UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas /UPV

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas

Tendencias y competitividad en la fabricación de muebles: Un estudio a partir de la huella digital

Trabajo Fin de Grado

Grado en Administración y Dirección de Empresas

AUTOR/A: Gil Buendia, Vicente

Tutor/a: Doménech i de Soria, Josep

CURSO ACADÉMICO: 2023/2024

o

Resumen

La creciente preferencia por la sostenibilidad, el diseño personalizado y el hogar inteligente, junto con los cambios en la estructura familiar hacen que los fabricantes de muebles tengan que innovar para adaptarse constantemente a estas tendencias. Este TFG tiene como objetivo estudiar cómo la adaptación de las empresas se relaciona con la competitividad de las mismas. Para alcanzarlo, se han empleado indicadores derivados de los sitios web corporativos y de información económica y financiera obtenida de SABI (Sistema de Análisis de Balances Ibéricos). El análisis cuantitativo se ha desarrollado mediante el lenguaje de programación R y técnicas de inteligencia de negocios. Este enfoque permite entender mejor cómo la presencia digital y la adaptación a las tendencias de consumo se asocian con la competitividad de los fabricantes de muebles. Gracias al estudio se ha podido observar como la adaptación de ciertas tendencias que parecen ser fundamentales para empresas de otros sectores a la hora de buscar ser más competitivas, como es la adaptación de productos sostenibles, en el caso del sector de la fabricación de muebles no parecen tener un impacto directo en la competitividad de las empresas.

Palabras clave: Muebles, Indicadores digitales, competitividad

Abstract

The growing preference for sustainability, personalized design, and smart homes, along with changes in family structure, require furniture manufacturers to innovate constantly to adapt to these trends. This Bachelor aims to study how the adaptation of companies relates to their competitiveness. To achieve this, indicators derived from corporate websites and economic and financial information obtained from SABI (Iberian Balance Sheet Analysis System) have been used. The quantitative analysis was conducted using the R programming language and business intelligence techniques. This approach allows for a better understanding of how digital presence and adaptation to consumer trends are associated with the competitiveness of furniture manufacturers. The study has shown that the adaptation of certain trends, which appear to be fundamental for companies in other sectors seeking to become more competitive, such as the adoption of sustainable products, does not seem to have a direct impact on the competitiveness of companies in the furniture manufacturing sector.

Key words: Furniture, Digital indicators, Competitiveness.

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

Hoy en día, el sector de la fabricación de muebles atraviesa una época en la que es fundamental para las empresas que lo conforman ser altamente competitivas. La entrada de mercados con los cuales es muy difícil o casi imposible competir en precios obliga a las empresas nacionales de este sector a buscar nuevos enfoques estratégicos.

Un enfoque típico que adoptan las empresas es abandonar la estrategia de precios y enfocarse en la calidad y el diseño para ser competitivas. Por lo tanto, para las empresas de este sector es fundamental mantenerse actualizadas en cuanto a tendencias.

Esto lo consiguen en gran parte a través del uso de herramientas digitales como las redes sociales, o la interacción con los consumidores mediante sitios web, los cuales les permiten conocer en detalle los gustos y prioridades de un número muy elevado de consumidores. Para ello las compañías utilizan diversas herramientas que les permitan recopilar este tipo de información a fin intentar aumentar su competitividad.

De esta manera para tratar de averiguar el camino que la empresa ha decidido tomar pueden usarse los indicadores digitales, como pueden ser las palabras clave, las cuales pueden encontrarse en su página web, y gracias a las cuales es posible conocer que estrategias y decisiones ha tomado la empresa.

Codina (2004) explica que las palabras clave se utilizan para optimizar y posicionar una web, algo que en el mundo de las empresas es de suma importancia, pues esto tendrá un impacto directo en el número de consumidores finales que comprarán o a los que les serán prestados algún servicio.

Es posible conocer la estrategia que seguirá una determinada empresa en función de las palabras más utilizadas en su página web, ya que es de esperar que en el sector del mueble una empresa que apueste por la fabricación a medida como estrategia de ventas incluirá esta información en su página web.

Por lo tanto, con esta información y dado que para este sector es muy importante la diferenciación y la adaptación de nuevas tendencias, será posible medir si realmente todas estas estrategias de calidad, diseño y adaptación de tendencias realmente están vinculadas a una mejora competitiva.

1.1 Objetivos

El objetivo principal de este trabajo será analizar datos tanto económicos como digitales, tratando de obtener o averiguar posibles relaciones entre ellos mediante técnicas de inteligencia de negocios, a fin de conocer cuáles son las estrategias que mejor funcionan para las empresas del sector del mueble. Para ello a lo largo de la investigación se seguirán los siguientes objetivos:

1. Construir y preparar la base de datos con la información obtenida de las empresas seleccionadas, para poder trabajar de manera adecuada.

2. Realizar análisis de relación de variables mediante aplicación de técnicas de Análisis de Componentes Principales y posteriormente agrupación de observaciones mediante técnicas de Clustering.

3. Construcción de modelos de aprendizaje supervisado de regresión y clasificación.

4.Analisis y evaluación de los resultados obtenidos en cada uno de los pasos ejecutados.

Realizado el estudio pueden obtenerse algunas recomendaciones útiles para las empresas del sector como pueden ser:

·La potenciación de la marca propia de cada una de las empresas.

·El enfoque en calidad y diseños actuales.

·La búsqueda de optimización de los recursos en todo momento

1.2·Estructura

Para la realización del trabajo el estudio está dividido en dos partes diferenciadas, siendo la primera una de contextualización, en la que se detallan los diferentes conceptos que pueden ser relevantes para la correcta compresión del estudio. Mientras que la segunda parte consiste en la realización del análisis, mediante la observación y la ejecución de diversas técnicas y pruebas de carácter estadístico.

La primera parte comprende tanto el capítulo 2, en el que se pone en contexto al lector explicando a grandes rasgos los conceptos que se trataran a lo largo del trabajo, como el capítulo 3, en el que se realiza una explicación de las diversas técnicas utilizadas a lo largo del estudio para la realización del análisis.

Por otro lugar la segunda parte comprende los capítulos 4 y 5, en los que se relatan los procesos seguidos durante el estudio, las diferentes pruebas y análisis, así como los resultados que estos mostraron, para la posterior obtención de conclusiones.

2. Marco contextual

2.1. El sector de la fabricación de muebles

El sector del mueble según lo define la guía de actividades empresariales de la Comunidad Valenciana (2012) es aquel que se dedica a la transformación de ciertos materiales semielaborados como pueden ser tableros o chapas, en productos finales o muebles, destinados al equipamiento de interiores, viviendas, locales o cualquier espacio.

Esta industria podría considerarse que tiene un comienzo en el antiguo Egipto, de donde aún se conservan muestras de mobiliario utilizado en aquella época, por lo que se trata de uno de los sectores más antiguos, que a pesar de que tradicionalmente ha sido considerado un 'arte menor', se empezó a revalorizar a finales del siglo XIX junto con las demás artes asociadas al diseño.

Hoy en día debido a la globalización y la entrada de mercados asiáticos como China, se ha provocado que industrias como la española deban centrarse en aspectos como el diseño o la calidad, puesto que para la mayoría es imposible competir en precios, centrando así competencia en países como Alemania o Italia, los cuales son históricamente grandes productores de mubles con una industria de características más similares a la española.

2.1.1. Dimensión

En España la industria de la fabricación de muebles conforma un sector que, según datos del Instituto Nacional de Estadística (INE), cuenta con un total de 10,008 empresas a fecha del 1 de enero de 2023. De estas empresas, únicamente 41 son grandes empresas, mientras que el resto son pymes, micropymes o empresas sin asalariados.

Tabla 1: Dimensión del sector de la fabricación de muebles en España

Empresas en el sector madera y mueble por estrato de asalariados en España. Año 2022			
CNAE 2009	CNAE 16	CNAE 31	TOTAL
TOTAL	8.864	10.008	18.872
Sin asalariados	2.843	3.652	6.495
Micropymes (1 a 9)	4.894	5.137	10.031
Pymes (10 a 199)	1.112	1.193	2.305
Grandes empresas (>= 200)	15	26	41

Fuente: INE. DIRCE (Empresas a 1 de enero de 2023)

Este sector podría clasificarse como no demasiado grande, con una aportación en 2022 de alrededor de un 0,4% del PIB, puesto que en comparación con sectores como el turismo, el cual representa un 12%, no supone un porcentaje demasiado elevado. Sin embargo, representa una buena parte de la industria manufacturera.

Este sector, además, está caracterizado por contar con buenas cifras de exportaciones, las cuales conforman una importante fuente de ingresos para el sector.

En cuanto a las exportaciones según la Asociación Nacional de Fabricantes y Exportadores de Muebles de España (ANIEME) (Mueble de España, 2022), rondaban los 2,000 millones de euros por año con un crecimiento positivo, el cual duró hasta 2020, cuando debido a la crisis sanitaria, las exportaciones bajaron. Sin embargo, en 2021 estas experimentaron un crecimiento sin precedentes del 25,2% respecto al año anterior, y en 2022 únicamente crecieron un 13%.

Tabla 2: Evolución exportaciones españolas muebles

Enero – Diciembre	2016	2017	2018	2019	2020	2021
Cantidad (Millones de Euros)	2.064	2.113	2.215	2.318	2.086	2.613
Crecimiento (%)	11,1	2,4	4,8	4,7	-10	25,2

Fuente: Mueble de España

Los principales importadores de muebles españoles son los países de la Unión Europea, ya que Francia, Portugal, Alemania y Reino Unido concentran más del 60% de las exportaciones. Por otro lado, Estados Unidos se posiciona como el primer mercado internacional con unas ventas españolas en 2021 de 181733 millones de euros.

Estos datos muestran el tipo de mercados a los que está dirigida la industria española: mercados con buen nivel adquisitivo, maduros, centrados en la calidad y el diseño, e influidos por las tendencias.

2.1.2. Evolución

Tradicionalmente, España ha sido una región productora de muebles, especialmente en localidades como Valencia o Cataluña, donde históricamente los artesanos han trabajado la madera fabricando muebles de alta calidad y diseño.

Sin embargo, a pesar de la tradición en la fabricación de muebles, no fue hasta la segunda mitad del siglo XX que el sector se consolidó completamente, y muchas empresas comenzaron a exportar sus productos, principalmente a países de Europa y América del Norte.

A finales de los años 80, surgió un auge del llamado 'diseño' como término casi artístico. “La acción de diseñar configura así la relación y la dialéctica entre ser y objeto y, por consiguiente, la calidad de esta relación. Y el concepto fundamental de la representación de lo real o imaginario es aún el paradigma rector en el diseño de un objeto sígnico.” (El 'boom' del diseño, 1988)

Gracias a este ‘boom’, España se posicionó como un referente en cuanto a calidad, diseño e innovación en múltiples disciplinas, incluido el sector de la fabricación de muebles.

En los últimos años, la creciente preocupación por la sostenibilidad y el medio ambiente, junto con la aparición y masificación de las nuevas tecnologías, ha hecho que las empresas dedicadas a la fabricación de muebles tomen un rumbo distinto y centren sus recursos en estos aspectos.

Los consumidores buscan cada vez más productos sostenibles, por lo que muchas empresas fabricantes de muebles han cambiado sus modelos de producción, adaptándose también a las nuevas tecnologías e implementando herramientas como el diseño asistido por ordenador (CAD).

“El software de diseño asistido por ordenador, mayormente conocido por las siglas CAD que provienen del inglés Computer-Aided Design, es un software para crear y editar modelos bidimensionales y tridimensionales de objetos físicos.” (Integral Innovation Experts, 2019).

2.2. Nuevas tendencias

Una tendencia se puede considerar como aquello que indica un cambio de dirección o que orienta hacia un sentido específico. Las tendencias, conocidas como 'modas' cuando son pasajeras, podrían definirse como los hábitos que se adquieren en un determinado sector y lo redireccionan.

Dado que la palabra tendencia puede llevar a varias interpretaciones en el contexto de un sector, en este caso se considerarán las tendencias como aquello que guía el rumbo del sector debido principalmente a las demandas de los consumidores.

Para comprender esto mejor, se podría decir que las tendencias influyen en un sector como el de la fabricación de muebles en aspectos como el tamaño de los muebles, los colores, los procesos de fabricación, etc.

Con el paso del tiempo, las tendencias cambian debido a que los consumidores adquieren nuevos gustos y preocupaciones. La era digital ha traído consigo una variedad de nuevas tendencias gracias a las oportunidades que ofrece el acceso a Internet, tanto en términos de información como en las posibilidades de comercio en línea.

Uno de los aspectos que ha ganado mayor relevancia gracias al acceso a la información proporcionada por Internet es la sostenibilidad.

2.2.1. Sostenibilidad

La sostenibilidad se refiere a la característica del desarrollo que abarca la satisfacción de las necesidades de las generaciones actuales sin comprometer la capacidad de satisfacer las necesidades de las generaciones futuras. (Munier, 2005).

El concepto de sostenibilidad tiene su aparición en el 1713 cuando Hans Carl von Carlowitz mencionaba en su libro Sylvicultura oeconomica (Carlowitz, 1713), que era necesario guardar un equilibrio entre la cosecha y el consumo de madera.

La sostenibilidad, como nueva tendencia, está muy relacionada con la conciencia ecológica que comenzó a preocupar a la población a finales del siglo XX y que, con la llegada de Internet y en los últimos años, se ha popularizado, convirtiéndose hoy en día en una de las cuestiones más importantes para la mayoría de la población.

La mayoría de las grandes compañías apuestan por utilizar procesos productivos sostenibles, ya sea por estar comprometidas con la situación global o por la repercusión que esto puede tener en su imagen ante los consumidores.

En cuanto a la sostenibilidad en la fabricación de muebles, existen ciertos aspectos destacados en los que deben centrarse las compañías que buscan adoptar prácticas respetuosas con el medio ambiente.

La primera y más evidente es la obtención de materiales de fuentes sostenibles, como por ejemplo la adquisición de maderas de bosques gestionados de manera responsable.

2.2.2. Estándares

Un estándar según explica Westreicher (2020) es un nivel referencia de algún factor de producción, y conocer estos en un mercado puede ser de vital importancia puesto que es una gran ayuda a la hora de la planificación y evaluación de resultados.

Por lo general, podemos referirnos a los estándares como un nivel de referencia cuantitativo, como sería por ejemplo el nivel de beneficios, o como una variable de tipo cualitativo, como sería por ejemplo la calidad de un producto. Para este trabajo, nos enfocaremos en esta segunda concepción.

En el caso del sector de la fabricación de muebles, si analizamos a qué podrían referirse los estándares, podríamos centrarnos en aspectos como la calidad, los materiales, el diseño o incluso los tamaños.

La gran mayoría de las empresas españolas fabricantes de muebles centran sus estándares en la calidad y el diseño, ya que es lo que más competitividad aporta a este mercado en comparación con otros. Por ejemplo, una empresa fabricante podría hablar de estándares de calidad refiriéndose a la resistencia, durabilidad, procedencia de los materiales, etc.

Por otro lado, en cuanto a los estándares en diseño, ciertas empresas buscan diferenciarse enfocándose en la creación de muebles que sean diferentes a lo que la población está acostumbrada, rompiendo de alguna manera este tipo de estándar. Buscan no solo atraer al consumidor por lo atractivo visualmente del mueble, sino en muchos casos por lo funcional que puede llegar a ser. Las crecientes tendencias hacia los hogares inteligentes, junto con los cambios en la estructura familiar, hacen que los consumidores necesiten, en muchos casos, nuevos diseños nunca antes conocidos que puedan satisfacer sus necesidades.

Por esto mismo, y dado que las tendencias podrían considerarse algo inestables debido a que son muy recientes y surgen casi a diario nuevos retos e inconvenientes para las empresas fabricantes de muebles (como la aparición de un nuevo dispositivo inteligente para el hogar), muchas empresas optan por ofrecer la fabricación de muebles a medida para el consumidor, pudiendo así satisfacer el mayor número posible de necesidades de sus clientes.

2.3. Sitios web de empresas

Los sitios web de empresas o webs corporativas sirven para aportar a las empresas presencia digital, y que estas cuenten con un espacio en internet donde representar tanto sus productos y servicios como información relevante sobre las empresas. Además, estos proporcionan en muchos casos una vía fácil de comunicación con el cliente.

En la era moderna gracias a la aparición de la web 2.0 las webs corporativas son una herramienta fundamental pues estas como explica (Dans, 2007) aparte de dotar de posicionamiento a las empresas permiten a los usuarios participar e interactuar con la propia empresa, lo que permite a las compañías obtener una valiosa información acerca que es lo que buscan, cuáles son los intereses, y que preocupaciones tienen sus consumidores.

