UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Análisis de la relación entre la competitividad de las empresas y su presencia online

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Soto Andreu, Sara

Tutor/a: Doménech i de Soria, Josep

CURSO ACADÉMICO: 2022/2023

Resumen

En la actualidad, la transformación digital ha llevado a que el comercio online tenga un peso cada vez mayor en comparación con las tiendas físicas, lo que ha obligado a las empresas a centrarse en su presencia digital para expandir sus ventas y atraer a más clientes. En este contexto, las estrategias de marketing online se han vuelto fundamentales para la captación de clientes y aumentar la competitividad. En este sentido, la huella digital que dejan las empresas en tiempo real puede ser una herramienta muy valiosa para predecir su grado de competitividad y aplicar estrategias de mejora.

El proyecto tiene como objetivo general analizar la correlación entre los indicadores de competitividad y la presencia online de las empresas. Para ello, se utilizarán los indicadores económicos obtenidos del Sistema de Análisis de Balances Ibéricos (SABI), incluyendo el número de empleados, la rentabilidad económica y el inmovilizado intangible. Además, se medirán las variables que miden la huella digital de las empresas a través de la minería de datos, utilizando el texto de la página web para crear un vector de embedding. A través de los métodos de aprendizaje automático, se podrán clasificar los datos en diferentes clusters en función de las similitudes encontradas en los datos. De esta manera, se podrá evaluar la relación entre los indicadores económicos y la presencia online de las empresas y determinar si la huella digital es un indicador válido para medir la competitividad de las mismas.

Los resultados obtenidos en este proyecto sugieren que existe una relación significativa entre la presencia online de las empresas y su competitividad. Los datos analizados se dividen en dos grupos claramente diferenciados: empresas competitivas y menos competitivas, cuya huella digital presenta características distintivas. Los métodos de aprendizaje automático, como el SVM y la regresión logística, resultaron ser eficaces para clasificar los datos según su similitud. Además, se logró obtener una función de regresión lineal que permite predecir el valor de la rentabilidad económica de una empresa en función del texto de su página web. Estos hallazgos indican que la huella digital de una empresa puede ser una herramienta útil para evaluar su competitividad y aplicar estrategias de mejora.

Palabras clave: Empresas, Competitividad, Huella digital, Clustering, Minería de Datos, Embeddings.

Resum

En l'actualitat, la transformació digital ha portat al fet que el comerç en línia tinga un pes cada vegada major en comparació amb les botigues físiques, la qual cosa ha obligat les empreses a centrar-se en la seua presència digital per a expandir les seues vendes i atraure a més clients. En aquest context, les estratègies de màrqueting en línia s'han tornat fonamentals per a la captació de clients i augmentar la competitivitat. En aquest sentit, l'empremta digital que deixen les empreses en temps real pot ser una eina molt valuosa per a predir el seu grau de competitivitat i aplicar estratègies de millora.

El projecte té com a objectiu general analitzar la correlació entre els indicadors de competitivitat i la presència en línia de les empreses. Per a això, s'utilitzaran els indicadors econòmics obtinguts del Sistema d'Anàlisi de Balanços Ibèrics (SABI), incloent-hi el nombre d'empleats, la rendibilitat econòmica i l'immobilitzat intangible. A més, es mesuraran les variables que mesuren l'empremta digital de les empreses a través de la mineria de dades, utilitzant el text de la pàgina web per a crear un vector de *embedding. A través dels mètodes d'aprenentatge automàtic, es podran classificar les dades en diferents clústers en funció de les similituds trobades en les dades. D'aquesta manera, es podrà avaluar la relació entre els indicadors econòmics i la presència en línia de les empreses i determinar si l'empremta digital és un indicador vàlid per a mesurar la competitivitat d'aquestes.

Els resultats obtinguts en aquest projecte suggereixen que existeix una relació significativa entre la presència en línia de les empreses i la seua competitivitat. Les dades analitzades es divideixen en dos grups clarament diferenciats: empreses competitives i menys competitives, l'empremta digital de les quals presenta característiques distintives. Els mètodes d'aprenentatge automàtic, com el *SVM i la regressió logística, van resultar ser eficaces per a classificar les dades segons la seua similitud. A més, es va aconseguir obtindre una funció de regressió lineal que permet predir el valor de la rendibilitat econòmica d'una empresa en funció del text de la seua pàgina web. Aquestes troballes indiquen que l'empremta digital d'una empresa pot ser una eina útil per a avaluar la seua competitivitat i aplicar estratègies de millora.

Paraules clau: Empreses, Competitivitat, Empremta digital, Clustering, Mineria de Dades, Embeddings.

Abstract

Nowadays, digital transformation has led to online commerce having an increasing weight compared to physical stores, which has forced companies to focus on their digital presence to expand their sales and attract more customers. In this context, online marketing strategies have become essential to attract customers and increase competitiveness. In this sense, the digital footprint left by companies in real time can be a very valuable tool to predict their degree of competitiveness and implement improvement strategies.

The general objective of the project is to analyze the correlation between competitiveness indicators and the online presence of companies. For this purpose, economic indicators obtained from the Iberian Balance Sheet Analysis System (SABI) will be used, including the number of employees, economic profitability and intangible fixed assets. In addition, variables that measure the digital footprint of companies will be measured through data mining, using the text of the web page to create an embedding vector. Through machine learning methods, it will be possible to classify the data into different clusters based on the similarities found in the data. In this way, it will be possible to evaluate the relationship between economic indicators and the online presence of companies and determine whether the digital footprint is a valid indicator to measure the competitiveness of companies.

The results obtained in this project suggest that there is a significant relationship between the online presence of companies and their competitiveness. The data analyzed are divided into two clearly differentiated groups: competitive and less competitive companies, whose digital footprint has distinctive characteristics. Machine learning methods, such as SVM and linear regression, proved to be effective in classifying the data according to their similarity. In addition, a logistic regression function was obtained to predict the value of a company's economic profitability based on the text on its website. These findings indicate that a company's digital footprint can be a useful tool for assessing its competitiveness and implementing improvement strategies.

Keywords: Companies, Competitiveness, Digital footprint, Clustering, Data mining, embeddings

Tabla de contenidos

Índice de figuras, tablas y ecuaciones

Índice de figuras

Índice de tablas

Índice de ecuaciones

1 - Introducción

En la era digital en la que nos encontramos, las empresas utilizan sus páginas web como herramienta fundamental para promocionar sus productos o servicios y llegar a un mayor número de clientes potenciales. En este proyecto de ciencia de datos, se explorará la relación entre el contenido textual de las páginas web de diferentes empresas y sus indicadores de competitividad, con el objetivo de descubrir si existe una relación significativa entre el lenguaje utilizado en la web y el rendimiento económico de las empresas. A través del análisis de datos, se buscará identificar patrones y correlaciones que nos permitan entender mejor cómo el contenido de la página web está relacionado con la competitividad de una empresa.

1.1 - Motivación

Después de considerar varias temáticas, me decidí por estudiar la relación entre el texto de la página web de las empresas y sus indicadores de competitividad. Esta elección se basó en mi interés por el análisis de texto y su impacto en la toma de decisiones empresariales. Además, esta temática tiene una gran relevancia en el mundo empresarial, ya que el contenido de la página web es uno de los principales medios de comunicación de una empresa con sus clientes y puede influir significativamente en su percepción y decisión de compra. Analizar la web de las empresas se presenta como una forma económica y rápida de supervisar la situación competitiva de las empresas en un determinado sector.

En cuanto a mis motivaciones personales, el desarrollo de este proyecto me permitió adquirir habilidades técnicas en el análisis de datos, aprendiendo a utilizar herramientas avanzadas de procesamiento de texto, modelos de regresión y clasificación.

Por otro lado, desde un punto de vista profesional, este proyecto me brindó la oportunidad de aplicar mis conocimientos en ciencia de datos a un problema real y relevante para el mundo empresarial. Esto me permitió desarrollar habilidades de resolución de problemas y toma de decisiones basadas en datos, que son muy valoradas en el mercado laboral actual.

Por lo tanto, el objetivo de este proyecto fue explorar la relación entre el contenido de la página web de las empresas y sus indicadores de competitividad, y utilizar esta información para desarrollar modelos predictivos de rentabilidad económica.

1.2 - Objetivos

El objetivo principal de este trabajo es analizar la relación entre la competitividad de las empresas y su presencia online. Para lograr esto, se plantean los siguientes objetivos específicos:

- Recopilar y analizar datos de diferentes empresas, obteniendo información de su sitio web y sus indicadores de competitividad.
- Realizar un preprocesamiento de los datos para eliminar información irrelevante y convertir el texto en variables numéricas utilizables para el análisis.
- Aplicar técnicas de aprendizaje automático para construir modelos que permitan predecir los indicadores de competitividad a partir del contenido del texto de la página web.
- Evaluar la calidad de los modelos mediante la comparación de sus predicciones con los valores reales de los indicadores de competitividad.
- Analizar los resultados obtenidos para determinar si existe una relación significativa entre el contenido del texto de la página web y los indicadores de competitividad.

Con el cumplimiento de estos objetivos, se espera obtener un conocimiento más profundo de la importancia del contenido del texto de la página web en la competitividad de las empresas, lo que podría contribuir a mejorar la estrategia de marketing y comunicación de las mismas.

1.3 - Impacto esperado

El producto final de este trabajo de TFG permitirá a las empresas analizar la relación entre el contenido textual de su página web y sus indicadores de competitividad, lo que les permitirá tomar decisiones más informadas sobre cómo mejorar su presencia en línea y aumentar su competitividad. Esto puede beneficiar a las empresas en términos de crecimiento económico, aumentando su capacidad de generar empleo y contribuyendo al desarrollo sostenible. Además, este trabajo también puede ser útil para los expertos en marketing y publicidad, ya que les brindará información valiosa sobre cómo las palabras y el contenido de un sitio web pueden influir en la percepción de una empresa por parte de sus clientes potenciales.

