = 3) # se corta el dendrograma para que salgan 3 clusters
datosC3<-datos
datosC3$cluster3<-cluster   # y se añade una columna a los datos con el cluster de cada registro
table(cluster)
```

```r
# Se hace un resumen estadístico de las variables por cluster. En concreto se analiza su media
View(datosC3)
resumen1=datosC3[,2:11] %>% group_by(cluster3)%>% summarise_all(mean)

# Se añade una columna con la proporción de observaciones que hay dentro de cada cluster
resumen1$prop<-table(cluster)/ nrow(datos)
resumen1

# Se grafican las observaciones en dos dimensiones haciendo PCA
# para ello se utiliza el paquete facto extra y la funcion fviz_cluster
View(datosC3)
fviz_cluster(object=list(data=datosC3[,2:11], cluster=cluster), geom="point", ellipse=TRUE)

####2 CLUSTERS####
cluster <-  cutree(hc, k = 2)# se corta el dendrograma para que salgan 2 clusters
datosC2<-datos
datosC2$cluster2<-cluster
head(datosC2)
table(cluster)

# Resumen estadístico de las variables por cluster. En concreto se analiza su media
View(datosC2)
resumen2=datosC2[,2:11] %>% group_by(cluster2)%>% summarise_all(mean)

# Se le añade una columna con la proporción de observaciones que hay dentro de cada cluster
resumen2$prop<-table(cluster)/ nrow(datos)
resumen2

# Se grafican las observaciones en dos dimensiones haciendo PCA
# para ello se utiliza el paquete facto extra y la funcion fviz_cluster
fviz_cluster(object=list(data=datosC2[,2:11], cluster=cluster), geom="point", ellipse=TRUE)

#ANÁLISIS DEL MÉTODO

# Se dibuja el dendrograma separando los clusters con rectángulos de colores y se quitan los labels de las hojas por cuestiones estéticas
plot(hc, hang=-1, main="Distancia euclídea, enlace Ward, k=2", labels=FALSE)
rect.hclust(hc, k=2, border=c("red", "blue"))

# Se visualiza el profile plot de los centroides
profile<-reshape2::melt(as.matrix(resumen2[,2:11]))
ggplot(profile, aes(x=Var2, y=value, group=Var1, colour=as.factor(Var1)) )+ geom_line()
```

```r
# Se visualizan los clusters usando PCA
fviz_cluster(km,data=datosnorm, geom="point", show.clust.cent=TRUE)
# Se puede analizar también la silhouette width y ver que observaciones tienen silueta negativa
require("cluster")
sil <- silhouette(km$cluster, dist(datosnorm))
head(sil)
fviz_silhouette(sil)
sil[sil[, "sil_width"] < 0,]

which(sil[, "sil_width"] < 0)

#average silhoutte width :0.2

#ANÁLISIS DEL MÉTODO
#Se ve la representación de los clusters
fviz_cluster(km,data=datosnorm, geom="point", show.clust.cent=TRUE)

#Se hace el profile plot de los centroides
centprofile<-melt(km$centers)
ggplot(centprofile, aes(x=Var2, y=value, group=Var1, colour=as.factor(Var1)) )+ geom_line()

# Se proyectan los datos en el espacio de los componentes principales (PCAs)
datos_pca <- predict(pcs, newdata = datosnorm)

# Se analiza el perfil de las variables dentro de cada cluster
#Se anade una columna al objeto "datos" con el cluster correspondiente a cada observacion  y se calcula su media y desv típica
datoskm<-datos
datoskm$clusterkm<-km$cluster
View(datoskm)

datoskm[,2:11] %>% group_by(clusterkm) %>% summarise_all(sd)

ggplot(datoskm, aes(x=clusterkm, y=Total_score , group=clusterkm)) +
  geom_boxplot()rm(list = ls())
```