Las webs corporativas también otorgan a las compañías un carácter internacional, ya que, como se mencionó, generan una gran exposición, lo que facilita el acceso a mercados internacionales.

Sin embargo, no basta con que exista un sitio web de la empresa; este también debe seguir ciertas normas o criterios que garanticen el éxito de la compañía en Internet. Por ejemplo, al analizar los sitios web de algunas de las empresas más exitosas, es posible encontrar ciertas características comunes, como un diseño profesional, atractivo y sencillo, facilidad de navegación y contenido relevante, entre otras.

Por ello, es importante que las páginas web estén bien estructuradas, y es necesario decidir qué información se incluirá en ellas. Por esta razón, este trabajo se centrará en el análisis de qué palabras o indicadores incluyen las empresas en sus sitios web y cómo influyen en ellos.2.3.2. Funciones

Como se explicaba anteriormente, las páginas web corporativas son de una gran utilidad para las empresas, puesto que estas proporcionan una variedad de funciones útiles para las compañías.

La primera y más evidente sería la presencia en línea, como explican Daries, Cristobal, Martín y Mariné (2017), en un artículo sobre las TIC (tecnologías de la información y comunicación), en el sector turístico los consumidores utilizan cada vez más internet como fuente de información sobre productos y servicios, y esto se ve no solo en este sector, sino que cada vez es más común recurrir a internet para obtener información sobre cualquier tipo de producto o servicio.

La presencia digital permite que los consumidores puedan, de manera sencilla, obtener información tanto corporativa (misión, visión, valores de la compañía…), como acerca de los productos que una empresa ofrece, ofertas o promociones, por lo que una buena presencia digital implica obtener un escaparate en línea.

Otra de las funciones, y siendo en este caso en torno a la que girará el desarrollo del trabajo, es la recogida de información a través de métricas.

Existen diversas herramientas para realizar esta recogida de información, como pueden ser los mapas de calor ‘Un mapa de calor o heatmap es una herramienta digital de análisis que muestra mediante un espectro de colores cálidos y fríos las áreas que atraen más y

menos atención, clics e interacciones de los usuarios de una página web o app.' (Gascó, 2020), los contadores de clics, o lo que se utilizara en esta investigación, el análisis de contenido a partir de las palabras clave.

2.3.3. Huella digital

La huella digital consiste en el rastro que deja cada usuario de internet al navegar con algún dispositivo electrónico, ya sea de manera activa o pasiva (Salgado, 2016)

· **Huella digital activa:** se trata de aquella información que los usuarios comparten deliberadamente en internet, bien sea en forma de publicaciones, interacciones en redes sociales o incluso registrándose en un sitio web o rellenando un formulario.

· **Huella digital pasiva:** esta consiste en aquella información que los usuarios no comparten de forma consciente, sino que son los sitios web los que recopilan información de los usuarios, como puede ser el número de veces que se ingresa en una página o el tiempo que pasa el usuario en cada sección de esta.

La huella digital puede ser una herramienta muy útil para las compañías a la hora de analizar los gustos, las tendencias y las preferencias de los consumidores, así como para poder optimizar la web corporativa de la empresa fijándose en qué partes o secciones son las que más llaman la atención y atraen a los usuarios.

Sin embargo, las compañías no solo deben analizar la huella digital de los compradores, sino que también deben prestar atención en cuidar la suya propia, dado que esta puede ser de gran utilidad para ayudar a la empresa a mantener una imagen positiva hacia los consumidores, así como obtener un buen posicionamiento en la web.

En el caso de la huella digital empresarial, esta estaría conformada tanto por la propia web corporativa de la empresa como por la presencia de esta en redes sociales o incluso las propias huellas digitales individuales de cada uno de los empleados que utilicen medios de la empresa, como puede ser un correo electrónico.

Es por esto que es crucial para la empresa generar una huella digital positiva, lo cual puede lograr a través de interacciones con los usuarios mediante reseñas, opiniones o atención al cliente en línea, o medios de comunicación digitales como pueden ser comunicados de prensa o noticias a través de una página web.

Por ello, es también necesario un buen posicionamiento en línea que haga visible la empresa y su comportamiento para el resto de los usuarios. Es aquí donde juegan un papel importante las palabras clave, fundamentales para el posicionamiento online y para la huella digital de la empresa.

3. Metodología

En este apartado se describirá la metodología o los pasos seguidos durante la investigación para la obtención de resultados. Se describirán tanto los datos como los métodos para el tratamiento de los mismos o los modelos de predicción utilizados.

La herramienta utilizada para el procesamiento de datos será R, un software estadístico y de análisis gráfico creada por Ross Ihaka y Robert Gentleman en 1993. R además de ser un programa estadístico cuenta con un lenguaje propio de programación, por lo que es posible implementar nuevas funciones en el programa constantemente. R es distribuido bajo los términos GNU (General Public License). (Ubeda, 2013)

3.1. Datos

Para la elaboración de la investigación se utilizarán datos de empresas del sector de la fabricación de muebles, los cuales proporcionan información tanto económica o financiera como información acerca de la huella digital, es decir, en este caso, indicadores digitales.

Estos datos servirán para elaborar gráficos, modelos y predicciones que nos ayuden a entender las relaciones que se generan entre las distintas variables. Mediante modelos estadísticos se tratará de obtener predicciones y comprender cómo afectan los indicadores digitales a las variables económicas que conforman las empresas.

3.1.1. Origen de datos

Los datos seleccionados para este trabajo fueron extraídos de la base de datos SABI o sistema de análisis de balances ibéricos, esta es una base de datos la cual cuenta con información tanto general como de las cuentas anuales de más de dos millones de empresas españolas así como de más de 800.000 portuguesas

Estos datos se seleccionaron mediante una búsqueda en la herramienta SABI de aquellas empresas que contaran con la suficiente información y lo más actualizada posible. Por ello se seleccionaron todas las empresas posibles que contasen con datos para el 2022 con el CNAE 2009, el cual hace referencia a la fabricación de muebles, puesto que para años posteriores no existía suficiente información, esta búsqueda dio como resultado 1487 empresas.

Se decidió no aplicar ningún tipo de filtro sobre el tamaño de las empresas, ya que se consideró interesante para el estudio contar con todos los escenarios de empresas posibles. Una de las características a tener en cuenta en la obtención de los datos fue que estos contaran con página web, puesto que los indicadores digitales fueron extraídos a través de estas.

Posteriormente estos datos fueron tratados debido a que no todos resultaron de utilidad para el estudio.

Además, se utilizaron las distintas páginas web de cada una de las empresas para obtener la frecuencia de aparición de los indicadores digitales, es decir las palabras clave.

3.1.2. Variables y observaciones

Una variable puede definirse como una característica, cualidad o propiedad de una observación, la cual puede tomar diversos valores, siendo susceptible de ser utilizada o cuantificada en un estudio. (Oyola-García, 2021)

Las variables pueden ser de diferentes tipos como numéricas, binarias, de carácter textual, etc.

La base de datos utilizada para el estudio cuenta con un total de 215 variables, las cuales podrían dividirse en cuatro grupos bien definidos.

Por un lado, estarían las variables que proporcionan información sobre la empresa como puede ser el nombre de esta, su página web, el cif de la misma, etc.

Tabla 3: Variables de información de la empresa

Nombre	Nombre de la empresa
Código NIF	Identificador de la empresa
website	Nombre de la web
urls_nuevas	Dirección de la web
URL	Dirección actualizada de la web
sizetext	Tamaño de la web

Fuente: Elaboración propia

Posteriormente se encontrarían las variables sobre información económica de la empresa, donde encontramos datos como pueden ser el resultado del ejercicio, el activo o el número de empleados de cada empresa entre otras.

Tabla 4: Variables económicas

IngExp	Ingresos de explotación	Miles de €
Result	Resultado del ejercicio	Miles de €
A	Activo	Miles de €
FP	Fondos propios	Miles de €
ROA	Rentabilidad económica	Porcentaje
EMP	Número de empleados	Unidades
DEUD	Endeudamiento	Porcentaje
VA	Valor agregado	Miles de €

Fuente: Elaboración propia

Finalmente estarían las variables de carácter de indicador digital las cuales se dividen en dos grupos siendo uno la frecuencia de aparición en la página web de una cierta palabra de manera exacta, mientras que el otro cuantifica la frecuencia de aparición de la misma palabra con concordancia amplia es decir que puede aparecer la palabra exacta o algún derivado, para el análisis solo se utilizara el grupo de variables que cuantifican la frecuencia de aparición de la palabra con concordancia amplia, puesto que ofrecen información muy similar al otro grupo pero de manera menos restrictiva, lo que puede aportar valor al análisis.

Tabla 5: Variables indicadores digitales concordancia exacta

keywords_adaptable	Frecuencia de aparición concordancia exacta
keywords_ahorro_de_recursos	Frecuencia de aparición concordancia exacta
keywords_ahorro_energetico	Frecuencia de aparición concordancia exacta
keywords_aislamiento_acustico	Frecuencia de aparición concordancia exacta
keywords_almacenamiento_inteligente	Frecuencia de aparición concordancia exacta
keywords_artesania	Frecuencia de aparición concordancia exacta
keywords_automatizacion	Frecuencia de aparición concordancia exacta
keywords_bienestar	Frecuencia de aparición concordancia exacta
keywords_biodegradable	Frecuencia de aparición concordancia exacta
keywords_calidad	Frecuencia de aparición concordancia exacta
keywords_cero_deshechos	Frecuencia de aparición concordancia exacta
keywords_cero_desperdicio	Frecuencia de aparición concordancia exacta
keywords_certificacion_ecologica	Frecuencia de aparición concordancia exacta
keywords_comercio_justo	Frecuencia de aparición concordancia exacta
keywords_compacto	Frecuencia de aparición concordancia exacta
keywords_competitividad	Frecuencia de aparición concordancia exacta
keywords_conectividad	Frecuencia de aparición concordancia exacta
keywords_consumo_responsable	Frecuencia de aparición concordancia exacta
keywords_control_por_voz	Frecuencia de aparición concordancia exacta
keywords_desarrollo	Frecuencia de aparición concordancia exacta
keywords_diseno	Frecuencia de aparición concordancia exacta

keywords_diseno_adaptable	Frecuencia de aparición concordancia exacta
keywords_diseno_personalizado	Frecuencia de aparición concordancia exacta
keywords_diseno_sostenible	Frecuencia de aparición concordancia exacta
keywords_diseno_unico	Frecuencia de aparición concordancia exacta
keywords_duradero	Frecuencia de aparición concordancia exacta
keywords_ecologico	Frecuencia de aparición concordancia exacta
keywords_economia_circular	Frecuencia de aparición concordancia exacta
keywords_economico	Frecuencia de aparición concordancia exacta
keywords_eficiencia_energetica	Frecuencia de aparición concordancia exacta
keywords_ergonomia	Frecuencia de aparición concordancia exacta
keywords_espacios_abiertos	Frecuencia de aparición concordancia exacta
keywords_espacios_de_trabajo_en_casa	Frecuencia de aparición concordancia exacta
keywords_espacios_multifuncionales	Frecuencia de aparición concordancia exacta
keywords_espacios_reducidos	Frecuencia de aparición concordancia exacta
keywords_estetica_sostenible	Frecuencia de aparición concordancia exacta
keywords_experiencia	Frecuencia de aparición concordancia exacta
keywords_exportacion	Frecuencia de aparición concordancia exacta
keywords_fabricacion_a_demanda	Frecuencia de aparición concordancia exacta
keywords_fabricacion_bajo_demanda	Frecuencia de aparición concordancia exacta
keywords_fabricacion_local	Frecuencia de aparición concordancia exacta
keywords_flexibilidad	Frecuencia de aparición concordancia exacta
keywords_funcionalidad	Frecuencia de aparición concordancia exacta
keywords_gama	Frecuencia de aparición concordancia exacta
keywords_gestion_de_residuos	Frecuencia de aparición concordancia exacta
keywords_hogar_inteligente	Frecuencia de aparición concordancia exacta
keywords_huella_de_carbono	Frecuencia de aparición concordancia exacta
keywords_iluminacion_led	Frecuencia de aparición concordancia exacta

keywords_iluminacion_natural	Frecuencia de aparición concordancia exacta
keywords_impacto_ambiental	Frecuencia de aparición concordancia exacta
keywords_importacion	Frecuencia de aparición concordancia exacta
keywords_innovacion	Frecuencia de aparición concordancia exacta
keywords_integracion_domotica	Frecuencia de aparición concordancia exacta
keywords_inteligencia_artificial	Frecuencia de aparición concordancia exacta
keywords_iot	Frecuencia de aparición concordancia exacta
keywords_larga_vida	Frecuencia de aparición concordancia exacta
keywords_local	Frecuencia de aparición concordancia exacta
keywords_lujo	Frecuencia de aparición concordancia exacta
keywords_mantenimiento_facil	Frecuencia de aparición concordancia exacta
keywords_manufactura	Frecuencia de aparición concordancia exacta
keywords_marca	Frecuencia de aparición concordancia exacta
keywords_materiales	Frecuencia de aparición concordancia exacta
keywords_materiales_innovadores	Frecuencia de aparición concordancia exacta
keywords_materiales_naturales	Frecuencia de aparición concordancia exacta
keywords_medida	Frecuencia de aparición concordancia exacta
keywords_mercado	Frecuencia de aparición concordancia exacta
keywords_minimalismo	Frecuencia de aparición concordancia exacta
keywords_movilidad_urbana	Frecuencia de aparición concordancia exacta
keywords_muebles_a_medida	Frecuencia de aparición concordancia exacta
keywords_muebles_inteligentes	Frecuencia de aparición concordancia exacta
keywords_muebles_modulares	Frecuencia de aparición concordancia exacta
keywords_muebles_transformables	Frecuencia de aparición concordancia exacta
keywords_muebles_versatiles	Frecuencia de aparición concordancia exacta
keywords_nacional	Frecuencia de aparición concordancia exacta
keywords_natural	Frecuencia de aparición concordancia exacta

keywords_nuevo	Frecuencia de aparición concordancia exacta
keywords_nuevos_materiales	Frecuencia de aparición concordancia exacta
keywords_optimizacion_del_espacio	Frecuencia de aparición concordancia exacta
keywords_pais	Frecuencia de aparición concordancia exacta
keywords_personalizacion	Frecuencia de aparición concordancia exacta
keywords_precio	Frecuencia de aparición concordancia exacta
keywords_produccion_bajo_demanda	Frecuencia de aparición concordancia exacta
keywords_produccion_etica	Frecuencia de aparición concordancia exacta
keywords_producicion	Frecuencia de aparición concordancia exacta
keywords_reciclado	Frecuencia de aparición concordancia exacta
keywords_recursos	Frecuencia de aparición concordancia exacta
keywords_recursos_renovables	Frecuencia de aparición concordancia exacta
keywords_reduccion_de_emisiones	Frecuencia de aparición concordancia exacta
keywords_resistente	Frecuencia de aparición concordancia exacta
keywords_reutilizacion	Frecuencia de aparición concordancia exacta
keywords_saludable	Frecuencia de aparición concordancia exacta
keywords_seguridad_del_hogar	Frecuencia de aparición concordancia exacta
keywords_sostenibilidad	Frecuencia de aparición concordancia exacta
keywords_sostenible	Frecuencia de aparición concordancia exacta
keywords_tecnologia_integrada	Frecuencia de aparición concordancia exacta
keywords_tecnologias_emergentes	Frecuencia de aparición concordancia exacta
keywords_tecnologia_sostenible	Frecuencia de aparición concordancia exacta
keywords_transparencia_en_la_cadena_de_suministro	Frecuencia de aparición concordancia exacta
keywords_ventilacion_mejorada	Frecuencia de aparición concordancia exacta