En términos más amplios, este trabajo puede contribuir a la consecución de los Objetivos de Desarrollo Sostenible de la ONU, en particular el objetivo número 8 (Trabajo decente y crecimiento económico) y el objetivo número 9 (Industria, innovación e infraestructura), ya que busca mejorar la capacidad de las empresas para competir y crecer de manera sostenible.

1.4 - Metodología

La metodología seguida en este proyecto es la Cross-Industry Standard Process for Data Mining (CRISP-DM). Esta metodología consta de 6 fases en las que se trata de guiar a profesionales a través de un proyecto de minería de datos. CRISP-DM se puede adaptar a las necesidades y requisitos específicos de cada proyecto (Shearer, 2000) . Por lo tanto, se siguen las siguientes fases:

- Fase 1: Comprensión del negocio: esta primera fase se centra en comprender los objetivos y requisitos del negocio, estudiando el contexto e identificando las preguntas clave que se debe responder mediante el análisis de datos.
- Fase 2: Comprensión de los datos: esta segunda fase se centra en obtener los datos que se van a tratar así como un análisis preliminar para comprender la calidad y distribución de los mismos.
- Fase 3: Preparación de los datos: esta tercera fase se centra en la preparación de los datos para estar listos para el análisis. Es necesario limpiar y transformar y procesar los datos del lenguaje natural antes de la fase de modelado.
- Fase 4: Modelado: esta cuarta fase se centra en seleccionar las técnicas más adecuadas de aprendizaje automático para modelar los datos y así obtener resultados.
- Fase 5: Evaluación: esta quinta fase se centra en evaluar los resultados obtenidos en cada uno de los resultados del aprendizaje automático, analizando la calidad y rendimiento de los mismos.
- Fase 6: Despliegue: esta sexta fase se centra en implementar el proyecto en el entorno de producción y al seguimiento que debe de tener para su correcto funcionamiento.

1.5 - Estructura

La introducción aborda la motivación del proyecto, los objetivos buscados y el impacto que se espera lograr. Luego, se presenta el marco teórico y el análisis del problema, donde se contextualiza la importancia de estudiar la competitividad y la huella digital de las empresas y se explica la razón de ser del proyecto, su plan de trabajo y presupuesto.

A continuación, se describe la base de datos y las variables estudiadas, detallando cómo se obtuvieron los datos, qué técnicas se aplicaron para procesarlos y se incluye un análisis exploratorio.

En el siguiente apartado, se exponen las técnicas de aprendizaje automático utilizadas y se presentan sus resultados.

Finalmente, se presentan las conclusiones obtenidas a lo largo del proyecto y se mencionan posibles trabajos futuros para el mantenimiento y desarrollo del mismo.

2 - Marco teórico

Este capítulo tiene como objetivo contextualizar el proyecto al presentar los conceptos de competitividad y huella digital, así como establecer la relación que existe entre ellos. Al abordar estos temas, se busca brindar al lector una comprensión completa del contexto en el que se desarrolla el proyecto.

2.1 - Competitividad

La competitividad empresarial se refiere a la capacidad de una empresa para competir con éxito en el mercado. (Porter, 1990). Se deben de tener ciertas habilidades y contar con unas estrategias que permitan a una empresa diferenciarse y destacar entre sus competidores.

La competitividad está directamente relacionada con el crecimiento económico de una empresa. Entre las estrategias más importantes para lograr la competitividad empresarial destacan las siguientes:

Según Porter (2008), una de las estrategias más cruciales para alcanzar la competitividad en la actualidad es la innovación. Este enfoque permite a las empresas generar nuevos productor y servicios, mejorar los existentes y desarrollar procesos de producción más eficientes. Asimismo, Chesbrough (2003) argumenta que la inversión en investigación y desarrollo, así como la capacidad de adaptarse a los cambios del mercado, son elementos fundamentales para fomentar la innovación y mantenerse actualizado en el entorno empresarial. La innovación se ha convertido en el principal impulsor del crecimiento económico, como sostiene Aghion y Howitt (2009), y es esencial para garantizar la competitividad en un mercado en constante evolución.

Otra clave para el éxito es la estrategia empresarial. Toda empresa necesita tener una estrategia clara y definida. La competitividad estratégica se mide a través de indicadores como la participación de mercado, la rentabilidad y la satisfacción del cliente. Por lo tanto, una buena estrategia implica conocer a los clientes, la competencia, el entorno y los recursos de la empresa, y saber cómo utilizarlos de forma efectiva (Porter, 1996). Conocer las necesidades del cliente es el primer paso para poder competir en el mercado. Un contacto constante con los consumidores ayuda a saber sus deseos y opiniones para poder adaptarse de forma efectiva a sus necesidades (Kotler, 2011).

Asimismo, las empresas deben contar con una eficiencia operativa, lo cual implica producir bienes o servicios de forma eficiente y a bajo costo, utilizando procesos y tecnologías de vanguardia y optimizando su cadena de suministro (Rothaermel,

2015). Una buena gestión de la calidad también es fundamental para la competitividad empresarial, ya que implica implementar prácticas y procesos que permitan mejorar la calidad de los productos o servicios ofrecidos por la empresa. Además, una buena gestión de la calidad puede generar satisfacción en los clientes y aumentar la reputación de la empresa (Flynn, 2010).

Por último, en la actualidad, el marketing social y digital se está asentando en todas las empresas y es un punto crucial para la competitividad de estas. Una buena estrategia permite a la empresa llegar a su público objetivo y destacarse en un mercado cada vez más saturado y competitivo (Kotler, 2000). Tanto en el marketing como en todos los sectores es de vital importancia la gestión del talento humano como puente para obtener los mejores resultados posibles. Esto no se consigue sólo teniendo en el equipo a los mejores profesionales, si no, aportando constante formación y generando un ambiente de trabajo en el que el empleado pueda explotar su rendimiento al 100%.

2.2 - Huella digital

La huella digital se compone de la información que se produce y se deja en línea durante el uso de dispositivos digitales y su interacción (Van Dijck, 2014). Esta información puede incluir datos de navegación, publicaciones en redes sociales, correos electrónicos, compras en línea, entre otros aspectos. La huella digital de una empresa abarca desde su página web, redes sociales y blogs, hasta su presencia en directorios en línea y foros de discusión.

Según el informe Digital 2021 de Hootsuite y We Are Social, a enero de 2021, hay 4.660 millones de personas en todo el mundo que utilizan Internet, lo que representa el 59% de la población mundial (Hootsuite & We Are Social, 2021). Debido a esto, cada vez son más las empresas que se suman al mundo digital para llegar a su público objetivo a través de las redes sociales y las páginas web.

Además, según el informe de la Asociación de Empresas de Tecnología de la Información y Comunicación (AETIC) de 2019, el 98,4% de las empresas en España tienen presencia en la web, ya sea a través de un sitio web propio, un perfil en redes sociales, o ambas opciones. Esta cifra muestra la importancia que las empresas le dan a estar presentes en el mundo digital para llegar a sus clientes potenciales y mantener una buena relación con ellos.

La amplitud y variedad de la huella digital de una empresa se convierte en una valiosa herramienta para evaluar su presencia y reputación en el entorno en línea, lo que puede tener un impacto determinante en su éxito a largo plazo.

De acuerdo con Greene y Goggins (2010), la huella digital activa de una empresa se extiende a través de varias plataformas digitales, donde las empresas pueden compartir información y lanzar campañas publicitarias. Las páginas web permiten una interacción más cercana con los clientes y brindan a las empresas la oportunidad de obtener información valiosa sobre los consumidores para desarrollar estrategias de marketing más efectivas. La capacidad de las empresas para gestionar y controlar su huella digital en línea es clave para garantizar una imagen positiva y evitar daños en su reputación.

2.3 - Análisis de la competitividad a través de la huella digital

En este estudio se plantea cómo los sitios web pueden servir como indicadores en tiempo real de las características económicas de las empresas. La información disponible en los sitios web es gratuita y, lo que es aún más crucial, está constantemente actualizada, lo que la convierte en una fuente valiosa para obtener información empresarial. A diferencia de las fuentes de datos tradicionales, el análisis de contenido web requiere un esfuerzo considerable en la recopilación, selección, limpieza y análisis de los datos. Sin embargo, proporciona una perspectiva única e independiente para comprender las empresas (Smith, 2023).

Estudios han demostrado que los cambios en la estructura y el contenido de un sitio web pueden influir en la percepción de los consumidores sobre la calidad de los productos o servicios de una empresa, así como en su confianza en la marca. Por ejemplo, investigaciones realizadas por Li (2019) revelaron que la calidad percibida de un sitio web y la facilidad de uso del mismo estaban positivamente relacionadas con la competitividad de la empresa.

Como Blazquez y Domenech (2018), clasificando el contenido de los sitios web se puede obtener los indicadores online de las organizaciones. Cambios o actualizaciones en las páginas web de las empresas pueden ser indicativos de cambios o variaciones en el riesgo de crédito de dichas empresas, por lo que es conveniente una monitorización basada en la observación continua del comportamiento de las empresas (Blazquez, 2018).

La presencia de información relevante y actualizada en el sitio web de una empresa puede influir en la toma de decisiones de los clientes y su intención de compra. Un estudio realizado por Han (2018) encontró que la disponibilidad de información detallada y precisa en el sitio web de una empresa aumentaba la confianza de los consumidores y su disposición a realizar transacciones con dicha empresa.

Por lo tanto, además de servir como indicadores en tiempo real de las características de las empresas, los cambios en los sitios web pueden tener un impacto significativo en su competitividad. La actualización y mejora constante de un sitio web puede reflejar la capacidad de una empresa para adaptarse a las demandas del mercado y mantenerse al día con las últimas tendencias y tecnologías (Smith, 2023).

3 - Análisis del problema

El objetivo del presente trabajo es analizar si existe una relación entre la presencia online de las empresas y sus indicadores de competitividad. Esta es una cuestión de gran interés para las empresas, ya que la página web es un canal importante para comunicarse con los clientes y para generar nuevas oportunidades de negocio.