Fuente: Elaboración propia

Tabla 6: Variables indicadores digitales concordancia amplia

kwstems_es_adapt	Frecuencia de aparición concordancia amplia
kwstems_es_ahorr_de_recurs	Frecuencia de aparición concordancia amplia
kwstems_es_ahorr_energet	Frecuencia de aparición concordancia amplia
kwstems_es_aislamient_acust	Frecuencia de aparición concordancia amplia
kwstems_es_almacen_inteligent	Frecuencia de aparición concordancia amplia
kwstems_es_artesani	Frecuencia de aparición concordancia amplia
kwstems_es_automatizacion	Frecuencia de aparición concordancia amplia
kwstems_es_bienest	Frecuencia de aparición concordancia amplia
kwstems_es_biodegrad	Frecuencia de aparición concordancia amplia
kwstems_es_calid	Frecuencia de aparición concordancia amplia
kwstems_es_cer_deshech	Frecuencia de aparición concordancia amplia
kwstems_es_cer_desperdici	Frecuencia de aparición concordancia amplia
kwstems_es_certificacion_ecolog	Frecuencia de aparición concordancia amplia
kwstems_es_comerci_just	Frecuencia de aparición concordancia amplia
kwstems_es_compact	Frecuencia de aparición concordancia amplia
kwstems_es_competit	Frecuencia de aparición concordancia amplia
kwstems_es_conect	Frecuencia de aparición concordancia amplia
kwstems_es_consum_respons	Frecuencia de aparición concordancia amplia
kwstems_es_control_por_voz	Frecuencia de aparición concordancia amplia
kwstems_es_desarroll	Frecuencia de aparición concordancia amplia
kwstems_es_disen	Frecuencia de aparición concordancia amplia
kwstems_es_disen_adapt	Frecuencia de aparición concordancia amplia
kwstems_es_disen_personaliz	Frecuencia de aparición concordancia amplia
kwstems_es_disen_sosten	Frecuencia de aparición concordancia amplia
kwstems_es_disen_unic	Frecuencia de aparición concordancia amplia
kwstems_es_durader	Frecuencia de aparición concordancia amplia

kwstems_es_ecolog	Frecuencia de aparición concordancia amplia
kwstems_es_economi_circul	Frecuencia de aparición concordancia amplia
kwstems_es_econom	Frecuencia de aparición concordancia amplia
kwstems_es_eficient_energet	Frecuencia de aparición concordancia amplia
kwstems_es_ergonomi	Frecuencia de aparición concordancia amplia
kwstems_es_espaci_abiert	Frecuencia de aparición concordancia amplia
kwstems_es_espaci_de_trabaj_en_cas	Frecuencia de aparición concordancia amplia
kwstems_es_espaci_multifuncional	Frecuencia de aparición concordancia amplia
kwstems_es_espaci_reduc	Frecuencia de aparición concordancia amplia
kwstems_es_estet_sosten	Frecuencia de aparición concordancia amplia
kwstems_es_experient	Frecuencia de aparición concordancia amplia
kwstems_es_exportacion	Frecuencia de aparición concordancia amplia
kwstems_es_fabricacion_a_demand	Frecuencia de aparición concordancia amplia
kwstems_es_fabricacion_baj_demand	Frecuencia de aparición concordancia amplia
kwstems_es_fabricacion_local	Frecuencia de aparición concordancia amplia
kwstems_es_flexibil	Frecuencia de aparición concordancia amplia
kwstems_es_funcional	Frecuencia de aparición concordancia amplia
kwstems_es_gam	Frecuencia de aparición concordancia amplia
kwstems_es_gestion_de_residu	Frecuencia de aparición concordancia amplia
kwstems_es_hog_inteligent	Frecuencia de aparición concordancia amplia
kwstems_es_huell_de_carbon	Frecuencia de aparición concordancia amplia
kwstems_es_iluminacion_led	Frecuencia de aparición concordancia amplia
kwstems_es_iluminacion_natural	Frecuencia de aparición concordancia amplia
kwstems_es_impact_ambiental	Frecuencia de aparición concordancia amplia
kwstems_es_importacion	Frecuencia de aparición concordancia amplia
kwstems_es_innovacion	Frecuencia de aparición concordancia amplia
kwstems_es_integracion_domot	Frecuencia de aparición concordancia amplia

kwstems_es_inteligent_artificial	Frecuencia de aparición concordancia amplia
kwstems_es_iot	Frecuencia de aparición concordancia amplia
kwstems_es_larg_vid	Frecuencia de aparición concordancia amplia
kwstems_es_local	Frecuencia de aparición concordancia amplia
kwstems_es_luj	Frecuencia de aparición concordancia amplia
kwstems_es_manten_facil	Frecuencia de aparición concordancia amplia
kwstems_es_manufactur	Frecuencia de aparición concordancia amplia
kwstems_es_marc	Frecuencia de aparición concordancia amplia
kwstems_es_material	Frecuencia de aparición concordancia amplia
kwstems_es_material_innov	Frecuencia de aparición concordancia amplia
kwstems_es_material_natural	Frecuencia de aparición concordancia amplia
kwstems_es_med	Frecuencia de aparición concordancia amplia
kwstems_es_merc	Frecuencia de aparición concordancia amplia
kwstems_es_minimal	Frecuencia de aparición concordancia amplia
kwstems_es_movil_urban	Frecuencia de aparición concordancia amplia
kwstems_es_muebl_a_med	Frecuencia de aparición concordancia amplia
kwstems_es_muebl_inteligent	Frecuencia de aparición concordancia amplia
kwstems_es_muebl_modular	Frecuencia de aparición concordancia amplia
kwstems_es_muebl_transform	Frecuencia de aparición concordancia amplia
kwstems_es_muebl_versatil	Frecuencia de aparición concordancia amplia
kwstems_es_nacional	Frecuencia de aparición concordancia amplia
kwstems_es_natural	Frecuencia de aparición concordancia amplia
kwstems_es_nuev	Frecuencia de aparición concordancia amplia
kwstems_es_nuev_material	Frecuencia de aparición concordancia amplia
kwstems_es_optimizacion_del_espaci	Frecuencia de aparición concordancia amplia
kwstems_es_pais	Frecuencia de aparición concordancia amplia
kwstems_es_personalizacion	Frecuencia de aparición concordancia amplia

kwstems_es_preci	Frecuencia de aparición concordancia amplia
kwstems_es_produccion_baj_demand	Frecuencia de aparición concordancia amplia
kwstems_es_produccion_etic	Frecuencia de aparición concordancia amplia
kwstems_es_producicion	Frecuencia de aparición concordancia amplia
kwstems_es_recicl	Frecuencia de aparición concordancia amplia
kwstems_es_recurs	Frecuencia de aparición concordancia amplia
kwstems_es_recurs_renov	Frecuencia de aparición concordancia amplia
kwstems_es_reduccion_de_emision	Frecuencia de aparición concordancia amplia
kwstems_es_resistent	Frecuencia de aparición concordancia amplia
kwstems_es_reutilizacion	Frecuencia de aparición concordancia amplia
kwstems_es_salud	Frecuencia de aparición concordancia amplia
kwstems_es_segur_del_hog	Frecuencia de aparición concordancia amplia
kwstems_es_sostenibil	Frecuencia de aparición concordancia amplia
kwstems_es_sosten	Frecuencia de aparición concordancia amplia
kwstems_es_tecnologi_integr	Frecuencia de aparición concordancia amplia
kwstems_es_tecnologi_emergent	Frecuencia de aparición concordancia amplia
kwstems_es_tecnologi_sosten	Frecuencia de aparición concordancia amplia
kwstems_es_transparent_en_la_caden_de_suministr	Frecuencia de aparición concordancia amplia
kwstems_es_ventilacion_mejor	Frecuencia de aparición concordancia amplia

Fuente: Elaboración propia

En cuanto a las observaciones estas serían las diversas empresas que conforman la base de datos y los valores que toman para cada una de las distintas variables.

Esta base de datos cuenta con 1487 observaciones, aunque es posible que no todas puedan utilizarse, ya que pueden existir datos erróneos o anómalos que distorsionen el estudio por lo que será necesario llevar a cabo un preprocesamiento de datos.

3.2. Preprocesamiento de datos

El primer paso a seguir con la base de datos es el preprocesamiento de estos, es decir, este paso consistirá en adecuar los datos para la correcta realización del estudio según lo requieran las distintas técnicas de análisis seleccionadas. Para ello, lo primero que se debe realizar es un filtrado de los mismos, con el objetivo de que la base quede lista para el análisis.

Para realizar el preprocesamiento, se debe hacer un análisis simple de las distintas variables y observaciones que componen la base, para así poder conocer la naturaleza de estos, entre qué valores oscilan los datos que componen cada variable, de qué tipo son, binarios, de texto, numéricos, etc.

Es importante saber qué tipo de variables son las que se van a utilizar en el análisis, puesto que es posible que no todas sean de utilidad o que alguna contenga datos erróneos.

De igual manera, se realiza un análisis mediante gráficos como el de caja y bigotes para conocer la distribución de los datos y si es necesario realizar alguna transformación o eliminación.

Por ello lo siguiente, será realizar los distintos procesos necesarios para dejar la base de datos preparada, en este caso dado la alta presencia de datos extremos se opta por eliminar aquellos que superen unos límites establecidos como 1.5 veces el rango intercuartílico por debajo del primer cuartil y por encima del tercer cuartil, una vez realizado esto deben imputarse los datos eliminados, en este caso mediante la librería 'mice' de R.

Una vez el preprocesamiento de datos se realice y la base de datos quede preparada se puede pasar a la elaboración del estudio de estos.

3.3. Métodos estadísticos

Una de las técnicas más utilizadas en el análisis de datos para la obtención de conclusiones es la aplicación de métodos estadísticos, los cuales pueden ser utilizados para diversos tipos de aplicaciones, como clasificar observaciones, realizar predicciones o identificar relaciones.

Cada modelo es utilizado para alguna de las aplicaciones antes mencionadas, pudiendo clasificar estos, además, según su naturaleza en modelos de aprendizaje supervisado y no supervisado. Algunos modelos usualmente utilizados son el análisis de componentes principales, el análisis cluster, el análisis de modelos de clasificación o el análisis de modelos de regresión.

Gracias a la combinación de estos pueden obtenerse resultados que resultan útiles a la hora de elaborar conclusiones.

3.3.1 Análisis PCA

El primer modelo para elaborar es el análisis de componentes principales o PCA, esta es, una técnica de aprendizaje no supervisado con la cual se trata de obtener predicciones, reduciendo la dimensionalidad sin perder demasiada información, además sirve como herramienta de visualización de datos (Tarazona Campos) .

Este modelo es utilizado para el estudio de variables numéricas debido a la implicación de cálculos matemáticos para su elaboración, aunque existen opciones para incluir variables categóricas en el modelo.

Para realizar el análisis PCA primero es necesario realizar el cálculo de las componentes principales las cuales podrían definirse como la combinación lineal de las variables originales. Es decir que la componente principal 1 es aquella que explica tanta variabilidad de los datos como sea posible, mientras que la componente principal 2 explicara la mayor parte de la variabilidad no explicada por la componente principal 1.

Para el análisis pueden extraerse varias componentes principales, sin embargo, solo será necesario utilizar aquellas que representen la mayor parte de la variabilidad de los datos.

Por lo tanto, en una matriz X con I observaciones y J variables el cálculo de las componentes principales se realizaría de la siguiente manera:

Ecuación 1: Cálculo componentes principales

$$PC_k = t_k = p_{k1}X_1 + p_{k2}X_2 + \ldots + p_{kJ}X_J$$

Fuente: Análisis de componentes Principales - Inteligencia de negocios I (Tarazona)

Donde:

· **t** = observaciones

·**p** = variables originales

·**X** = matriz de datos con I observaciones y J variables

Una vez obtenidas las componentes principales y seleccionado el número de estas que se utilizaran en el estudio se puede realizar el modelo PCA, cuyo calculo sería el siguiente:

Ecuación 2: Cálculo matemático PCA

$$t_k = p_{k1}X_1 + p_{k2}X_2 + \ldots + p_{kJ}X_J = Xp_k$$

$$T = XP$$

Fuente: Inteligencia de negocios I

Donde:

· **T** es la matriz de scores

· **P** es la matriz de loadings

Siendo la matriz de scores aquella que contiene las coordenadas de los datos u observaciones originales sobre el espacio que ha sido creado con las componentes principales. Mientras que la matriz de loadings es la que contiene los coeficientes de las variables originales en el espacio de las componentes principales.

El porcentaje de variabilidad de x que explica la componente principal PCa se calcularía como:

Ilustración 1: Porcentaje de variabilidad de X explicado por PCa

$$100\lambda_a / \sum_{k=1}^{K} \lambda_k$$

Fuente: Análisis de componentes Principales - Inteligencia de negocios I (Tarazona)

Donde:

·**λk** = varianza asociada a la PCk

3.3.2 Análisis cluster

Otro modelo que servirá para realizar el estudio es análisis cluster o de "conglomerados", este se trata, como explica , de un método descriptivo no supervisado el cual hace una clasificación de un conjunto de datos heterogéneos en grupos o 'clusters' según las similitudes y diferencias que pueda encontrar en los valores que conforman cada una de las observaciones (Tarazona Campos, Inteligencia de negocios II).

Existen diversos tipos de técnicas de clustering como son el clustering jerárquico, el cual consiste en estructurar los diversos clusters de manera jerárquica lo que deja como resultado un dendograma indexado, o los métodos de partición en los cuales se clasifican las observaciones en un número prefijado de clusters.

También existen otras técnicas como son los métodos híbridos, fuzzy clustering o el clustering basado en modelos.

Antes de realizar el análisis cluster es posible conocer si existe alguna tendencia de agrupamiento en los datos a través del estadístico de Hopkins, el cual selecciona un número concreto de observaciones y calcula la distancia existente entre cada uno de ellos para posteriormente simular el mismo número de observaciones a partir de una distribución uniforme con la misma variación que los datos originales, calcular la distancia entre los elementos simulados más cercanos y así obtener el estadístico H.

Ecuación 3: Estadístico de Hopkins

$$H = \frac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} y_i + \sum_{i=1}^{m} x_i}$$

Fuente: Análisis Clustering - Inteligencia de negocios II (Tarazona)

·**m**: Es el número de puntos seleccionados aleatoriamente del conjunto de datos.

·**yi:** Es la distancia mínima desde cada punto aleatorio generado artificialmente iii en el espacio de características al punto más cercano en el conjunto de datos original.

·**xi:** Es la distancia mínima desde cada punto aleatorio iii seleccionado del conjunto de datos original al punto más cercano dentro de ese mismo conjunto de datos.

Para realizar el análisis antes es necesario seleccionar una medida de similitud o distancia, la cual será utilizada posteriormente por el algoritmo que se seleccione.

·**Medidas de similitud**: Estas cuantifican lo parecidas que son las distintas observaciones.

·**Medidas de distancia**: Estas cuantifican el grado de diferencia que existe entre las observaciones.

lustración 2: Medidas de distancia

- Distancia euclídea: $d_{ii'} = \sqrt{\sum_{j=1}^{J}(x_{ij} - x_{i'j})^2} = \sqrt{(\vec{x}_i - \vec{x}_{i'})'(\vec{x}_i - \vec{x}_{i'})}$
- Distancia euclídea estandarizada:

$$d_{ii'} = \sqrt{\sum_{j=1}^{J}\frac{1}{s_j^2}(x_{ij} - x_{i'j})^2} = \sqrt{(\vec{x}_i - \vec{x}_{i'})'\overline{\overline{D}}^{-1}(\vec{x}_i - \vec{x}_{i'})}$$

- Distancia de Mahalanobis: $d_{ii'} = (\vec{x}_i - \vec{x}_{i'})'\overline{\overline{S}}^{-1}(\vec{x}_i - \vec{x}_{i'})$
- Distancia de Manhattan: $d_{ii'} = \sum_{j=1}^{J}\left|x_{ij} - x_{i'j}\right|$
- Distancia Gi-dos: $\chi^2_{ii'} = \sum_{j=1}^{J}\frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$

Fuente: Inteligencia de negocios II

Con la medida de distancia seleccionada es necesario seleccionar el algoritmo que se aplicara para la elaboración de los clusters.

En cuanto a los algoritmos basados en partición, dos de los más comunes son el k-medias y el k-medioides.

·**K-medias**: Utiliza un número de puntos como centros de los clusters o 'centroides', además de la distancia euclídea para así formar los clusters.

·**K-medoides**: Utiliza medoides, los cuales son elementos de un cluster para los cuales la suma de distancias entre el y el resto de los elementos es mínima. Este algoritmo puede utilizar cualquier medida de distancia para su elaboración.

3.3.3 Métodos de clasificación

Los métodos de clasificación son un tipo de aprendizaje supervisado, es decir se conoce la naturaleza o clase real para cada una de las observaciones que se utilizaran para construir el clasificador, el cual necesita una variable respuesta categórica cualitativa. (Debón, 2023).

Una vez construido el modelo o clasificador, este tratará de asignar a cada una de las observaciones que no fueron utilizadas para crear el modelo en la clase o categoría que le corresponda en función de los valores que presenten para las diversas variables que las conformen.

Estos modelos pueden ser muy útiles en diversos ámbitos, ya sea en medicina para diagnosticar enfermedades según los síntomas que presenten los pacientes, en finanzas, evaluando y clasificando a los clientes para predecir el riesgo de impago de un préstamo o en ámbitos como la consultoría y asesoría proporcionando recomendaciones basadas en la clasificación de empresas, entre otros.

Existen una diversidad de métodos los cuales pueden utilizarse para la clasificación de observaciones, entre los más usados estarían:

1. **Regresión Logit:** La regresión logística es un método utilizado para clasificación el cual trata de predecir la probabilidad de pertenencia a una clase a través de variables independientes.

   Este método utiliza la función logística para modelar la relación entre las variables.

2. **Arboles de decisión**: Este método en el cual la salida que proporciona sería un árbol el cual clasifica las observaciones de la raíz a las hojas y en el que cada nodo especifica el test de algún atributo.

3. **Random forest:** Se basa en la construcción de varios árboles de decisión cada uno con un conjunto de datos de entrenamiento diferente, elaborados a partir de un boostrap, el cual consiste en la selección aleatoria de muestras con remplazo.

   Posteriormente el resultado de la clasificación se obtiene a partir de los 'votos' o elecciones que hicieron la mayoría de árboles para una clase.

4. **Vecino más próximo:** Esta técnica a diferencia de las anteriores no genera un modelo como tal, el cual deba 'entrenarse' sino que se basa en la predicción de un valor a través de las observaciones más cercanas.

   Para la utilización del algoritmo debe decidirse una distancia de medida, como pueden ser la euclídea o la de Manhattan entre otras. Posteriormente el algoritmo selecciona un nímero de observaciones para las cuales buscara cuales son las más cercanas y así contara la clase a la que estas pertenecen para asignar la clase más común a la observación desconocida.

5. **Naive Bayes:** Este algoritmo se basa en la aplicación del teorema de bayes.

Ecuación 4: Teorema de Bayes

$$P(C/X_1, X_2, \ldots, X_p) = \frac{P(X_1, X_2, \ldots, X_p/C)P(C)}{P(X_1, X_2, \ldots, X_p)}$$

Fuente: Inteligencia de Negocios

Donde:

· P(C|X) es la probabilidad de que la clase C sea la correcta dada la evidencia X.

· P(X|C) es la probabilidad de que la evidencia X ocurra dado que la clase C es correcta.

· P(C) es la probabilidad a priori de que la clase C sea correcta.

· P(X) es la probabilidad a priori de que ocurra la evidencia X.

6. **SVM:** Las máquinas de soporte vectorial son un conjunto de algoritmos los cuales pueden ser utilizados para clasificación.

   Estos tratan de buscar el hiperplano que mejor separe las observaciones de una de las clases de la otra.

Una vez elegidos los modelos, pueden utilizarse diversas técnicas de validación, en este caso se utilizó el 'holdout repetido', el cual consiste en repetir el proceso de holdout o entrenamiento y test de los modelos un número determinado de veces con el objetivo de mejorar la fiabilidad de la estimación de los modelos.

Posteriormente una vez obtenidos los resultados, estos podrán ser comparados a través de diversos métodos, como la tasa de acierto de cada uno de los modelos o el área bajo la curva de las curvas ROC, la cual es una herramienta de validación del rendimiento de modelos. Un valor de 1 indica un modelo perfecto, mientras que un valor de 0.5 indica un rendimiento no mejor que el azar.

3.3.4. Modelos de Regresión

Al igual que los métodos de clasificación, los de regresión son métodos de aprendizaje supervisado; sin embargo, en este caso no se predice la clase a la que pertenece una observación, sino el valor de una variable numérica ordinal o continua.

Para realizar esto, se trata de construir un modelo que permita predecir valores para observaciones nuevas a partir de un conjunto de ejemplos de los cuales se conoce el valor de la predicción.

Al igual que en los modelos de clasificación, en la regresión existen diversos métodos que sirven para llevar a cabo las predicciones. Algunos de los más utilizados serían los ya mencionados anteriormente en clasificación, puesto que, a pesar de que existen diferencias en el tipo de predicción, el funcionamiento es muy similar entre ambos.

4. Resultados

En este apartado se pasará a describir los resultados obtenidos tras realizar el análisis de la base de datos seleccionada. Se tratarán de encontrar e interpretar aquellos resultados que puedan proporcionar conclusiones interesantes en relación al objetivo del estudio.

4.1. Análisis descriptivo univariante

El primer análisis necesario a realizar, y que puede proporcionar información interesante para el estudio, sería un análisis descriptivo univariante, en el cual se debe observar cada una de las variables que conforman la base de datos con el objetivo de conocer mejor las observaciones con las que se cuenta.

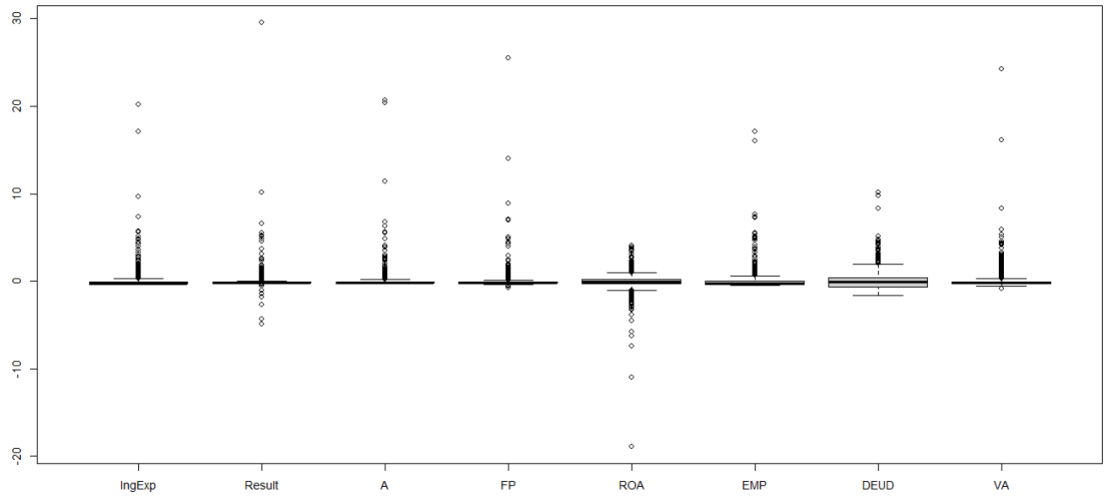
Dado que la base de datos se divide, como se mencionó anteriormente, en cuatro grupos de variables diferenciadas, es necesario seleccionar qué variables se analizarán, puesto que no todas proporcionan información relevante para el estudio.

Las variables que generan un mayor interés para realizar este tipo de análisis, en este caso serían aquellas que proporcionan información económica acerca de las empresas, dado que analizando estas se podrá conocer algo mejor el tipo de observaciones con las que se está trabajando, ya que al ser empresas en este caso no sería lo mismo obtener conclusiones acerca de una pyme que de una gran multinacional, en cuanto al resto de variables de la base de datos, si bien son necesarias para la elaboración del estudio, pues contienen tanto información que permite identificar a las empresas como la frecuencia de aparición de las palabras, no tendría sentido realizar este tipo de análisis univariante en concreto, puesto que muchas son variables de tipo texto, y en el caso de las numéricas, las cuales serían las frecuencias, valores como la distribución, los valores máximos, la media, etc, son irrelevantes, estas variables se utilizaran para otra parte del estudio.

Por esto, tras separar en distintos dataframes las variables y escalar mediante la técnica de estandarización, la cual se consigue restando la media de cada variable y luego dividiendo por su desviación estándar. Las variables que fueron escaladas son aquellas que contenían información económica de la empresa, se realizó un gráfico de caja y bigotes con el cual pudiera conocerse entre que rangos oscilaban los valores de dichas variables.

En este caso se realizó el escalado debido a que para realizar un estudio gráfico y poder comprobar la existencia de anómalos todas las variables deben estar en la misma escala de manera que puedan apreciarse todas juntas con un mismo eje, ya que no podrían representarse juntas por ejemplo la variable de empleados con la del activo, puesto que los valores para uno son mucho más elevados, y esto imposibilitaría el encontrar datos anómalos mediante el estudio gráfico.

Ilustración 3: Boxplot 1- Análisis de anómalos



Fuente: Elaboración propia

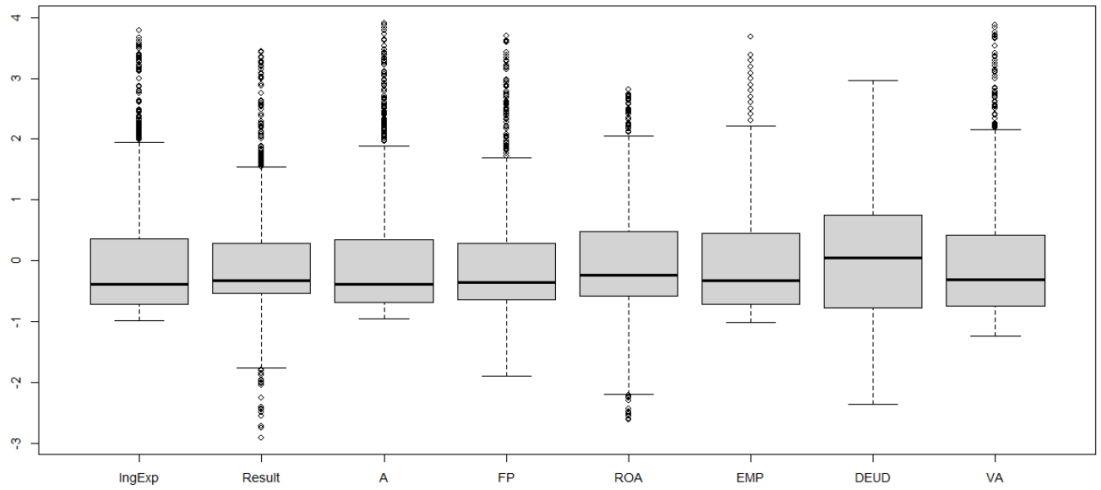
Como se puede apreciar en el gráfico, en el cual se ven las diversas variables, junto con los valores escalados de las diferentes variables, los cuales están representados en el eje y, de manera que puedan visualizarse todas las variables en un mismo gráfico. Existe una gran diversidad de valores para las distintas variables, tanto en las de tamaño como las de rendimiento. Esto puede apreciarse mediante la presencia de los datos anómalos, tomando como anómalos los puntos que pueden apreciarse en el gráfico, los cuales representan valores inusualmente altos o bajos, los cuales no serían una representación fiable de los valores que habitualmente toman estas variables para la mayoría de empresas del sector.

Por lo tanto, para la realización del estudio se consideró necesario la eliminación de este tipo de datos, puesto que a pesar de no ser necesariamente erróneos pueden distorsionar los resultados obtenidos por las distintas técnicas de análisis. Además, este estudio busca analizar el comportamiento general de empresas de este sector.

Una vez eliminados los valores anómalos e imputados aquellos datos para las observaciones que no contaban con más de un 20% de datos faltantes, sustituyendo el valor 'NA' que deja la eliminación de la observación por el valor imputado mediante la librería 'mice', la cual realiza la imputación de manera iterativa, generando varios conjuntos de datos completos basándose en otras variables del conjunto de datos. Se volvió a realizar un nuevo gráfico de caja y bigotes.

En cuanto a las observaciones que contaban con más de un 20% de datos faltantes se optó por eliminarlas, ya que realizar la imputación para los valores de estas observaciones podría causar errores, dado que existe poca información disponible y el valor imputado podría alejarse mucho del real.

Ilustración 4: Boxplot 2- Análisis de anómalos tras imputación



Fuente: Elaboración propia

A pesar de que en este nuevo gráfico pueden seguirse apreciando puntos los cuales indicarían presencia de datos anómalos, estos son mucho menos extremos que los eliminados anteriormente, por lo que no deberían suponer un problema para el análisis.

Con los datos listos para el estudio se realiza un primer análisis visual de las distintas variables económicas.

Tabla 7: Summary variables económicas

IngExp		Result		A		FP	
Min	3,676	Min	-115,908	Min	21,32	Min	-574,63
1ºCuartil	348,958	1ºCuartil	2,852	1ºCuartil	301,78	1ºCuartil	84,48
Mediana	783,640	Mediana	13,089	Mediana	614,81	Mediana	227,08
Media	1276,520	Media	29,470	Media	1020,33	Media	411,08
3ºCuartil	1736,315	3ºCuartil	44,362	3ºCuartil	1378,21	3ºCuartil	557,43
Max	6162,733	Max	201,878	Max	5075,93	Max	2270,19

ROA		EMP		DEUD		VA	
Min	-10,112	Min	1	Min	0,513	Min	-65,83
1ºCuartil	0,806	1ºCuartil	4	1ºCuartil	39,836	1ºCuartil	123,55
Mediana	2,735	Mediana	8	Mediana	60,031	Mediana	287,70
Media	3,965	Media	11,32	Media	58,840	Media	408,26
3ºCuartil	6,531	3ºCuartil	16,49	3ºCuartil	76,977	3ºCuartil	565,80
Max	19,082	Max	49	Max	132,213	Max	1881,78

Fuente: Elaboración propia

Con un simple vistazo se puede conocer algo más acerca de las observaciones que componen la base de datos.

Se puede observar como la base de datos está conformada por pequeñas empresas las cuales tienen una media de unos 11 empleados y alrededor de 1 millón de euros de ingresos de explotación.

Estos datos son positivos para el análisis, ya que como se explicó en el apartado 2.1.1 de este mismo estudio la mayoría de las empresas que componen este sector son pequeñas y medianas empresas.

En cuanto a las variables que informan del tamaño de la empresa es posible observar los ingresos de explotación, variable para la cual la empresa más pequeña tiene un valor de tan solo 3.676€, mientras que la más grande 6.162.733€ lo que parecería indicar una dispersión muy alta, sin embargo, al observar el rango intercuartílico, es posible observar cómo esta dispersión es considerablemente menor, puesto que el primer cuartil se situaría en 348.958€ mientras que el tercero en 1.736.315€.

Para el resto de las variables ocurre algo similar, la variable activo ofrece un valor mínimo de 21.320€, y un máximo de 5.075.930€, sin embargo el primer cuartil es de 301.780€ mientras que el tercero es de 1.378.210€, por lo que mientras el rango de toda la variable es de 5.054.610, el rango intercuartílico es de 1.076.430, es decir casi 5 veces menor, lo que indica que la dispersión no es tan elevada como parecería observando los valores extremos, y que por lo tanto estos valores máximos y mínimos se corresponden efectivamente con alguno de los puntos que podían observarse en el gráfico de caja y bigotes, los cuales ya se comentó que se trataban de observaciones puntuales.

Pasa lo mismo con el resto de las variables de tamaño. En cuanto a aquellas que indican rentabilidad en la empresa, encontramos el resultado del ejercicio y en especial el ROA, estas de igual manera presentan algún valor extremo, siendo los rangos de 317.786 para el resultado y de 29,194 para el ROA, mientras que los rangos intercuartílicos son de 41.510 y de 3,796. Las medias de estas variables son de 29.470€ para el resultado del ejercicio y de 3,965% para el ROA.

Por otra parte, la base de datos cuenta con la variable deuda, la cual no está directamente relacionada con ni con el tamaño ni con la rentabilidad de las empresas, aunque sí que podría informar acerca de la salud financiera de las empresas y tal vez explicar por que alguna empresa a pesar de estar siguiendo una que estrategia que podría parecer correcta no es la más competitiva.

Los valores que muestra esta variable son un mínimo de 0,513% y un máximo de 132,213%, con un rango intercuartílico de 37.141 y una media de 58,840%, lo que parece indicar que las empresas de este sector suelen tener un nivel de endeudamiento que estaría en el límite de lo que habitualmente suele considerarse ideal, es decir entre 40% y 60%.

4.2. Análisis descriptivo bivariante

Una vez se conocen bien las distintas variables que componen la base de datos puede realizarse un primer estudio de las relaciones bivariantes que existen entre estas.