Las instituciones públicas y políticas, pueden promover un entorno favorable para el crecimiento empresarial a partir de analizar los sitios web de las empresas ya que pueden identificar necesidades y desafíos. Además, proporcionar información sobre la actividad económica de un área geográfica específica conlleva a identificar sectores en desarrollo y comprender la dinámica empresarial de la región. Por lo tanto, el seguimiento de los cambios en los sitios web de las empresas puede ser útil para evaluar el impacto de las políticas públicas en el tejido empresarial.

Para llevar a cabo este análisis se seguirán las recomendaciones de las asignaturas de proyectos I, II y III. Se realizará una identificación sistemática de las oportunidades de innovación o de negocio que pueden derivarse del estudio de la relación entre el texto de la página web y los indicadores de competitividad de las empresas.

En concreto, se llevará a cabo un análisis del problema identificando los indicadores de competitividad más relevantes para el estudio y se recopilando los textos de las páginas web de las empresas seleccionadas para el análisis. A continuación, se analizará el texto de estas páginas web utilizando técnicas de procesamiento del lenguaje natural y se extraerán características relevantes del mismo. Finalmente, se aplicarán técnicas estadísticas y de aprendizaje automático para analizar la relación entre el texto de la página web y los indicadores de competitividad de las empresas.

3.1 - Plan de trabajo y presupuesto

Es esencial contar con una adecuada organización en cualquier proyecto, ya que solo así se pueden cumplir las fechas límite. Para lograr nuestros objetivos dentro del plazo establecido, es fundamental diseñar un plan de trabajo realista y bien estructurado.

Se establece un plazo de tres meses en los que la planificación es la siguiente:

- Semana 1-2: Investigación preliminar y definición de los objetivos
    - Investigación de fuentes de datos disponibles y posibles técnicas de minería web para la adquisición de los datos que miden la huella digital.
    - Definición de objetivos específicos del proyecto.
- Semana 3-4: Adquisición de los datos
    - Selección de fuentes de datos y técnicas de minería web.
    - Extracción del vector de embeddings y limpieza de los datos
- Semana 5-6: Análisis exploratorio de datos
    - Análisis descriptivo y exploratorio de los datos
    - Identificación de patrones y relaciones entre las variables
- Semana 7-8: Preparación de los datos
    - Selección y preparación de las variables relevantes
    - Creación de las variables adicionales si es necesario
- Semana 9-10: Selección y entrenamiento de modelos
    - Selección y entrenamiento de varios modelos de aprendizaje automático.
    - Evaluación de la eficacia de cada modelo.
- Semana 11-12: Validación y refinamiento del modelo.
    - Validación de los modelos.
    - Refinamiento del modelo en función de los resultados de la validación
- Semana 13: Redacción de la memoria

En función de la planificación establecida se establecen unos costes por hora trabajada. El salario medio de un científico de datos en España es de 35000€/anuales lo que correspondería a 18,5€/hora trabajada. En la siguiente tabla observamos el coste del proyecto:

Tabla 1: Duración y presupuesto de las actividades a realizar en el proyecto.

Tarea	**Tiempo**	**Coste**
Investigación preliminar	**45 horas**	**832,5 €**
Documentación y obtención de datos	5 horas	92,5 €
Documentación OpenAI	10 horas	185 €
Documentación minería web	10 horas	185 €
Documentación procesado de textos	10 horas	185 €
Documentación métodos aprendizaje automático	10 horas	185 €

Adquisición de los datos	**75 horas**	**1387,5 €**
Selección variables en SABI	5 horas	92,5 €
Obtención texto de las webs	20 horas	370 €
Limpieza del texto	20 horas	370 €
Extracción vector embeddings	30 horas	555 €
Análisis exploratorio de datos	**45 horas**	**832,5 €**
Análisis descriptivo	30 horas	555 €
Identificación de patrones	15 horas	277,5 €
Preparación de los datos	**35 horas**	**647,5 €**
Preparación de las variables	30 horas	555 €
Creación variables adicionales	5 horas	92,5 €
Selección y entrenamiento de modelos	**30 horas**	**555 €**
Selección y entrenamiento de modelos de AA	20 horas	370 €
Evaluación de los modelos	10 horas	185 €
Validación y refinamiento del modelo	**30 horas**	**555 €**
Redacción de la memoria	**40 horas**	**740 €**
TOTAL	**300**	**5550 €**

Fuente: Elaboración propia

A parte de considerar el coste de mano de obra del científico de datos, se debe de tener en cuenta el coste asociado del uso de la API de OpenAI, como parte integral de los requisitos del proyecto. Aunque en este caso se empleó la licencia gratuita proporcionada por la Universidad Politécnica de Valencia, es un coste que debe ser asumido. Teniendo en cuenta el uso de esta cuenta gratuita, a razón de 0.0001 dólares por cada mil palabras, el gasto habría ascendido a 5 euros.

Además, se debe incluir en el presupuesto el desgaste del ordenador. Este coste se estima tomando en cuenta que el valor promedio de un portátil es de 1500€ y su vida útil oscila entre 3 y 5 años, lo que representa un coste de uso de 80€ para la duración del proyecto. Cabe destacar que no se incluye en el presupuesto el costo del software, ya que se utilizarán programas gratuitos.

En este proyecto, no se incluyen los gastos directos derivados de las empresas, como el alquiler de oficinas, los suministros o la conexión a internet, entre otros. Además, es importante tener en cuenta el uso de la licencia de SABI para obtener los datos de competitividad de las empresas. Sin embargo, dado que la universidad cuenta con una licencia gratuita, no es necesario considerar el costo que normalmente estaría asociado a esta herramienta.

En conclusión, el presupuesto total del proyecto asciende a 5635€, incluyendo la mano de obra del científico de datos y el costo de hardware.

3.2 - Análisis del marco legal y ético

En todo proyecto de análisis de datos se debe garantizar un uso adecuado y legal de los datos. Al manipular datos delicados, es estrictamente necesario considerar diversos aspectos relacionados con la recopilación, análisis y uso de datos empresariales. Es importante tener en cuenta que los datos no son sólo números, sino que representan a organizaciones reales, por lo que se deben respetar sus derechos.

La privacidad y el consentimiento siguen siendo cuestiones importantes a considerar. La obtención de datos de este proyecto viene de dos fuentes distintas. En primer lugar, se accede a un registro público en el cual las empresas están obligadas a depositar sus cuentas. Estos datos se obtienen ya que la Universidad Politécnica de Valencia cuenta con una licencia para utilizarlos. En segundo lugar, se obtiene información de las páginas web de las empresas, la cual es proporcionada de manera voluntaria. Es crucial que estas empresas sean responsables y conscientes de la información que están brindando.

Por lo tanto, es importante que se realice un uso responsable de los datos empresariales, a pesar de que la información sea proporcionada de forma obligatoria o voluntaria por parte de las empresas.

4 - Preparación y comprensión de datos

En este apartado se explica de qué manera se han obtenido los datos, que variables se han decidido estudiar y el procesado previo que se aplica a la muestra antes de utilizar los datos en como entrada a los métodos de aprendizaje automático. Todo este análisis se ha realizado con Python.

4.1 - Obtención de la base de datos

La base de datos consta de 4006 observaciones y 23 variables entre las que encontramos el nombre, la web, indicadores de competitividad e indicadores que miden la presencia online. Para la obtención de esta, se decide filtrar por empresas activas (es decir, que no están en liquidación o en suspensión de pagos) de la industria agroalimentaria con sede en España y con menos de 250 trabajadores.

4.1.1 - Indicadores de competitividad

Estas variables se obtienen utilizando el Sistema de Análisis de Balances Ibéricos o más conocido como SABI. Un software de análisis financiero que recoge información general y de las cuentas anuales de más de 2.000.000 de empresas españolas de una forma rápida y sencilla. Esta base de datos permite hacer comparaciones entre empresas, investigar a la competencia del sector u obtener el estado de las cuentas anuales (CEA, 2021).

En este estudio se pretenden tomar variables que sean representativas en cuanto a la medición de la competitividad. En la elección de estas hubiese sido interesante estudiar variables internas de la empresa pero, como nos ceñimos a información pública, las variables que se estudiarán para medir la competitividad son las siguientes:

a. **Valor del inmovilizado intangible (inmovilizado_intangible_año)**

Esta cifra recoge todos aquellos activos que no tienen apariencia física, que son susceptibles de valoración económica y que no se pueden liquidar en un periodo inferior a un año. El área encargada de evaluar los activos intangibles y los resultados que provoca en la empresa es el área de tecnología (Álvarez, 2005).

Teniendo en cuenta el Plan General de Contabilidad de 2007, estaríamos hablando de los gastos de inversión y desarrollo, la propiedad industrial o software.

Se ha considerado con especial importancia estudiar esta variable porque a pesar de no tener soporte material, es capaz de generar gran valor para la empresa. Aporta grandes beneficios a la empresa como incremento en las ventas, aumento en la productividad o reducción de los costes. La inversión en activos intangibles conlleva a un mayor riesgo, y también potencialmente por ello, una mayor rentabilidad (CEDE, 2009). Por tanto, cuanto mayor sea esta cifra, más competitividad tendrá la empresa.

Hay una variable de tipo continua para cada uno de los siguientes años: 2017, 2018, 2019, 2020 y 2021.

b. Número de empleados (numero_empleados_año):

El número de empleados de una empresa tiene una vital importancia ya que representa la magnitud de esta. Esta cifra se asocia a la capacidad de producción por lo que, cuantos más empleados tenga la organización, más productiva será y en consecuencia más competitiva.

Hay una variable de tipo discreta para cada uno de los siguientes años: 2017, 2018, 2019, 2020 y 2021.

c. Rentabilidad económica (rentabilidad_economica_año):

La rentabilidad económica es la ganancia que han dejado las inversiones efectuadas por una compañía. Es un indicador de la buena gestión administrativa de cualquier empresa. Para aumentar este valor las organizaciones deberían aumentar los ingresos y/o reducir los costes en la compañía (Braeley, 2017).

Conocer la rentabilidad anual ayuda a saber si se han tomado buenas decisiones o no en materia financiera. De igual manera que las otras variables seleccionadas para el estudio, una organización será considerada más competitiva cuanto mayor sea su porcentaje de rentabilidad económica.

Hay una variable de tipo continua para cada uno de los siguientes años: 2017, 2018, 2019, 2020 y 2021.

4.1.2 - Indicadores de la huella digital

En este apartado se explicarán las nuevas variables que se han decidido crear para almacenar la información de la página web de cada una de las empresas. Entre la información que se desea obtener se encuentra el texto de la web, el número de imágenes, de videos, de enlaces o de formularios que pueda tener la misma así como los topics de lo que trata el texto.

a. Variables de contenido web.

i. **Vector de embedding (v_embedding):**

Para procesar el texto de la página web se decide obtener un vector numérico que recogerá toda la información. Para generar este vector es necesario que el texto haya sufrido un procesamiento previo. Obtener un formato operativo que permita analizar y aprovechar la información del texto sería el objetivo de esta variable.

La API de OpenAI nos permitirá llevar a cabo este ejercicio. OpenAI es una empresa de investigación de inteligencia artificial (está detrás del revolucionario ChatGPT) y se enfoca en la exploración, el desarrollo y la implementación de tecnologías de IA avanzadas para resolver problemas complejos ("Economia 3", 2021) . Una de las funciones que tiene es poder generar vectores de embedding en forma de lista de números reales que miden la relación entre las palabras de la cadena de texto (OpenAI, 2022).

Se utiliza el modelo ‘text-embedding-ada-002’ ya que es la versión más reciente y por lo tanto la más óptima ya que permite convertir textos de hasta 8191 tokens en un vector de 1536 valores de dimensión.

ii. **Palabras clave del texto de la web (top10):**

Variable en forma de vector que almacena las 10 palabras más comunes en la web de cada una de las empresas estudiadas.

b. Variables de presencia web.

i. **Número de imágenes (num_imagenes):**

Variable discreta que representa el número de imágenes que tiene la página web de la empresa a la fecha de recopilación de los datos. Este dato se obtiene a partir de la etiqueta ‘<img>’ del código HTML.

ii. **Número de videos (num_videos):**

Variable discreta que representa el número de videos que tiene la página web de la empresa a la fecha de recopilación de los datos. Este dato se obtiene a partir de la etiqueta <video> del código HTML.

iii. **Número de formularios (num_formularios):**

Variable discreta que representa el número de formularios que tiene la página web de la empresa a la fecha de recopilación de los datos. Este dato se obtiene a partir de la etiqueta <form> del código HTML.

iv. **Número de enlaces (num_enlaces):**

Variable discreta que representa el número de enlaces que tiene la página web de la empresa a la fecha de recopilación de los datos. Este dato se obtiene a partir de la etiqueta <a> del código HTML.

4.2 - Procesado de datos

Una vez se ha obtenido la base de datos con todas las empresas a analizar procedemos a realizar el procesamiento de datos de la misma. En este apartado se tiene el objetivo de preparar la base de datos para próximamente aplicar los modelos que nos ayudarán a analizar que de manera influye la información que se proporciona en la página web en la competitividad online de las empresas.

El hecho de que varias empresas utilicen la misma página web puede suponer un problema. Hasta el momento tenemos 4006 empresas diferentes por lo que se comprueba cuántas de estas comparten URL para posteriormente, decidir de qué manera tratar estos casos específicos o si hay que tomar la decisión de dejar de tener en cuenta estas organizaciones.

Tras el análisis, obtenemos que 91 empresas comparten página web con otras. El tamaño de la muestra con la que estamos tratando se considera lo suficientemente grande como para que no representen un problema significativo.

Otro de los puntos a destacar en este procesado de los datos es la forma en la que viene la URL de cada página web. Entre las diferentes formas que se encuentran podemos ver los siguientes tres modelos:

- www.1989sit.com : no tiene prefijo
- http://www.accesoaltura.es : su prefijo es ‘http://’
- https://agirrelekue.com/ : su prefijo es ‘https://’

Para que sea más sencillo acceder a ellas se decide que ninguna de las webs tenga prefijo y que por lo tanto todas presenten el mismo formato. En el caso de los tres ejemplos mencionados quedarían de la siguiente manera: ‘www.1989sit.com’, ‘www.accesoaltura.es’, ‘agirrelekue.com/’.

Una de las técnicas más conocidas para la extracción del texto de una URL es el web scraping. De esta manera se puede acceder al contenido y extraer el código HTML. Una de las mejores bibliotecas de Python para la extracción de datos no estructurados y la que se utilizará para realizar esta tarea es Beautiful Soup, librería que permite la extracción del contenido de una URL y lo transforma en una lista, matriz o diccionario.

Acceder a las diferentes webs de la muestra sufre sus complicaciones principalmente porque no todas las URL de la muestra exportada de SABI están actualizadas o continúan existiendo a fecha de recopilación de los datos. Se descartan 210 observaciones por los siguientes motivos:

- Error de conexión (Connection Error): este error ocurre cuando no se encuentra la página web debido a que no existe, no está disponible por problemas técnicos o requiere credenciales de autenticación de usuario para acceder a ellas.
- Error de certificado SSL (SSLError): este error ocurre cuando existe algún problema con el certificado, configuración de seguridad SSL del cliente, es decir, hay un error en la capa de seguridad SSL al realizar una solicitud HTTPS.
- Demasiados redireccionamientos (TooManyRedirects): este error ocurre cuando se supera el límite de redirecciones permitido por el cliente cuando se realiza una solicitud HTTP. Este problema se produce cuando una URL redirige a otra URL, que a su vez redirige a otra, y así sucesivamente en un bucle sin fin.
- Error de decodificación del contenido (ContentDecodingError): este error ocurre cuando no se puede decodificar correctamente el contenido de respuesta HTTP debido a contenido no compatible o problemas en la transferencia de los datos de la respuesta.
- Error de decodificación por trozos (ChunkedEncodingError): este error ocurre cuando la codificación ‘chunked’ utilizada por el servidor no puede decodificar o manejar correctamente cuando se está descargando contenido a través de una conexión HTTP.
- Error de tiempo de espera (Timeout): este error ocurre una conexión lenta, un servidor ocupado o inactivo, o un tiempo de espera de solicitud demasiado corto.

Ya que cada una de nuestras empresas tiene una página web diferente y por tanto, la estructura del código HTML varía demasiado entre unas y otras, sólo se tomará el contenido de texto de las siguientes etiquetas:

- Head: etiqueta principal que incluye metadatos; atributos que no se muestran al usuario, solo describen referencias de la página web, por ejemplo el título.

- h1, h2, h3, h4, h5, h6: representan las etiquetas del título
- p: indica la apertura y cierre de un párrafo.

Una vez se haya obtenido, se procede a tratar el texto para que su formato sea apto para la generación del vector. Entre estas transformaciones encontramos eliminar los saltos de línea, las tildes, los números y carácteres especiales y eliminar los espacios en blanco.

Por otro lado, al tener datos para 5 años diferentes en las variables que miden la competitividad de las organizaciones, se decide que las variables a estudiar sean la media de todos estos años, por lo tanto, las variables resultantes son las siguientes: media del inmovilizado intangible, media del número de empleados, media de la rentabilidad económica de cada una de las empresas.

La distribución que siguen los datos es la siguiente:

Tabla 2: Distribución de los valores medios de las variables que miden la competitividad.

Variable	Total	Media	Mínimo	Máximo
media_rentabilidad_economica (%)	3373	3.96	-106.58	66.70
media_inmovilizado_intangible (miles de €)	1644	250.60	-28.83	51874.60
media_numero_empleados	3179	30.27	1.00	384.80

Fuente: elaboración propia.

Imputación de los datos

Si nos centramos en los resultados de la tabla anterior podemos observar como la variable del inmovilizado intangible presenta 2061 valores faltantes, lo que puede afectar significativamente al análisis y reducir la precisión de los resultados. En este caso, resulta conveniente realizar una imputación de valores para completar los datos faltantes y evitar posibles sesgos o errores en el análisis. La imputación de valores puede ayudar a mejorar la calidad de los datos y a obtener resultados más precisos y fiables.

Para esta ocasión, se ha decidido utilizar una imputación de datos por la media (MDA). Este método se basa en la descomposición de la matriz de covarianza de los datos y la selección de variables predictoras altamente correlacionadas para imputar los valores faltantes (Li et al., 2016).

Normalización de los datos

Una vez los datos han sido imputados y por lo tanto, la muestra no presenta valores faltantes se decide normalizar los datos ya que las variables no están medidas en las mismas unidades. Mediante la normalización se escalan los valores de las diferentes variables a un rango similar, evitando que una variable domine sobre las demás en el análisis y sobre todo se consigue mejorar la eficacia de ciertos algoritmos de aprendizaje automático. En el presente proyecto se normalizan los datos con el método min-max, que escala los valores a un rango de 0 a 1 (James, 2013).

Tras este proceso, las variables media del inmovilizado intangible, media del número de empleados y media de la rentabilidad económica están imputadas y normalizadas.

Finalmente, se obtiene una muestra de 3705 empresas y 26 columnas.

5 - Metodología

5.1 - Balanceo de datos

El balanceo de datos es una técnica que aborda problemas de desequilibrio de clases en conjuntos de datos, es decir, una clase de interés puede tener muchas menos muestras que las otras clases, lo que puede llevar a un modelo a tener un rendimiento deficiente en la clasificación de esta clase minoritaria (He, 2009).

El balanceo de datos ayuda a solventar este problema creando una distribución más equilibrada de las muestras en cada clase con el fin de obtener modelos más precisos y robustos en la tarea de clasificación.