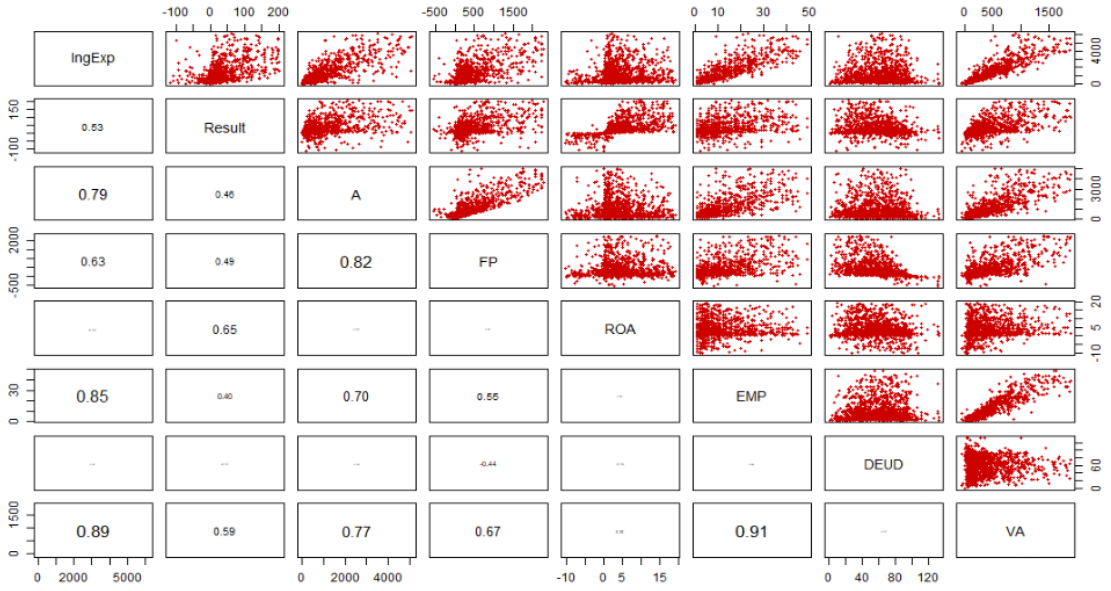
Este estudio si bien no tiene por qué ofrecer conclusiones muy profundas puede proporcionar un contexto para el análisis de manera que se conozca mejor los puntos de mayor interés a estudiar.

Este punto se realizará mediante un simple análisis de correlaciones que muestre si existe algún tipo de relación positiva entre las distintas variables económicas de la base de datos.

4.2.1. Análisis de correlaciones

Para realizar esto se realizó un gráfico en el cual se representó la correlación lineal de todas las posibles combinaciones de variables junto con sus respectivos gráficos de dispersión, para de esta manera comprobar si alguna variable afecta directamente algún otra.

Ilustración 5: Gráfico de correlaciones



Fuente: Elaboración propia

En un primer vistazo pueden apreciarse ciertas relaciones que podrían ser de esperar, por ejemplo, aquellas variables que miden el tamaño de la empresa están altamente correlacionadas.

Por ejemplo, Activo y Fondos propios muestran un coeficiente de 0.70, o Empleados y Va los cuales presentan un coeficiente de correlación de 0.91.

En este caso parece que las relaciones que más llaman la atención son precisamente las muy pequeñas o en algunos casos incluso negativas.

En especial, el ROA muestra unas relaciones muy bajas con todas las variables de tamaño.

Si bien es cierto que el ROA no viene definido por el tamaño de la empresa ya que este es calcula dividiendo el resultado entre el activo de la empresa, es de esperar que aquellas empresas que están más establecidas en el sector cuenten con mayores rentabilidades, ya que conocen mejor el funcionamiento del sector.

Esto parece indicar que el tamaño no es necesariamente un indicador de rendimiento para el sector de la fabricación de muebles. Posteriormente se analizará esto más en profundidad.

Por otra parte, se puede apreciar que la variable deuda no presenta relación con ninguna otra variable.

En cuanto a la relación con las variables de tamaño esto podría ser de esperar, ya que teniendo en cuenta que la deuda está representada por el ratio de endeudamiento no sería de extrañar las empresas tengan niveles de endeudamiento independientes de su tamaño.

4.3. Análisis descriptivo multivariante

Una vez analizado tanto las variables individuales como las relaciones que se establecen entre estas, el siguiente paso sería realizar un análisis descriptivo multivariante.

Este análisis tratara de analizar las relaciones del conjunto de variables a fin de encontrar resultados que proporcionen conclusiones de interés para el estudio.

Para realizar este análisis se aplicarán diversas técnicas con distintos objetivos para así poder tener una visión más amplia en el estudio.

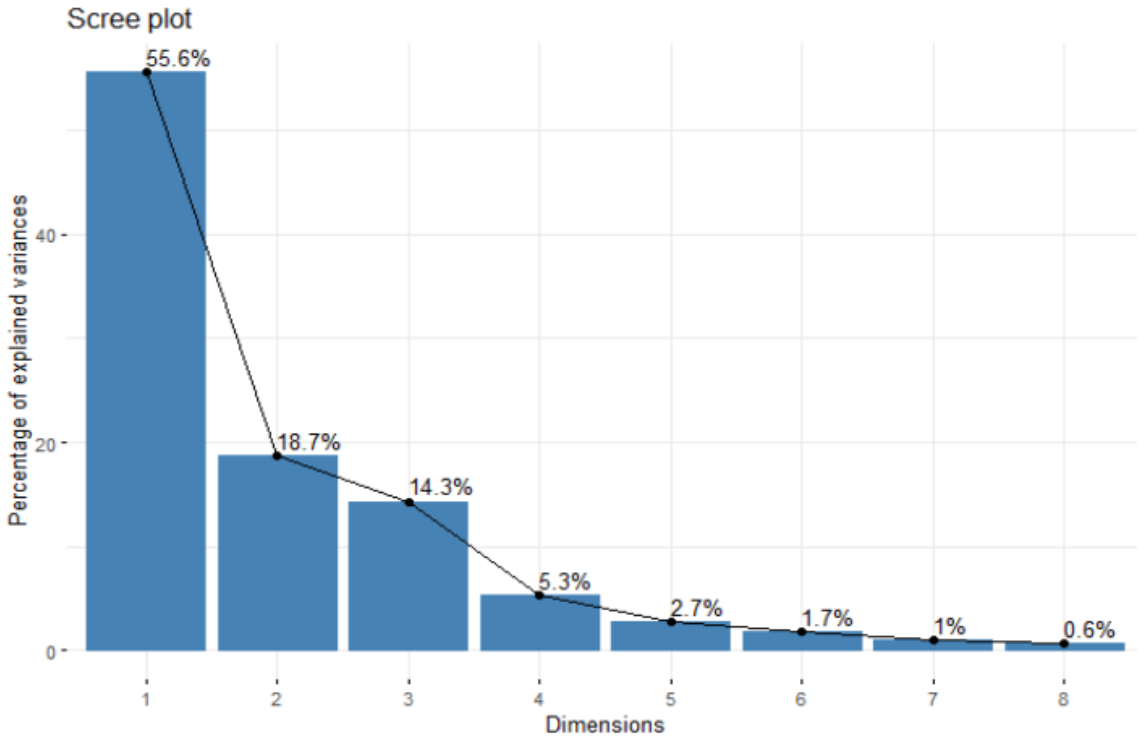
4.3.1. Resultados PCA

La primera técnica a aplicar será el análisis de componentes principales o PCA, el cual, como se explicó en el apartado 3.3.1, es una técnica de aprendizaje automático no supervisado que busca obtener predicciones reduciendo la dimensionalidad sin perder demasiada información, a través de la creación de un nuevo conjunto de variables conocidas como ‘componentes principales’.

Para realizar este análisis el primer paso será seleccionar el número de componentes principales a analizar. Para ello las variables que se utilizaran son las económicas, puesto que son las que nos proporcionan información acerca de la competitividad de cada una de las empresas, y de esta manera será posible observar cómo se distribuyen las observaciones en función de las diferentes variables económicas a fin de determinar que observaciones son más o menos competitivas, para posteriormente buscar por qué lo son.

Para ello se puede utilizar un 'Scree Plot' o gráfico de codo, el cual muestra el porcentaje de varianza explicada por cada componente principal.

Ilustración 6: Scree plot

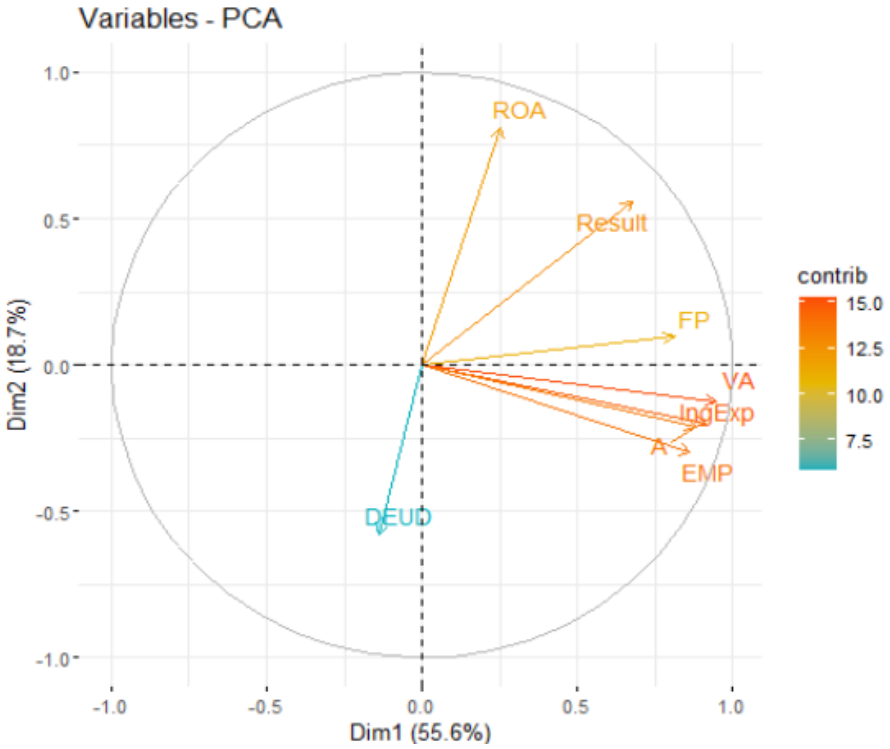


Fuente: Elaboración propia

El criterio que se estableció para la selección de componentes principales fue que estas explicaran más de un 70% de la variabilidad, por lo que como se puede apreciar en el gráfico se seleccionaran las dos primeras, las cuales explican un 74.3% de la variabilidad, el cual podría considerarse como suficiente dado el contexto del análisis en un ámbito de ciencias sociales.

Con el número de componentes principales establecidos se realiza la interpretación del análisis, a través de un 'Loading plot' el cual muestra las distintas variables y su contribución a las componentes principales.

Ilustración 7: Loading plot



Fuente: Elaboración propia

El gráfico muestra como las variables que más contribuyen a la primera componente principal sería el valor agregado, los ingresos de explotación, el activo y los empleados, por lo que se podría deducirse que la primera componente principal explica el tamaño de las empresas.

Por otra parte, puede observarse como la variable que más contribuye a la segunda componente principal sería la ROA, por lo que podría concluirse que esta representa variables relacionadas con la rentabilidad de la empresa.

En cuanto a la variable que representa el resultado de las empresas puede apreciarse como esta contribuye tanto a la primera como a la segunda componente principal, esto puede deberse a que la rentabilidad de las empresas se mide en gran parte en función del resultado de estos, pero también es cierto que por lo general empresas con un mayor tamaño tenderán a tener resultados mayores.

También es posible apreciar como la variable que representa la deuda tiene una correlación negativa con el ROA, lo que a priori quiere decir que para este sector en concreto aquellas empresas que presentan un alto endeudamiento tienen una menor rentabilidad.

Sin embargo, esto no siempre debería ser necesariamente así puesto que las empresas en ciertas ocasiones pueden aprovecharse del endeudamiento para mejorar su rentabilidad.

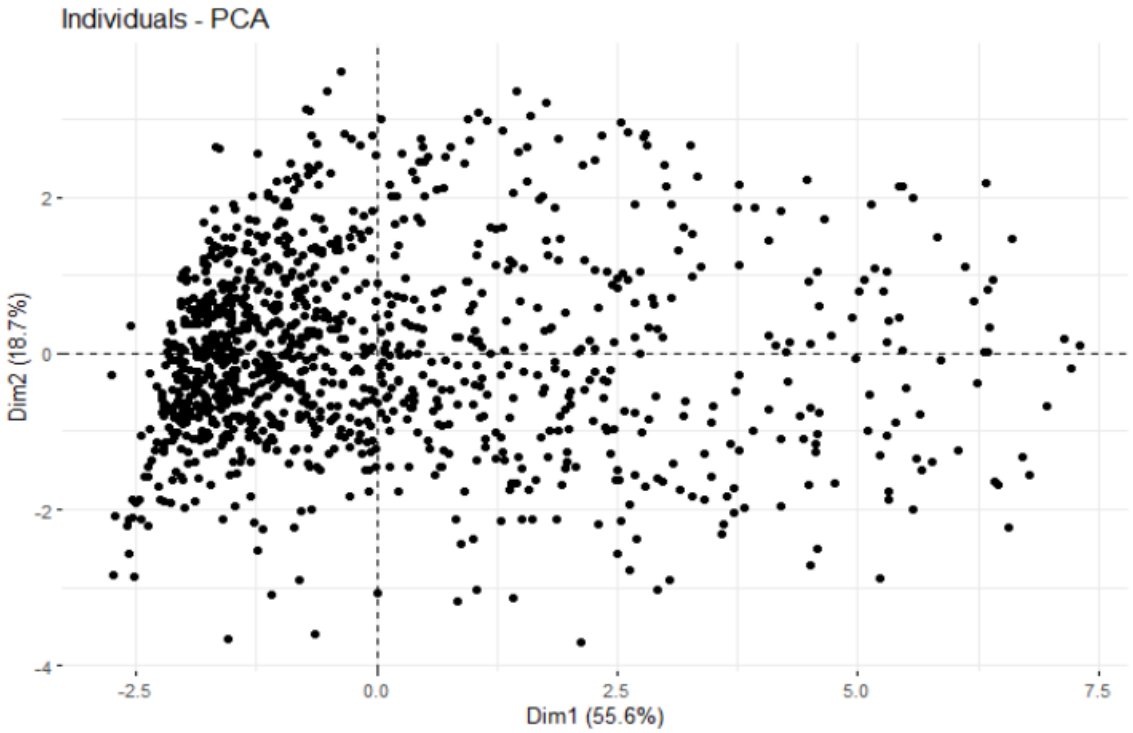
Si bien es cierto que el ROA no tiene directamente en cuenta el endeudamiento la relación no tendría por qué ser necesariamente negativa, sino casi inexistente.

Sin embargo, el color que muestra la variable nos indica que la deuda tiene una contribución muy escasa en el modelo, por lo que podría concluirse que a pesar de que ciertas empresas con un nivel alto de deuda tienen bajos ROA no siempre será de esta manera.

Si se tienen en cuenta los datos analizados en el análisis univariante es posible apreciar como existían ciertas empresas las cuales presentaban pérdidas, por lo que es probable que aquellas empresas con resultados negativos, que presentaran por tanto un ROA negativo hayan tenido que recurrir a endeudamiento, lo que explicaría esta relación.

El siguiente gráfico que puede aportar información es el gráfico de observaciones o 'Score plot' en el que se pueden observar cómo se distribuyen las observaciones en el plano generado con las componentes principales.

Ilustración 8: Score plot



Fuente: Elaboración propia

En este gráfico es posible corroborar que la base de datos está conformada en su mayoría por empresas de tamaño no demasiado grande puesto que las observaciones se muestran agrupadas en los cuadrantes 2 y 3, presentando valores negativos para la dimensión 1, (lo cual no quiere decir que necesariamente estas empresas presenten valores negativos para estas variables, ya que para realizar este análisis es necesario escalar los datos).

En cuanto al resto de empresas con mayor tamaño es posible observar cómo se distribuyen a lo largo de todo el plano, habiendo empresas de una gran variedad de tamaños y rentabilidades.

4.3.2 Resultados Clustering

El siguiente paso en el análisis multivariante fue el análisis clustering, en el cual se trata de un método de aprendizaje no supervisado con el cual se busca agrupar las observaciones en función de las similitudes o diferencias que estas presentan para las diversas variables.