Algoritmo SMOTE

El algoritmo SMOTE es una técnica utilizada para abordar el problema de desequilibrio de clases en los datos de entrenamiento en el aprendizaje automático.

Este algoritmo se encarga de generar nuevos ejemplos sintéticos de la clase minoritaria a partir de los ejemplos existentes mediante la combinación de diferentes ejemplos de la clase minoritaria.

El proceso del algoritmo SMOTE es el siguiente: primero, se elige un ejemplo al azar de la clase minoritaria y se buscan los k ejemplos más cercanos a este en la clase minoritaria. Luego, se generan ejemplos sintéticos para cada uno de estos ejemplos cercanos, eligiendo un punto al azar en la línea que conecta el ejemplo seleccionado y su vecino y creando un nuevo ejemplo en esa posición. Este proceso se repite varias veces hasta alcanzar el equilibrio deseado entre las clases (Chawla, 2002).

5.2 - Aprendizaje no supervisado

El proyecto se enfoca en el análisis de la relación de la presencia en línea de las empresas y su nivel de competitividad. Con el propósito de alcanzar este objetivo, se emplea un enfoque basado en el uso de vectores de embedding, para explorar la capacidad predictiva de los mismos en la determinación del grado de competitividad de las empresas (Gulati, 2015).

A continuación, se presenta el procedimiento llevado a cabo durante esta sección del proyecto:

Análisis Clustering

El análisis de clusters o conglomerados es una técnica de aprendizaje automático que se utiliza para agrupar las observaciones en clusters, de tal forma que las observaciones dentro de cada clúster sean más similares entre sí que con aquellas en otros clusters. El algoritmo de clustering K-means se ha empleado en la presente investigación, dado que permite agrupar las observaciones de manera eficiente, sin requerir etiquetas previas en los datos. En esencia, este método de clustering busca encontrar patrones y estructuras dentro del conjunto de datos, para agrupar las observaciones que presentan similitudes en cuanto a sus características y atributos.

Algoritmo k-means

El objetivo principal del algoritmo K-means es la partición de los datos en k clusters, con la finalidad de que los datos dentro de cada clúster sean similares y los datos entre clusters sean lo más distintos posible. Para lograr este objetivo, el algoritmo utiliza una estrategia de optimización iterativa que comienza con la asignación de un conjunto aleatorio de centroides iniciales, que se utilizan como punto de partida para la creación de cada clúster. Posteriormente, se llevan a cabo cálculos iterativos para mejorar la posición de los centroides, con el fin de minimizar la distancia intra-clúster y maximizar la distancia inter-clúster. Para determinar qué agrupamiento es el más adecuado y recalibrar los centroides iterativamente, se utiliza una función de distancia que mide la similitud entre dos puntos, x e y.

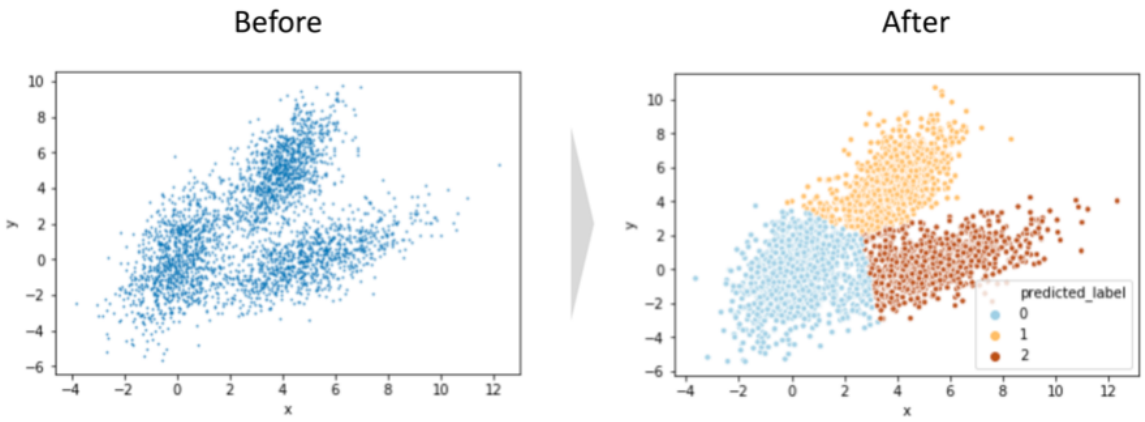


Figura 1: Proceso de creación de clusters con el algoritmo k-means (Hiros G, 2022).

Selección y validación del número de clusters

Los algoritmos empleados en este proyecto no cuentan con la capacidad de determinar automáticamente el número de clusters óptimo, por lo que es necesario especificar este valor durante el proceso de entrenamiento del modelo. Para abordar este desafío, se han utilizado cuatro indicadores distintos para evaluar, seleccionar y validar el número k de clusters más adecuado para el algoritmo.

1. Método del codo

La idea principal detrás de este método es identificar la "curva del codo" en un gráfico que muestra la variación de la suma de los cuadrados de las distancias intra-clúster en función del número de clusters.

En concreto, el método del codo implica graficar la suma de los cuadrados de las distancias intra-clúster en función del número de clústers utilizados en el modelo. A medida que el número de clústers aumenta, la suma de los cuadrados de las distancias intra-clúster tiende a disminuir. Sin embargo, a partir de cierto punto, los beneficios de agregar más clusters disminuyen significativamente, lo que se refleja en la "curva del codo" en el gráfico.

El número óptimo de clusters se selecciona en el punto en el que la adición de un clúster adicional no proporciona un beneficio significativo en la reducción de la suma de los cuadrados de las distancias intra-clúster.

2. Índice Davies-Bouldin

Este índice se basa en la distancia media entre los centroides de los clusters y la dispersión dentro de cada clúster. El objetivo del índice Davies-Bouldin es encontrar el número óptimo de clusters que maximice la distancia entre los clusters y minimice la dispersión dentro de cada clúster. El valor del índice Davies-Bouldin varía de 0 a infinito, siendo 0 el mejor resultado posible.

3. Análisis de la silueta

El objetivo del análisis de la silueta es determinar qué tan bien se ajusta cada objeto a su propio clúster y qué tan separado está su clúster de los demás clusters.La silueta de un objeto se calcula como la diferencia entre la distancia media del objeto a los otros objetos en su propio clúster y la distancia media del objeto a los objetos en el clúster más cercano (distinto del propio clúster). Si la silueta de un objeto es alta, significa que ese objeto está bien clasificado en su propio clúster y que está alejado de otros clusters. Si la silueta de un objeto es baja o negativa, significa que ese objeto podría estar mejor clasificado en otro clúster.

4. Estadístico de Gap

El estadístico de Gap compara la dispersión intra-clúster con la dispersión inter-clúster para diferentes valores de k (número de clusters).

El proceso comienza generando un número determinado de datos aleatorios a partir del conjunto de datos original y ajustando el modelo de clustering correspondiente. Se repite este proceso varias veces y se compara la distribución de la dispersión intra-clúster obtenida a partir de los datos aleatorios con la obtenida a partir del conjunto de datos original. El número óptimo de clusters se determina en el punto en el que el estadístico de Gap alcanza su valor máximo.

Test de Kruskal-Wallis para comparación de medias

En este estudio se ha empleado el test de Kruskal-Wallis para determinar si existen diferencias estadísticamente significativas entre las medias de las variables en los distintos clusters. Esta técnica estadística se utiliza para evaluar la hipótesis de que las medias de dos grupos son equivalentes. El procedimiento implica comparar el valor del estadístico de la prueba con el valor de la distribución chi-cuadrado. El estadístico de la prueba se calcula a partir de los rangos promedios por grupos y el tamaño de cada grupo. Concretamente, el estadístico de prueba sigue una distribución chi-cuadrado con k-1 grados de libertad, donde k representa el número de grupos. Si el valor del estadístico de prueba supera el valor crítico, se rechaza la hipótesis nula y se concluye que los grupos difieren significativamente entre sí (Conover, 2012).

Análisis de componentes principales

Esta técnica multivariante reduce la dimensionalidad de un conjunto de datos hasta encontrar una dirección de proyección en la que la varianza sea máxima. El objetivo es poder mantener la mayor cantidad de información posible en un espacio de menor dimensión llamado espacio de componentes principales (Esbensen, 2009).

Este método calcula una matriz de covarianzas, a partir de unos datos estandarizados, que muestra las relaciones lineales entre las variables originales. Más tarde, se calculan las direcciones de mayor variabilidad en los datos llamados vectores propios y los valores propios que corresponden a la cantidad de variabilidad explicada por cada dirección. Estos vectores propios son ordenados en función de sus valores propios y, se seleccionan los primeros componentes principales capaces de explicar la mayor variabilidad en los datos. Finalmente se generan las componentes principales, nuevas variables que proyectan los datos originales en el espacio de componentes principales utilizando los vectores propios seleccionados.

Por lo tanto, el método PCA busca una dirección de proyección en la que la variabilidad de la variable latente sea máxima, a partir de una matriz de datos con dimensiones NxK (N corresponde al número de observaciones y K al número de variables explicativas).

5.3 - Aprendizaje supervisado

Regresión logística

La regresión logística es una técnica estadística utilizada para modelar la relación entre una variable dependiente binaria y una o más variables independientes continuas o categóricas.

Este modelo se basa en la transformación logística de la relación lineal entre las variables predictoras y la variable dependiente binaria. La función logística utilizada es una curva en forma de "S" que representa la probabilidad de pertenecer a una de las dos clases binarias como una función de las variables predictoras, asignando un valor entre 0 y 1.

Una vez que se ha ajustado el modelo de regresión logística, se pueden hacer predicciones sobre la probabilidad de pertenecer a cada una de las dos clases binarias para nuevos datos. La predicción se realiza a partir de los valores de las variables predictoras conocidas y los coeficientes estimados en el modelo (Berlanga Silvestre & Vilà-Baños, 2014).

Máquinas de vectores soporte

El objetivo de las máquinas de vectores soporte es encontrar el hiperplano que mejor separa dos clases en un espacio de características de alta dimensionalidad. El hiperplano que se busca es aquel que maximiza la distancia entre los puntos de las dos clases más cercanos a él. El margen entre el hiperplano y los vectores de soporte es el más amplio posible, lo que hace que el modelo tenga una buena capacidad de generalización (Suárez, 2014). Estos puntos que separan el hiperplano se conocen como vectores de soporte.

Cabe destacar que este método utiliza un parámetro de regularización donde, en caso de empate y que el algoritmo no se decida entre las clases a predecir, se penalizan los datos mal clasificados.

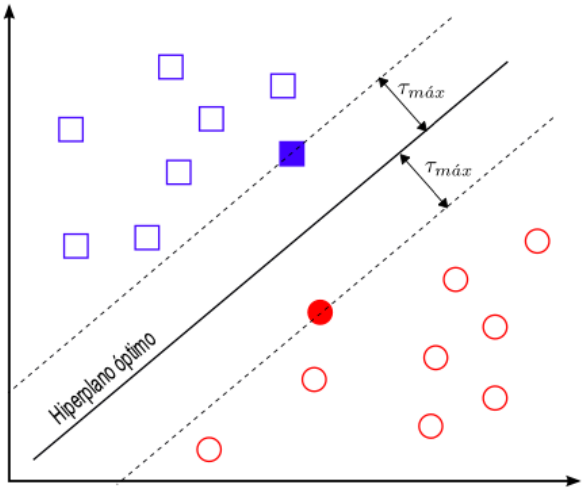


Figura 2: Representación del hiperplano de separación óptima y su margen asociado máximo (Suárez, 2014).

Árboles de decisión

Los árboles de decisión representan una forma de visualizar decisiones en el que cada nodo interno del árbol representa una pregunta sobre una característica del conjunto de datos, y cada rama que sale del nodo representa una posible respuesta a la pregunta. El algoritmo busca dividir el conjunto de datos en subconjuntos más pequeños y homogéneos, hasta que se alcanza un punto en el que las instancias de cada subconjunto son lo suficientemente similares para ser clasificadas en la misma categoría.

El proceso de clasificación o predicción comienza en la raíz del árbol y se mueve hacia abajo siguiendo las ramas correspondientes a las características de la nueva instancia hasta llegar a una hoja del árbol, que representa una decisión sobre la categoría a la que pertenece la instancia o el valor numérico que se predice.

Los árboles de decisión tienen la ventaja de ser fáciles de entender e interpretar, ya que la estructura del árbol es muy intuitiva. Sin embargo, pueden ser sensibles a pequeñas variaciones en los datos y tienden a sobre ajustarse si se les permite crecer demasiado (Breiman, 1984).

Random Forest

El algoritmo de Random Forest se basa en la construcción de múltiples árboles de decisión aleatorios, donde cada árbol es entrenado con una muestra aleatoria de datos y un subconjunto aleatorio de características (Breiman, L. 2001).

Se caracteriza por tener dos fases, la fase de entrenamiento donde se construye un conjunto de árboles de decisión a partir de diferentes muestras de datos y características aleatorias. En cada árbol de decisión, se selecciona una muestra aleatoria de datos con reemplazo y se construye el árbol con una selección aleatoria de características. De esta manera, cada árbol se entrena en una muestra diferente de datos y características. Y una fase de predicción en la que cada árbol del bosque aleatorio realiza una predicción individual y luego se calcula la predicción final a partir de la combinación de todas las predicciones individuales. En el caso de problemas de clasificación se utiliza una votación mayoritaria entre los árboles.

Redes Neuronales

Este algoritmo intenta imitar el funcionamiento del cerebro humano. La estructura se organiza en capas formadas por un gran número de unidades de procesamiento llamadas neuronas.

La primera capa recibe la entrada al modelo, realiza una operación matemática y produce una salida que se transmite a la siguiente capa, así sucesivamente hasta llegar a la última capa que produce la salida. El proceso de entrenamiento de la red neuronal ajusta los pesos y sesgos de las conexiones entre las neuronas para minimizar el error entre la salida producida por la red y la salida esperada.

Una vez entrenada la red neuronal, se puede utilizar para hacer predicciones sobre nuevos datos. La red recibe la entrada y la procesa a través de sus capas, produciendo una salida que se interpreta como la predicción del modelo (Goodfellow, 2016).

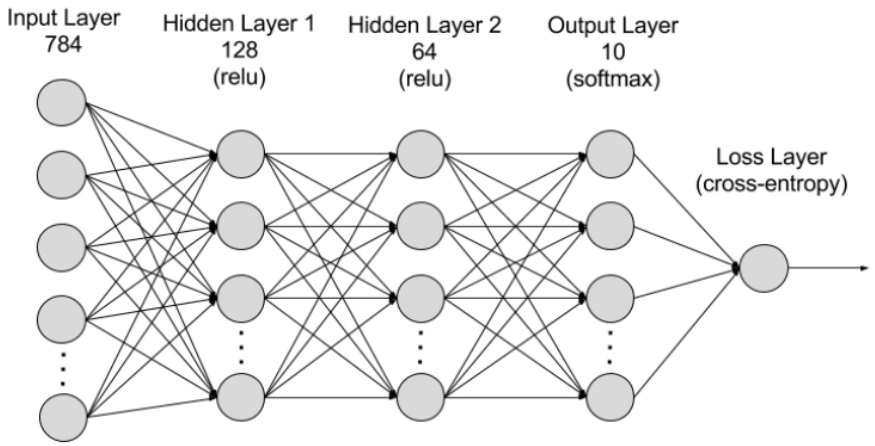


Figura 3: Funcionamiento de una red neuronal. (Amazon Web Services, s.f).

5.4 - Evaluación de modelos de clasificación

Accuracy

El accuracy es un valor que determina la calidad de un modelo de clasificación. Esta medida corresponde a la proporción de casos que el modelo clasifica correctamente.

Para calcular este valor se comparan las predicciones del modelo con los valores reales de la variable objetivo en el conjunto de datos de prueba. El accuracy se calcula dividiendo el número de predicciones correctas entre el número total de predicciones.

Matriz de confusión

La matriz de confusión es una matriz de orden nxn (donde n es el número de clases). Las columnas de esta tabla se refieren a las clases reales y las filas a las predicciones devueltas por el modelo.

Es una forma visual de representar el número de observaciones que han sido predichas de manera correcta e incorrecta a través de cuatro apartados diferentes:

- Verdaderos positivos o sensibilidad (TP)
- Falsos positivos (FP)
- Verdaderos negativos o especificidad (TN)
- Falsos negativos (FN)

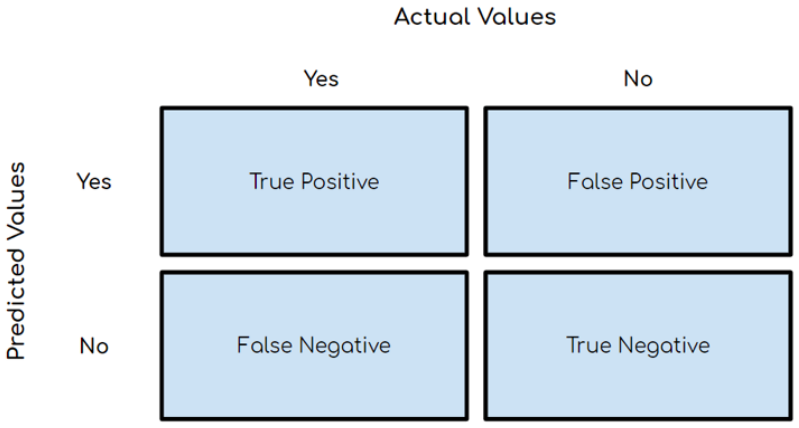


Figura 4: Matriz de confusión (Datasource.ai. , 2022).

A partir de esta matriz obtenemos las medidas de sensibilidad y especificidad.

- Sensibilidad: capacidad del modelo para identificar correctamente los casos positivos, es decir, aquellos que pertenecen a la clase que se está buscando predecir.

$$Sensibilidad = \frac{VP}{(VP + FN)}$$

Ecuación 1: Fórmula sensibilidad

- Especificidad: capacidad del modelo para identificar correctamente los casos negativos, es decir, aquellos que no pertenecen a la clase que se está buscando predecir.

$$Especificidad = \frac{VN}{VN + FP}$$

Ecuación 2: Fórmula especificidad

Curvas ROC y valor AUC

El gráfico de la curva ROC relaciona la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR). En otras palabras, la curva ROC muestra cómo cambia la sensibilidad y la especificidad del modelo a medida que se ajusta el umbral de clasificación. Un modelo con una curva ROC que se acerca a la esquina superior izquierda del gráfico se considera mejor, ya que indica una alta sensibilidad y una baja tasa de falsos positivos.

Además de la curva ROC, también se utiliza el área bajo la curva (AUC, del inglés Area Under the Curve) como medida del rendimiento del modelo. El valor AUC varía entre 0 y 1, donde cuanto más alto el valor mejor será la clasificación.

6 - Conocimientos extraídos y evaluación de los modelos

En este apartado, se presentan los conocimientos extraídos y evaluación de los modelos de aprendizaje automático que se han utilizado para extraer conocimiento de los datos recopilados en el estudio. Primero se muestra un análisis descriptivo de los datos, seguido de la evaluación de los modelos y finalmente los resultados de una función de regresión logística.

6.1 - Análisis descriptivo

El análisis descriptivo tiene como objetivo obtener una visión global de los datos con los que se está trabajando. Lo más importante en este apartado es saber qué tipo de empresas estamos estudiando, ver las parecidos entre estas y analizar las variables para cada tipo de ellas.

Se pretende examinar los indicadores de competitividad de las empresas en relación a las similitudes que estas puedan tener en su página web. Para ello, se utiliza el método de clustering, una técnica de aprendizaje no supervisado que se utiliza para agrupar observaciones en grupos, en este caso el vector de embeddings, con la intención de descubrir patrones ocultos en los datos.