Antes de comenzar el análisis clustering se realizó el cálculo del estadístico de Hopkins

Tabla 8: Estadístico de Hopkins

Min	1º Cuartil	Mediana	Media	3ºCuartil	Max
0,9575	0,9844	0,9866	0,9844	0,9896	0,9936

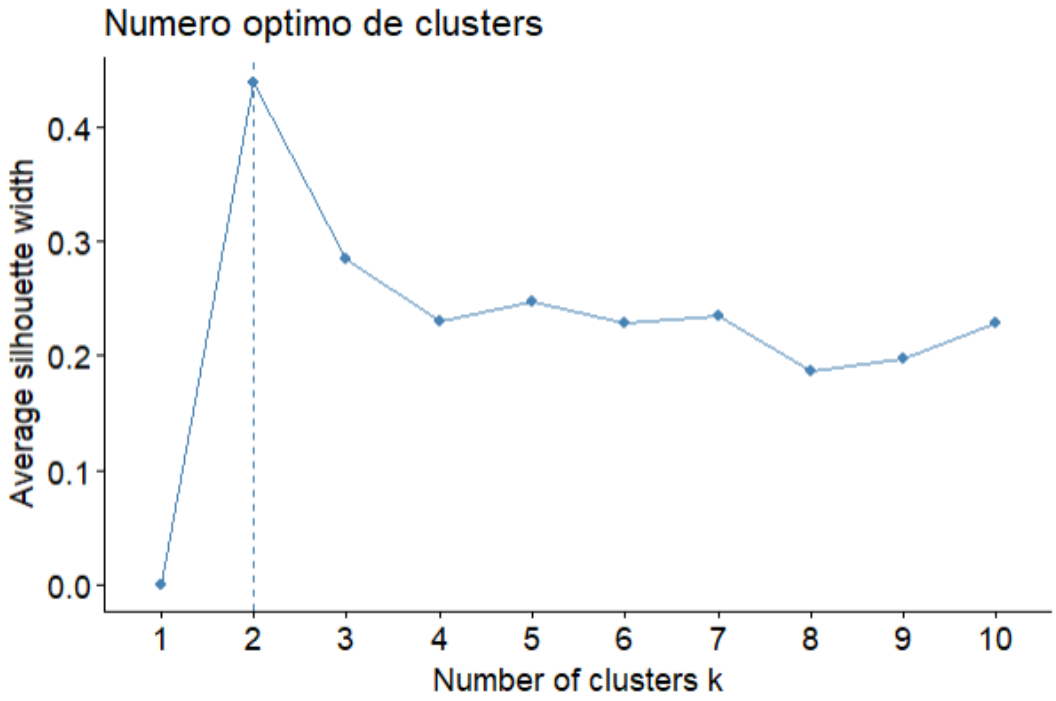
Fuente: Elaboración propia

Este mostró unos valores cercanos a 1, lo cual indica tendencia a agrupamiento.

Posteriormente se seleccionó el algoritmo de agrupamiento el cual se utilizará, para ello se compararon los algoritmos K-medias y K-medoides, por lo que la medida de distancia seleccionada fue la distancia Euclidea, puesto que se buscó la agrupación de las observaciones clasificando aquellas empresas con valores parecidos para las variables económicas.

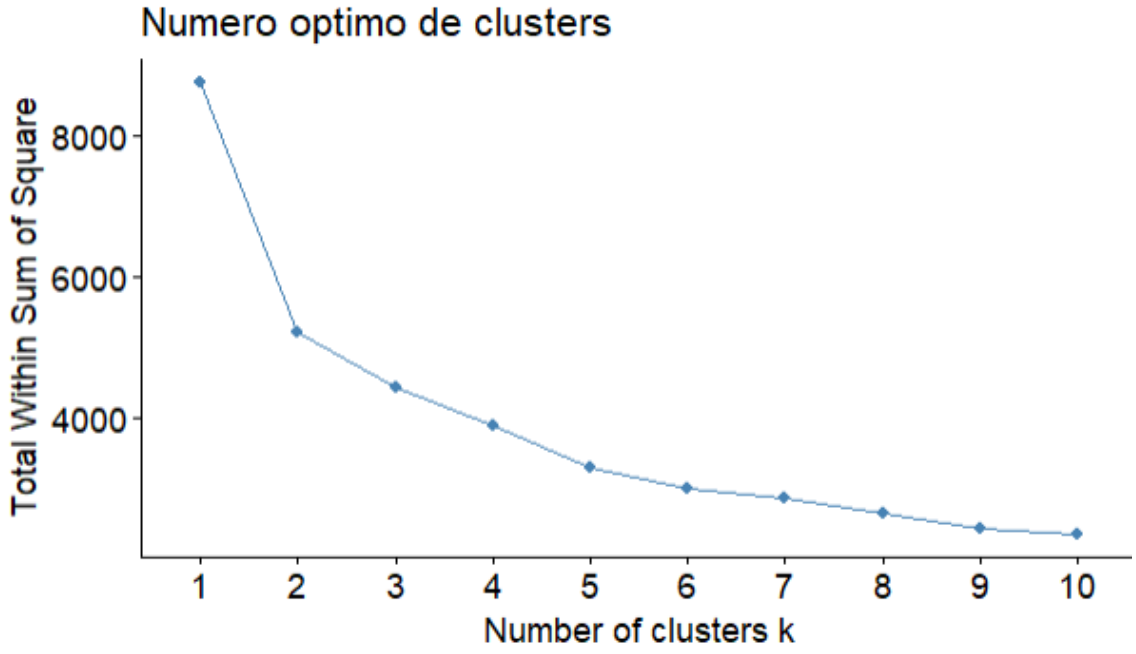
Para realizar esto primero se hallaron el número óptimo de clusters según el coeficiente de Silhouette y la varianza intra cluster.

Ilustración 9: Coeficiente de Silhouette medio k-medias



Fuente: Elaboración propia

Ilustración 10: Varianza intra cluster k-medias



Fuente: Elaboración propia

En cuanto al algoritmo de k-medias, atendiendo al valor que presenta el coeficiente medio de silhouette el número óptimo de clusters sería 2.

Por otra parte, el gráfico de la ilustración número 16, que representa la varianza intracluster muestra como la varianza se ve bastante reducida con dos clusters.

Si bien es cierto que la varianza podría ser menor, para la realización del estudio resulta interesante tener solo dos clusters, por lo tanto, será este el número que se seleccione.

De esta manera según el algoritmo de k-medias los clusters quedarían repartidos de la siguiente manera.

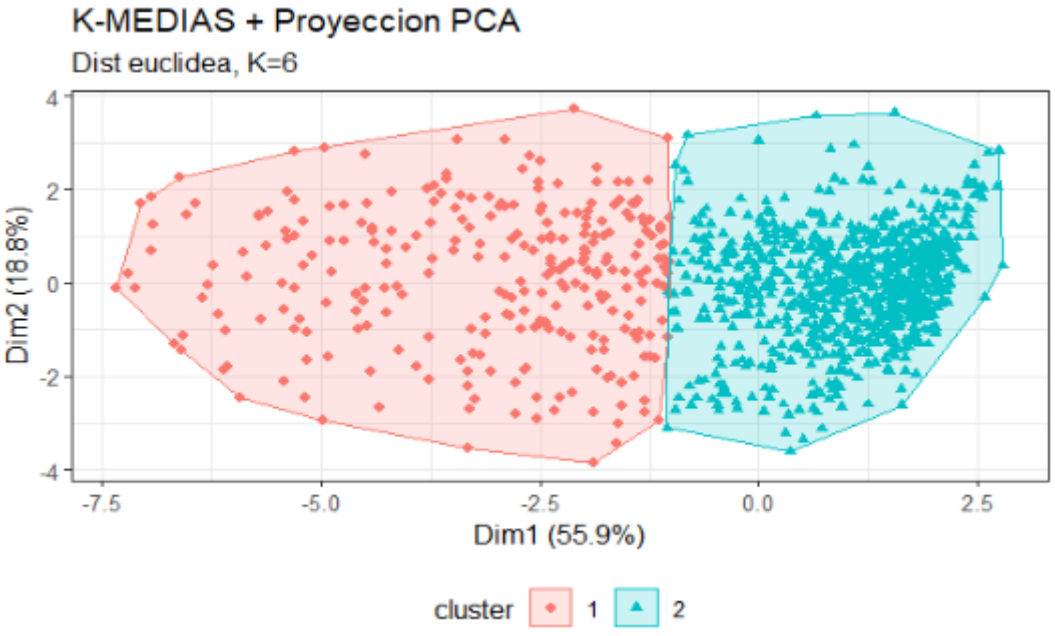
Tabla 9: Reparto clusters k-medias

1	2
274	824

Fuente: Elaboración propia

Con 274 observaciones en el primer grupo y 824 en el segundo, lo cual quiere decir que el primer grupo contaría aproximadamente con un 25% de los datos mientras que el segundo tendría el 75% restante.

Ilustración 11: Clusters + Scores

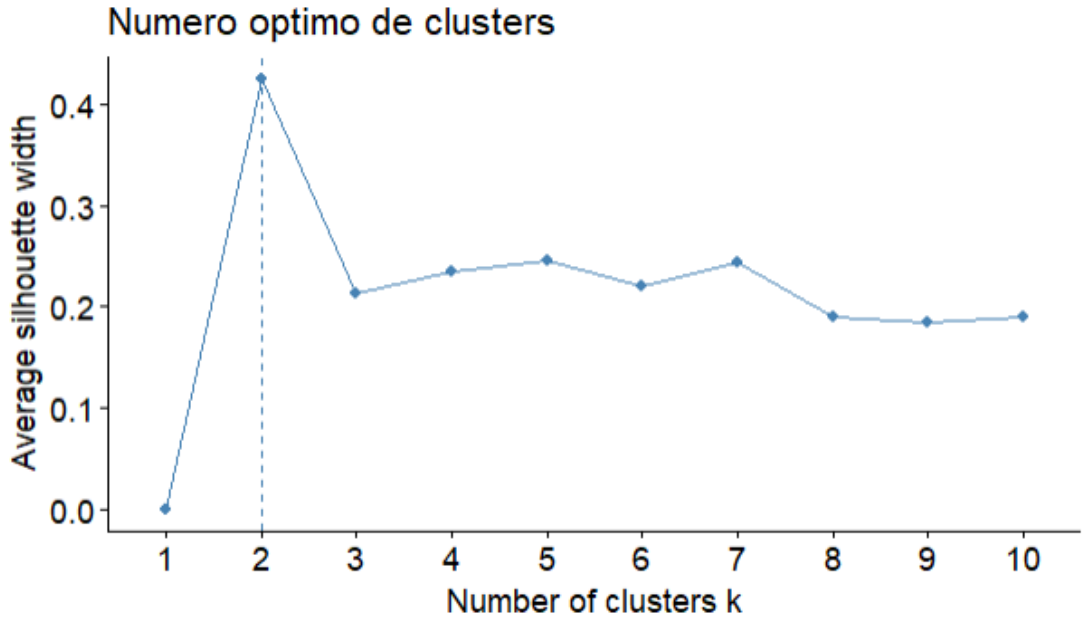


Fuente: Elaboración propia

Observando el gráfico de Scores puede apreciarse mejor la distribución por grupos de las observaciones, se puede apreciar cómo no existe apenas solapamiento, por lo que a priori este parecería un buen método de agrupación.

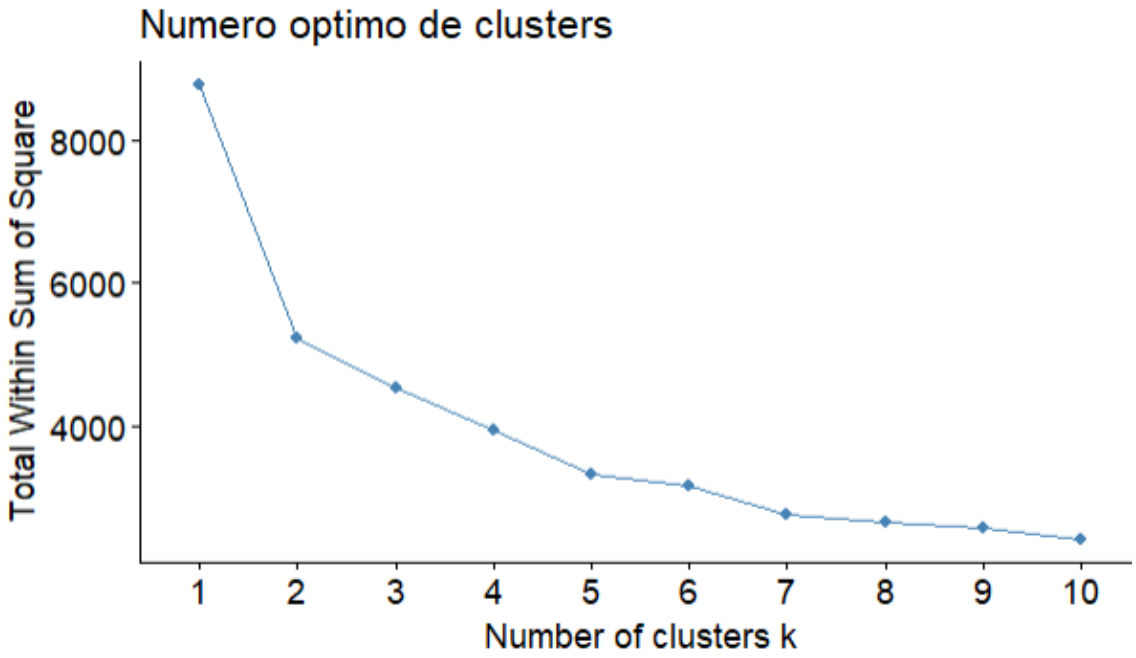
A continuación, se analiza el número óptimo de clusters a seleccionar con el algoritmo k-medoides.

Ilustración 12: Coeficiente medio de Silhouette k-medoides



Fuente: Elaboración propia

Ilustración 13: Varianza intra clusters k-medoides



Fuente: Elaboración propia

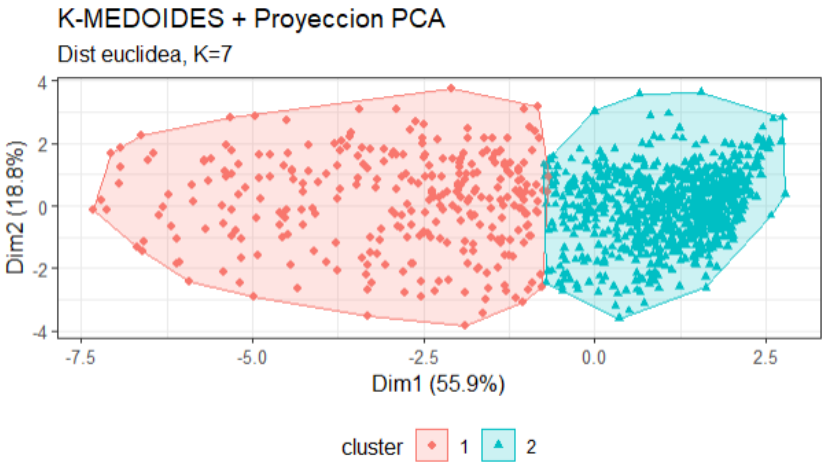
Al igual que en el algoritmo de k-medias ambos gráficos muestran que el número óptimo de clusters es 2, por lo que se toma este y se pasa a analizar la distribución de los grupos mediante este algoritmo.

Tabla 10: Reparto clusters k-medoides

1	2
300	798

Fuente: Elaboración propia

Ilustración 14: Cluster + Scores

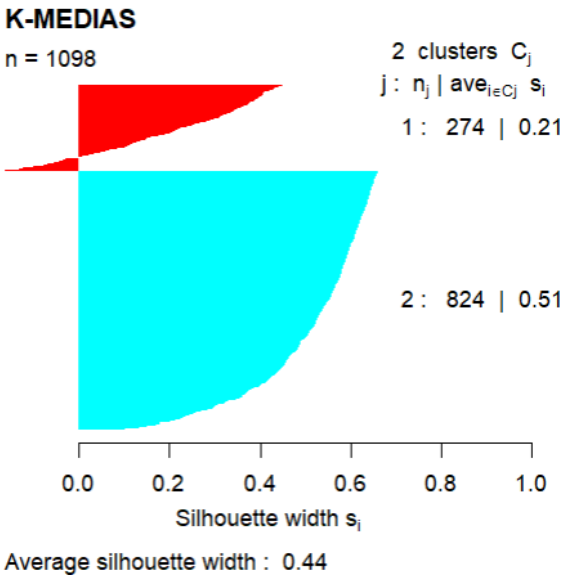


Fuente: Elaboración propia

En este caso las observaciones quedan repartidas de manera que el grupo 1 cuenta con 300 observaciones y el 2 con 798, es decir un reparto de aproximadamente un 30% y 70%.