El proceso de clustering implica la selección de un algoritmo adecuado, la elección de la medida de distancia, la determinación del número óptimo de clúster y la interpretación y visualización de los resultados del clustering.

Uno de los principales problemas a la hora de abordar este ejercicio es la forma de tratar el vector de embedding para que sea óptimo para el análisis. Una alta dimensionalidad puede aumentar la complejidad del análisis aportando información irrelevante o aumentando el costo computacional y por lo tanto afectando negativamente al rendimiento.

Para obtener el valor óptimo de componentes aplicamos un PCA a nuestra columna de vectores. Esta técnica de reducción de la dimensionalidad transforma un conjunto de componentes en un conjunto más pequeño de componentes no correlacionadas llamadas componentes principales que explican la variabilidad en los datos.

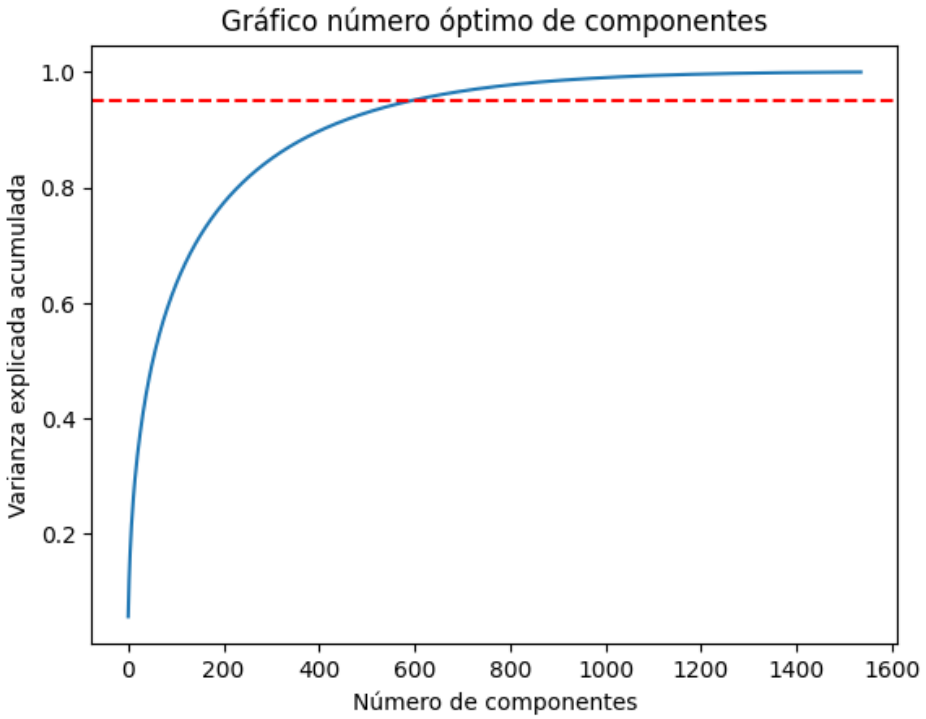


Figura 5: Gráfico del número óptimo de componentes del vector de embeddings.

Fuente: Elaboración propia.

Cómo se puede observar en el gráfico con 600 componentes del vector se puede explicar el 95% de la varianza explicativa acumulada.

Una vez se ha reducido la longitud del vector a 600, se aplican tres métodos distintos para determinar el número óptimo de clusters: el método del codo, el método de la silueta y el método de Gap.

Cada uno de ellos arroja un resultado diferente, ya que el método del codo sugiere que el número óptimo de clusters podría ser 5, mientras que el método de la silueta sugiere k = 3 y el método de Gap sugiere k = 1. Debido a que este análisis se realiza para comprender mejor los datos, se decide dividir la muestra en 5 clusters. Se replica el método de clustering y se obtiene un score de 0.032034. Debido a que el resultado no es muy bueno, se evalúa la división en 1 a 10 clusters y se observa que los scores resultantes son prácticamente iguales. Por lo tanto, se decide mantener la división en 5 clusters. El histograma resultante se muestra en la figura 6:

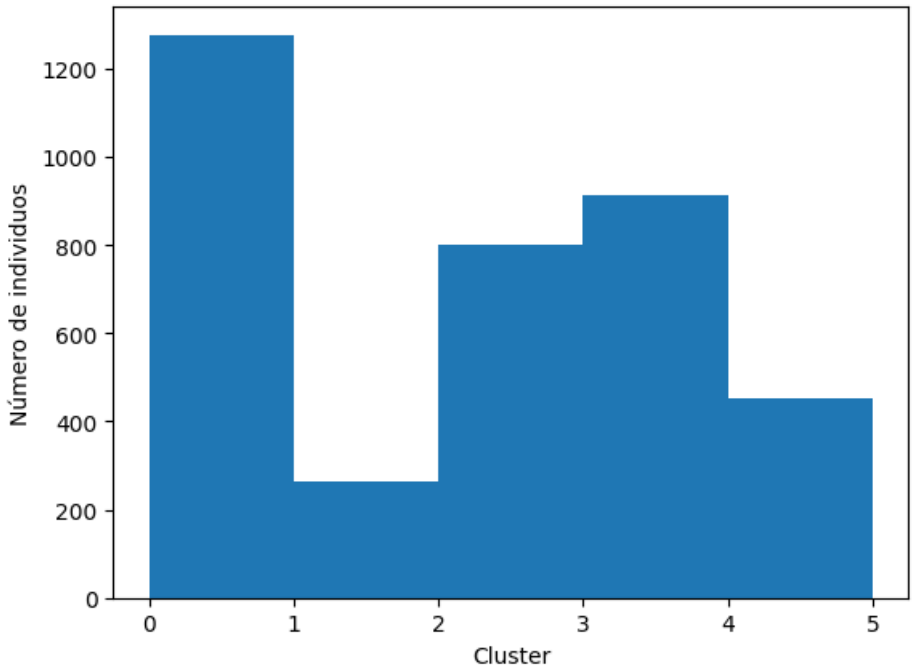


Figura 6: Histograma del número de individuos que pertenecen a cada cluster.

Fuente: Elaboración propia.

El recuento de empresas por clúster es el siguiente: 1277 empresas para el clúster 0, 265 empresas para el clúster 1, 799 empresas para el clúster 2, 913 empresas para el clúster 3 y 451 empresas para el clúster 4.

Además, para entender las diferencias que entre los diferentes clusters según el texto de su página web, se obtienen las palabras más recurrentes de cada uno de los grupos creados:

- Clúster 0 : este grupo parece estar relacionado con empresas que se dedican a la producción y venta de productos de calidad, especialmente conservas. Puede incluir empresas que destacan por la calidad de sus productos y brindan información detallada sobre ellos en sus sitios web. Entre las palabras más repetidas en este clúster destacan: productos, calidad, web, empresas, conservas e información.
- Clúster 1: las palabras pertenecientes a este grupo sugieren que está compuesto por empresas relacionadas con la producción y comercialización de productos derivados del olivo, como el aceite de oliva virgen extra y aceitunas de calidad. Entre las palabras más repetidas en este clúster destacan: aceite, oliva, virgen, extra, calidad y aceitunas.
- Clúster 2: las palabras pertenecientes a este clúster indican que podría representar empresas especializadas en productos cárnicos y lácteos. Entre las palabras más repetidas en este clúster destacan: productos, embutidos, jamón, calidad, quesos e ibérico.

- Clúster 3: las palabras relacionadas con este grupo sugieren que podría estar compuesto por empresas del sector de la panadería, pastelería y café. Podrían ser empresas especializadas en la producción y venta de productos horneados, dulces, chocolates y cafés. Entre las palabras más repetidas en este clúster destacan: productos, pan, calidad, pastelería, panadería, chocolate y café.
- Clúster 4: este grupo parece ser diferente de los anteriores, ya que no parece relacionarse directamente con un tipo específico de empresas. Estas palabras podrían indicar que este grupo incluye empresas relacionadas cuyas webs no están operativas. Entre las palabras más repetidas en este clúster destacan: web, página, server, found, www, error y forbidden.

Variables de presencia web

Se analizan las variables de número de imágenes, videos, formularios y enlaces en cada página web con el objetivo de determinar si existe alguna relación entre estos resultados y los indicadores de competitividad correspondientes a su clúster.

A pesar de las expectativas iniciales, los resultados obtenidos no muestran ninguna relación significativa entre las variables analizadas (número de imágenes, videos, formularios y enlaces) y los indicadores de rentabilidad económica, número de empleados o activos intangibles para ninguno de los clusters identificados. Estos hallazgos sugieren que las variables estudiadas no son predictivas ni influyentes en los aspectos mencionados en los diferentes clusters. A continuación se muestran los resultados de la distribución de las variables de la presencia web en función de los indicadores de competitividad y clúster:

Tabla 3: Valores medios del número de imágenes, videos, enlaces y formularios por clúster en función de la media del número de empleados, la media del inmovilizado intangible y la media de la rentabilidad económica.

		Número de empleados	Inmovilizado intangible (miles de €)	Rentabilidad económica (%)
Número medio de imágenes	Clúster 0	22.97	22.97	22.97
	Clúster 1	24.13	24.13	24.13
	Clúster 2	24.51	24.51	24.51
	Clúster 3	24.05	24.05	24.05
	Clúster 4	6.00	6.00	6.00

Número medio de videos	Clúster 0	0.13	0.13	0.13
	Clúster 1	0.11	0.11	0.11
	Clúster 2	0.09	0.09	0.09
	Clúster 3	0.06	0.06	0.06
	Clúster 4	0.05	0.05	0.05
Número medio de enlaces	Clúster 0	62.39	62.39	62.39
	Clúster 1	65.26	65.26	65.26
	Clúster 2	64.88	64.88	64.88
	Clúster 3	63.41	63.41	63.41
	Clúster 4	16.22	16.22	16.22
Número medio de formularios	Clúster 0	7.35	7.35	7.35
	Clúster 1	7.05	7.05	7.05
	Clúster 2	7.33	7.33	7.33
	Clúster 3	7.93	7.93	7.93
	Clúster 4	3.39	3.39	3.39

Fuente: Elaboración propia.