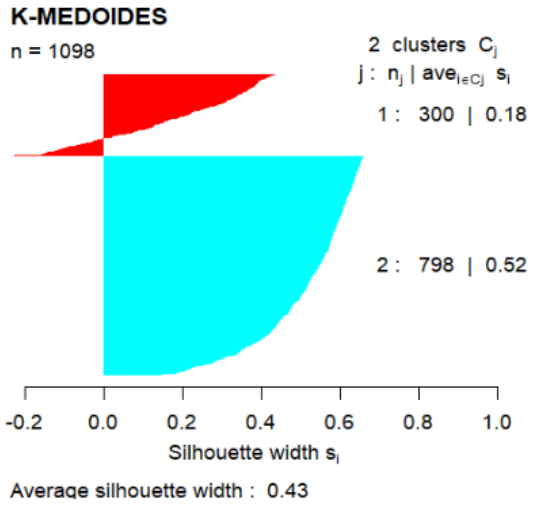
En cuanto al gráfico de Scores se aprecia una distribución muy similar a la obtenida con el algoritmo de k-medias, aunque parece que en este caso existe un mayor solapamiento, por lo que a continuación se realizó la selección del mejor método según el coeficiente de Silhouette.

Ilustración 15: Coeficiente de Silhoutte K-medias



Fuente: Elaboración propia

Ilustración 16: Coeficiente de Silhoutte K-medoides



Fuente: Elaboración propia

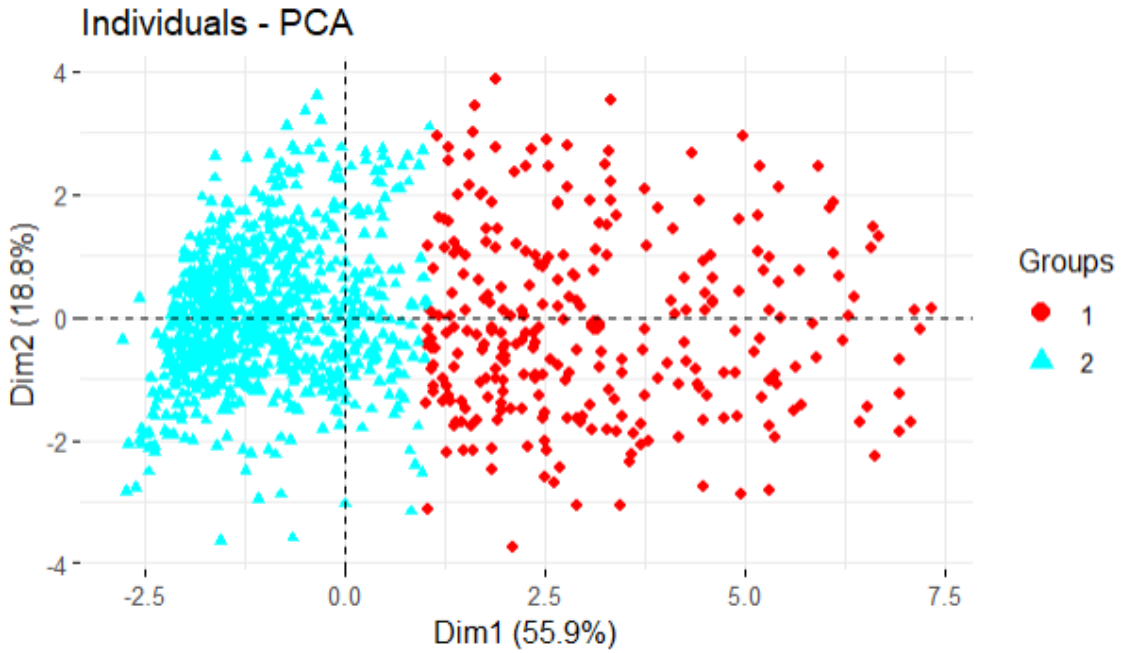
Atendiendo al coeficiente de Silhouette puede determinarse que el algoritmo que mejor clasifica las observaciones es el k-medias, puesto que presenta un coeficiente ligeramente mayor así como un menor número de observaciones mal clasificadas.

Por lo tanto, el método de agrupación seleccionado fue el k-medias.

A continuación, se realizó la interpretación de los resultados obtenidos con el análisis cluster.

Para ello se utilizó el gráfico de observaciones del PCA coloreando las observaciones en función del grupo al que pertenecían.

Ilustración 17: PCA con clusters



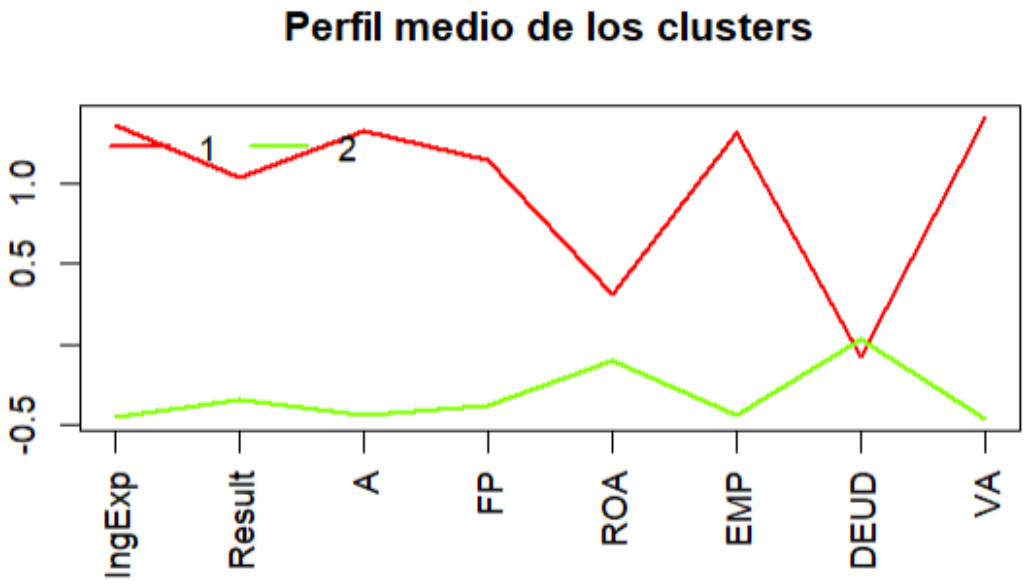
Fuente: Elaboración propia

De esta manera se puede apreciar como la división en grupos se realizó principalmente atendiendo al tamaño de las empresas.

De igual manera se puede observar como el grupo que cuenta con un mayor número de empresas es el que engloba las empresas pequeñas.

Otra forma de interpretación del análisis es a través de un gráfico que represente el perfil medio de los gráficos para las diferentes variables que los conforman.

Ilustración 18: Perfil medio de los clusters



Fuente: Elaboración propia

De esta manera puede obtenerse más información acerca de cómo están agrupadas las observaciones.

Se puede apreciar claramente la diferencia del tamaño medio de las empresas entre los dos grupos.

Algo interesante a analizar es como las empresas pequeñas presentan un endeudamiento mayor.

Además, se puede ver que el grupo de empresas grandes presenta un ROA pequeño atendiendo al resto de variables mientras que el grupo de empresas pequeñas presentan un ROA más elevado teniendo en cuenta el resto de los valores.

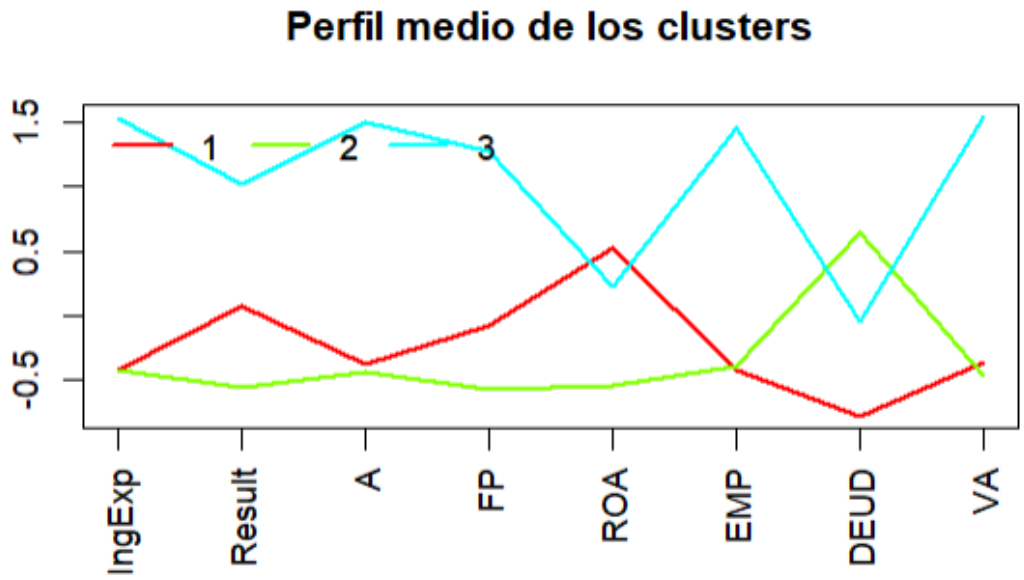
Si se tienen en cuenta los resultados obtenidos en el análisis PCA puede concluirse que habría algún tipo de incongruencia, puesto que la relación entre ROA y deuda casi inexistente.

Esto incita a pensar que puede existir una subagrupación dentro de las empresas pequeñas existiendo un grupo de empresas que probablemente han obtenido una serie de malos resultados y han recurrido a la deuda para intentar sobrellevar la situación.

Y un grupo de empresas que a pesar de ser pequeñas han conseguido optimizar sus recursos obteniendo así unas rentabilidades muy positivas.

Para comprobar esto se realizó una nueva agrupación mediante el método de K-medias, esta vez seleccionando 3 clusters.

Ilustración 19: Perfil medio de los clusters 2



Fuente: Elaboración propia

De esta manera el gráfico confirma las suposiciones puesto que se puede ver como el grupo uno presentando valores muy escasos para variables de tamaño como pueden ser el activo obtiene rentabilidades más altas que empresas pertenecientes al grupo de las clasificadas como grandes.

Por otra parte, el grupo 2 representaría aquellas empresas en declive, las cuales presentan valores bajos tanto para el tamaño como para la rentabilidad, pero no para el endeudamiento siendo este superior incluso al del grupo de las grandes.

El reparto de los grupos con esta partición quedaría de la siguiente manera:

Ilustración 20: Repartos clusters k-medias 2

1	2	3
390	473	235

Fuente: Elaboración propia

4.3.3 Resultados clasificación

Tras realizar la agrupación de las observaciones el siguiente paso lógico es el de elaborar modelos de clasificación los cuales permitan conocer a que grupo de empresas pertenecerá una determinada observación en función de la estrategia que esta emplee.

En este caso se utilizarán los indicadores digitales para tratar de clasificar las observaciones, para ello la clasificación se llevó a cabo utilizando la primera agrupación obtenida por los clusters, donde las empresas fueron catalogadas en función de su tamaño.

Se realizó un Hold-out repetido 100 veces a fin de encontrar cuál es el modelo que mejor clasifica las observaciones, habiendo realizado previamente una partición de los datos para obtener de esta manera datos de entrenamiento y de test. Tras la realización del Hold-out se recopilaron las áreas bajo la curva de ROC de cada uno de los modelos así como sus respectivas tasas de acierto.

Tras realizar el hold-out repetido los resultados obtenidos fueron los siguientes:

Tabla 11: Resultados clasificación 2

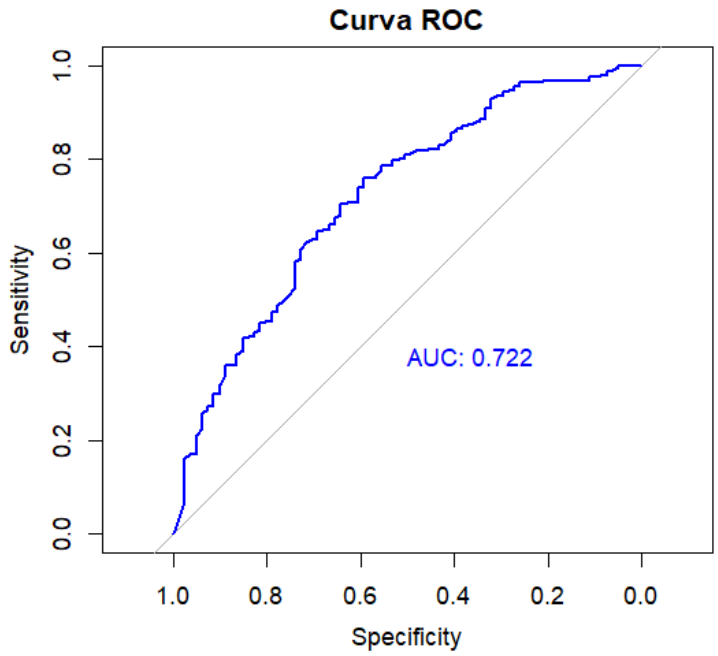
Medias	Logit	Tree	Random Forest	Naive Bayes	SVP	SVM	Vecino
AUC	0,6064003	0,6125136	0,7255296	0,5870520	0,6688412	0,5860571	0,6812888
Tasa de acierto	0,3264024	0,5888110	0,7739634	0,0070427	0,0050305	0,7520427	0,7511280

Fuente: Elaboración propia

Estos resultados muestran valores interesantes, siendo el modelo 'Random Forest' el que mejor clasifica en este caso las observaciones según tanto el área bajo la curva como la tasa de acierto.

Si se representa la curva ROC del modelo Random Forest quedaría de la siguiente manera:

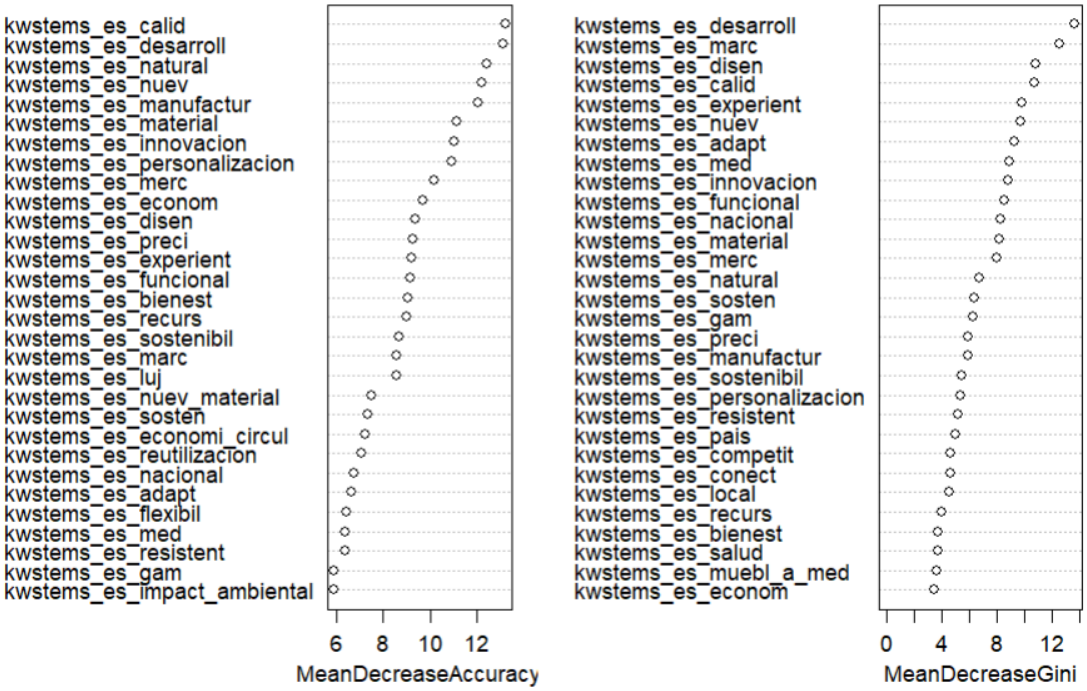
Ilustración 21: Curva ROC modelo Random Forest



Fuente: Elaboración propia

Dado que el modelo Random Forest es el que mejores resultados muestra para la clasificación de empresas por tamaño se pasará analizar cuáles son las variables que más contribuyen a esta clasificación.

Ilustración 22: Importancia variables Random Forest



Fuente: Elaboración propia

El gráfico de la izquierda muestra la disminución media en la precisión del modelo, es decir, cuanto disminuiría la precisión del modelo en caso de que una variable fuera eliminada pudiendo así comparar que variables son las que más impacto tienen en el modelo.

El gráfico de la derecha por otra utiliza el índice de Gini para medir la importancia de las variables.

Atendiendo a los resultados mostrados por el gráfico las palabras que más contribuyen a la clasificación serían 'calidad', 'desarrollo', 'natural', 'nuevo' y 'marca'.

Así se puede concluir que la estrategia de las empresas grandes reside en tratar de ofrecer productos de calidad, teniendo en cuenta el diseño, el desarrollo del producto, etc.

Otro resultado interesante es la aparición de marca entre las palabras que más contribuyen a las predicciones del modelo.

Esto podría deberse a que aquellas empresas grandes usualmente, son las más conocidas por los consumidores y utilizan nombre como marca comercial, lo cual podría indicar que una estrategia interesante a seguir por empresas más pequeñas que estén incursionando en el sector sea la creación y potenciación de una marca propia, o en este caso de la empresa, que los consumidores sean capaces de identificar.

Dado que el análisis de clasificación sirvió para conocer la estrategia que utilizan las empresas más grandes, pero no aquellas que a pesar de su tamaño consiguen maximizar su rentabilidad, se tratara de usar modelos de regresión para cumplir este objetivo.

4.3.4 Resultados regresión

Como se comentaba anteriormente para tratar de conseguir comprender que estrategia utilizan las empresas con altas rentabilidades independientemente de su tamaño se realizará un análisis de regresión, tomando como la variable respuesta la segunda componente principal del PCA que se realizó con anterioridad, puesto que como se vio en su respectivo apartado esta mide principalmente la rentabilidad de las empresas.

Para realizar este análisis, al igual que en clasificación, se utilizaron los indicadores digitales, y se elaboró un hold-out repetido, partiendo los datos para entrenamiento y para test.

Dado que los resultados de regresión no pueden medirse con una tasa de acierto se utilizaron dos medidas de error, el NMAE y el RMSE.

- **NMAE**: Medida de bondad de ajuste relativa que permite medir la distancia entre las predicciones y las observaciones con independencia de la magnitud de los datos.

- **RMSE**: Medida de bondad de ajuste absoluta que permite evaluar la precisión del modelo.

Tabla 12: Resultados regresión

	Regresión	Árbol	Random Forest	Vecino	SVP	SVM
NMAE	1,553627	1,041344	1,002117	1,075663	1,013231	0,999743
RMSE	6,13947	1,270240	1,219472	1,310809	1,228957	1,222626

Fuente: Elaboración propia

Los resultados en este caso muestran cómo según el NMAE el modelo que mejor predice los valores para la segunda componente principal es el SVM, mientras que según el RMSE sería el Random Forest.

Sin embargo, como la diferencia entre el NMAE que presentan el modelo Random Forest y el modelo SVM no es demasiado elevado, se comprobará mediante un test ANOVA si existen diferencias significativas respecto a la medida NMAE entre los dos modelos.

Tabla 13: Test ANOVA

	Df	Sum 1sq	Mean Sq	F value	Pr(>F)
Modelo.reg	1	0,000282	0,0002818	2,311	0,13
Residulas	198	0,024145	0,0001219		

Fuente: Elaboración propia

Tabla 14: Test ANOVA 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Modelo.reg	1	0,000282	0,00028179	3,789	0,0544
Bloque.reg	99	0,016782	0,00016952	2,279	0,0000275
Residulas	99	0,007363	0,00007437		

Fuente: Elaboración propia

Las hipótesis del test ANOVA son:

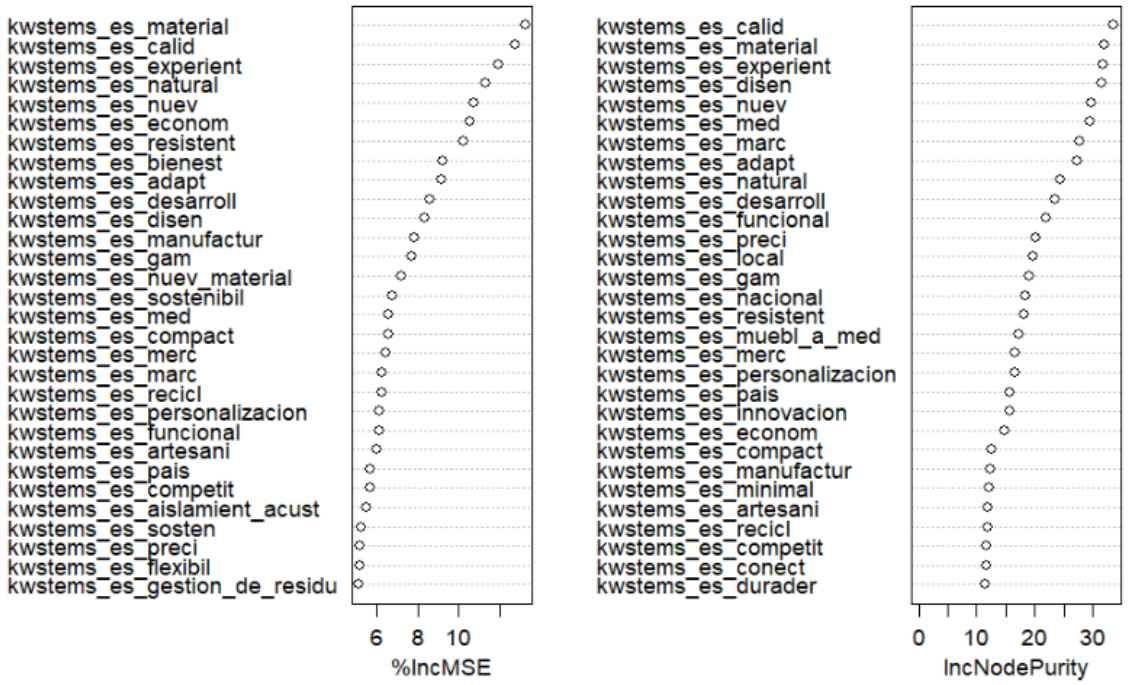
·H0= No existe diferencia significativa entre los dos modelos.

·H1= Existe diferencia significativa entre los dos modelos.

Dado que el p-valor es mayor que 0.05 no se puede rechazar la hipótesis nula, por lo que se puede aceptar que no existen diferencias significativas entre los modelos.

Teniendo esto en cuenta, se utilizará el modelo Random Forest para analizar la estrategia de las empresas más competitivas, ya que a diferencia del SVM este si proporciona una herramienta visual que permita comprobar que variables son las que más contribuyen a la predicción del modelo.

lustración 23: Importancia variables Random Forest 2



Fuente: Elaboración propia

En este caso, atendiendo a los resultados que presenta el gráfico puede apreciarse como las variables que más contribuyen a altos rendimientos en las empresas son ‘material’, ‘calidad’, ‘experimental’, ‘diseño’, ‘natural’.

Teniendo esto en cuenta puede llegarse a la conclusión de que las variables que más contribuyen a un alto rendimiento para las empresas del sector de la fabricación de muebles son en general las relacionadas con la calidad y el diseño.

En este caso, a diferencia de los resultados obtenidos en clasificación la variable marca no presenta una importancia tan relevante, lo que puede indicar que, si bien puede ser un indicador de tamaño, el, de alguna manera explotar una marca conocida por los consumidores, no es necesariamente el punto o la estrategia en la que más deban fijarse aquellas empresas pequeñas que buscan maximizar su rentabilidad.

5. Conclusiones

Una vez comprendido el sector del estudio y analizados los resultados obtenidos es posible establecer conclusiones que resulten de interés.

El objetivo principal del estudio fue el análisis de la competitividad de las empresas del sector de la fabricación de muebles a través de las nuevas tendencias tomando indicadores digitales para medir el impacto de estas.

Esto se logró analizando tanto el tamaño de las empresas como la rentabilidad en función de que estrategia toman estas empresas con relación a las nuevas tendencias.

El estudio además proporcionó algo de información que puede resultar útil para la gestión de estas empresas en relación no solo a su estrategia comercial sino financiera, puesto que como se pudo apreciar en las empresas que se analizaron la deuda no parece ser de utilidad para aumentar el rendimiento de las empresas, y aquellas empresas que necesitan recurrir a esto es debido a una mala situación.

Por otra parte, otra conclusión interesante se generó con la división por grupos del análisis clustering, donde se pudo apreciar como existe un grupo de empresas pequeñas con valores de rentabilidad más elevados incluso que las grandes empresas del sector.

Esto quiere decir por un lado que las empresas más grandes de este sector no están optimizando sus recursos de la misma manera que consiguen hacerlo empresas más pequeñas, lo cual puede deberse en cierta parte al 'acomodamiento' que pueden sufrir las empresas más grandes. Como ya tienen un nombre y un tamaño no realizan el mismo esfuerzo por maximizar sus recursos, sino que simplemente sobreviven con lo que tienen.

Por otra parte, comparando los resultados de los modelos de clasificación y regresión puede apreciarse como si bien es cierto que las empresas grandes comparten una buena parte de la estrategia con las empresas que más rentabilidad obtienen, hay ciertos aspectos a los que prestan posiblemente más importancia de la necesaria.

En este caso se estaría hablando de la marca, lo cual podría confirmar lo mencionado anteriormente sobre el 'acomodamiento'. Aquellas empresas grandes utilizan el hecho de ser conocidas como estrategia de diferenciación. Podría resumirse como que a la empresa con marca X le compran por ser X.

Por esto parece que sí que sería una buena idea para empresas más pequeñas tratar de generar una marca la cual los consumidores asocien, pero sin descuidar el resto de aspectos.

Finalmente, respecto a la estrategia seguida por las empresas más rentables del mercado y por lo tanto más competitivas indica que los aspectos más fundamentales en los que deben enfocarse son aquellos relacionados con la calidad y el diseño, no siendo así para factores como la sostenibilidad, dado que según los datos esta no juega un papel tan importante en la rentabilidad de la empresa.

En conclusión, las empresas del sector de la fabricación de muebles que busquen optimizar su rendimiento deben tener en cuenta las nuevas tendencias, pero no tanto en cuanto a ser sostenibles sino a ofrecer diseños actualizados, 'que estén a la moda'.

Bibliografía

Hütt Herrera, H. (2012). Las redes sociales: Una nueva herramienta de difusión. *Reflexiones*, 1021-1209.

Oyola-García, A. E. (2021). La variable. *Revista del cuerpo médico del HNAAA*.

¿Qué es una huella digital? ¿Cómo puedes protegerla de los hackers? (s.f.). Obtenido de Kaspersky: https://www.kaspersky.es/resource-center/definitions/what-is-a-digital-footprint

Carlowitz, H. C. (1713). *Sylvicultura oeconomica*.

Codina, L. (2004). Posicionamiento Web: Conceptos y Ciclo de Vida. *Hipertext.net*.

Dans, E. (2007). La empresa y la Web 2.0. *Marketing y ventas*, 36-43.

Daries-Ramon, N., Cristóbal-Fransi, E., Martín-Fuentes, E., & Mariné-Roig, E. (2017). Desarrollo de las TIC en el turismo de nieve:Análisis de la presencia en línea de las estaciones de esqui de España y Andorra. *Documents d'Anàlisi Geogràfica*, 399-426.

Debón, A. (26 de Noviembre de 2023). Aprendizaje supervisado: clasificacion. *Inteligencia de negocios*.

Enprenemjunts. (17 de Mayo de 2012). Obtenido de https://www.emprenemjunts.es/?op=13&n=2046

Flores, M. J. (31 de 3 de 2021). *Técnicas Multivariadas con R*. Obtenido de https://bookdown.org/jsalinas/tecnicas_multivariadas/presentacion.html

Gascó, V. (2020). Herramientas de mapas de calor y eye tracking para optimizar ventas. *Sales Layer*.

Hugo Cardenas , F., Jimenez Rosero, C., Holovatyi, M., & Lara Pazos, P. (1 de Enero de 2020). El impacto de las redes sociales en la administración de empresas. *Recimundo*, 173-182.

Integral Innovation Experts. (20 de Agosto de 2019). *Integral Innovation Experts*. Obtenido de https://integralplm.com/blog/2019/08/20/que-es-cad/

Martinez del Rio, F. (29 de 1 de 2024). *Programacion con R*. Obtenido de https://www4.ujaen.es/~fmartin/R/

Mueble de España. (13 de Junio de 2022). Obtenido de https://muebledeespana.com/es/sala-de-prensa/informe-de-comercio-exterior-del-mueble-espanol-2021

Munier. (2005). *Introducción al concepto de sostenibilidad*. Obtenido de https://uapa.cuaieed.unam.mx/sites/default/files/minisite/static/693ee8e8-f02c-43c2-8222-

498e1e8b8814/ConceptoSostenibilidad/index.html#:~:text=La%20sostenibilidad%20se%20refiere%20a,futuras%20(Muiner%2C%202005).

Rocchi, S., & Boada Ortíz, A. (2005). Sostenibilidad, negocios y marca. *poliantea, 2*(4), 37-50.

Salgado, D. (2016). La huella digital. *Indice*, 14-17.

Tarazona Campos, S. (s.f.). Inteligencia de negocios I. *PCA Analisis de Componentes Principales*.

Tarazona Campos, S. (s.f.). Inteligencia de negocios II. *Análisis clustering*.

Treyes, J. (2 de Agosto de 2022). *LInkedin*. Obtenido de Datos Anomalos: https://www.linkedin.com/pulse/datos-an%C3%B3malos-jhon-treyes?originalSubdomain=es#:~:text=Los%20datos%20an%C3%B3malos%20o%20at%C3%ADpicos,que%20se%20considera%20como%20at%C3%ADpico.

Ubeda. (8 de Mayo de 2013). *Oficina de Software y Hardware Libre*. Obtenido de Universidad Miguel Hernandez de Elche: https://oshl.umh.es/2013/05/08/r-project/

Universitat Oberta de Catalunya. (s.f.).

Westreicher, G. (1 de Marzo de 2020). *economipedia*. Obtenido de https://economipedia.com/definiciones/estandar.html

Zimmermann, Y. (2 de Enero de 1988). El 'boom' del diseño. *El País*.

Anexo I. Objetivos de Desarrollo Sostenible

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza**				**X**
ODS 2. **Hambre cero**				**X**
ODS 3. **Salud y bienestar**				**X**
ODS 4. **Educación de calidad**				**X**
ODS 5. **Igualdad de género**				**X**
ODS 6. **Agua limpia y saneamiento**				**X**
ODS 7. **Energía asequible y no contaminante**				**X**
ODS 8. **Trabajo decente y crecimiento económico**	**X**			
ODS 9. **Industria, innovación e infraestructuras**	**X**			
ODS 10. **Reducción de las desigualdades**				**X**
ODS 11. **Ciudades y comunidades sostenibles**		**X**		
ODS 12. **Producción y consumo responsables**			**X**	
ODS 13. **Acción por el clima**		**X**		
ODS 14. **Vida submarina**				**X**
ODS 15. **Vida de ecosistemas terrestres**				**X**
ODS 16. **Paz, justicia e instituciones sólidas**				**X**
ODS 17. **Alianzas para lograr objetivos**				**X**

Fuente: Elaboración propia.

Este Trabajo de fin de grado muestra un cierto grado de implicación con algunos de los ODS.

En primer lugar, estaría el ODS 8, el cual es 'Trabajo decente y crecimiento económico', ya que el principal objetivo del estudio es entender cómo afecta la implementación de las nuevas tendencias a la competitividad de las empresas, buscando así las decisiones o estrategias que deben seguir las empresas de este sector para maximizar su rentabilidad y conseguir de esta manera un mayor crecimiento económico.

Por otra parte, está muy relacionado con el ODS 9, 'Industria, innovación e infraestructuras', ya que un punto clave del trabajo es el estudio de las nuevas tendencias

así como de las estrategias clave que deben seguir las empresas para mejorar su rentabilidad, entre las cuales destacó la innovación.

Como se explicó en el trabajo las empresas españolas de este sector deben enfocar sus esfuerzos y estrategias en diferenciarse por medios como la innovación en diseños.

Posteriormente el trabajo estaría relacionado con los ODS 11 y 13, 'Ciudades y comunidades sostenibles' y 'Acción por el clima', esto es debido a que una de las nuevas tendencias más populares hoy en día es la sostenibilidad, y fue un punto importante de análisis para este trabajo.

Si bien los resultados mostraron que la sostenibilidad no figura entre las estrategias que mejor optimizan la rentabilidad de las empresas, sí que aparece entre las palabras que contribuyen a esto, y teniendo en cuenta la importancia no solo económica sino de responsabilidad social es un hecho que la sostenibilidad es un aspecto fundamental a cuidar por las empresas no solo de este sector sino de cualquier otro.

Finalmente, si bien este trabajo no muestra demasiada relación con el resto de ODS, sí que puede resultar útil para que las empresas del sector se responsabilicen respecto a las cuestiones que recogen los ODS.