Variable número de empleados

A través del análisis de la evolución del número de empleados en función del tiempo y de los distintos clusters, se observa que la cantidad de empleados se mantiene prácticamente constante en todos los años, independientemente del clúster al que pertenezca la empresa. En efecto, si bien se aprecia un ligero aumento en la media del número de empleados para cada clúster a lo largo del tiempo, exceptuando los clusters 3 y 4 en los años 2019-2020 y 2020-2021 respectivamente. El valor medio más elevado se encuentra en el clúster 0 para el año 2021, con una cifra de 40 empleados.

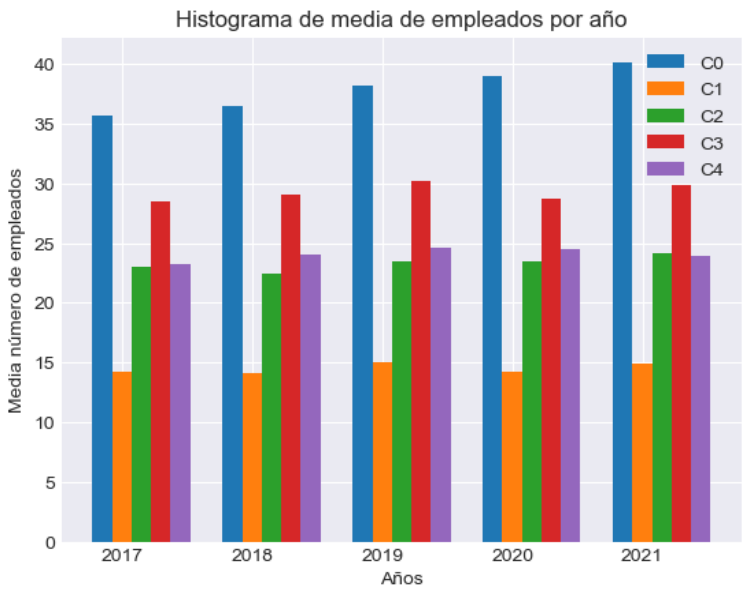


Figura 7: Histograma de media de empleados por año y clúster.

Fuente: Elaboración propia.

Variable rentabilidad económica.

Si nos enfocamos en los valores medios de rentabilidad económica por clúster y año, parece que los datos no siguen un patrón claro. Sin embargo, es importante destacar que el clúster 0 siempre obtiene los valores más altos en términos de rentabilidad. Incluso el clúster 3 en 2020 llega a obtener valores negativos.

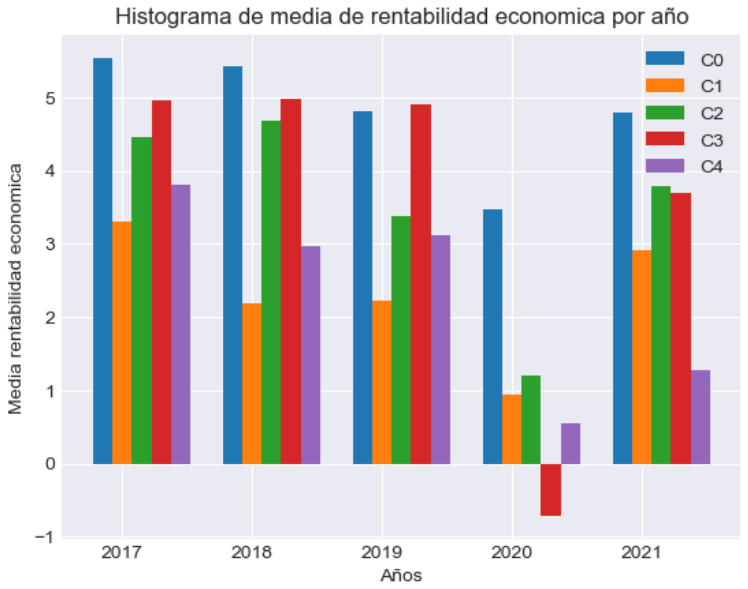


Figura 8: Histograma de media de rentabilidad económica por año y clúster.

Fuente: Elaboración propia.

Variable inmovilizado intangible.
Para concluir, la media de los valores del inmovilizado intangible por año y clúster muestra una cierta tendencia en su evolución. En particular, el clúster 0 exhibe una trayectoria creciente en sus datos, mientras que los clusters 1 y 2 mantienen valores constantes a lo largo de los años. Por otro lado, se observa que los valores del inmovilizado intangible del clúster 3 y 4 disminuyen con el tiempo.

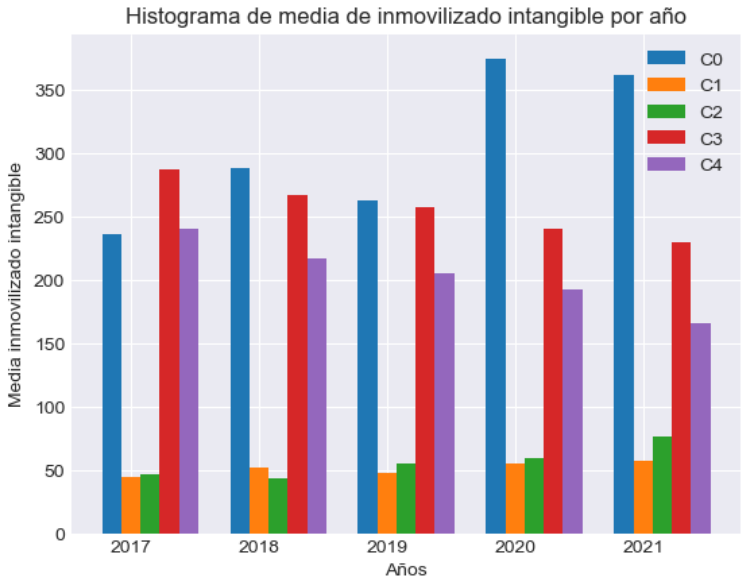


Figura 9: Histograma de media de inmovilizado intangible por año y clúster.
Fuente: Elaboración propia.

Se ha observado que el clúster 0 sobresale en los indicadores de competitividad, a pesar de tener un mayor número de observaciones. Esto se debe a que las empresas dedicadas a aspectos como la calidad y que se centran en dar una información detallada del producto tienden a ser más competitivas.

Una sección relevante para estudiar es cómo el número de imágenes, videos, formularios y enlaces en la página web influyen en los indicadores de competitividad. Sin embargo, a pesar de las expectativas de encontrar resultados o conclusiones significativas sobre este tema, los datos revelan que no existe una relación aparente entre el número de elementos multimedia o formularios y los valores de empleados, rentabilidad económica o inmovilizado intangible.

6.2 - Evaluación de modelos

6.2.1 - Análisis Clustering

Con el fin de predecir si una empresa es competitiva o no en función de los valores de su vector de embedding, aplicamos el método de clustering a las variables de competitividad (media del inmovilizado intangible, media del número de empleados y media de la rentabilidad económica) para dividir las empresas en clusters en función de su competitividad.

En primer lugar, se presentan los procedimientos utilizados para la selección del número óptimo de clusters. El método del codo (ilustrado en la figura 10), indica que los datos deberían ser agrupados en un rango de 2 a 3 clusters, dado que es en este punto donde la curva comienza a presentar un aplanamiento significativo en función del incremento de un clúster adicional.

La tabla 4 muestra los resultados del índice Davies-Baulvin, los resultados muestran que el número óptimo de clusters es k = 2 ya que es el valor mínimo.

El tercer método estudiado es el método de la silueta (figura 11). En este caso el coeficiente alcanza su mayor valor cuando el número de clusters es igual a 2.

Por último, el método de Gap devuelve que el número óptimo de clusters se encuentra en k=1.

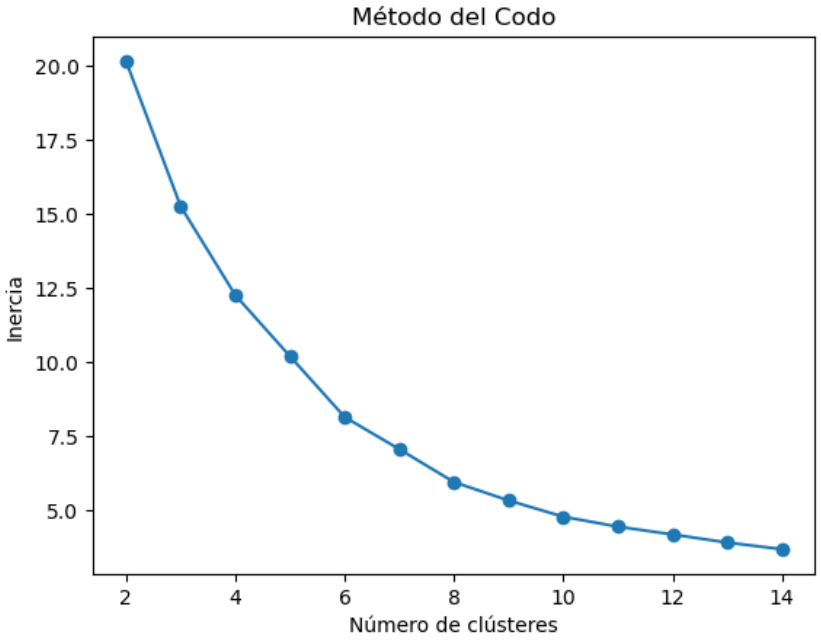


Figura 10: Gráfico del método del codo.

Fuente: Elaboración propia.

Tabla 4: Índices Davies-Baulvin

Número de clusters	Índice Davies-Baulvin
2	0.603
3	0.798
4	0.903
5	0.760
6	0.716
7	0.714
8	0.697
9	0.730
10	0.762

Fuente: Elaboración propia.

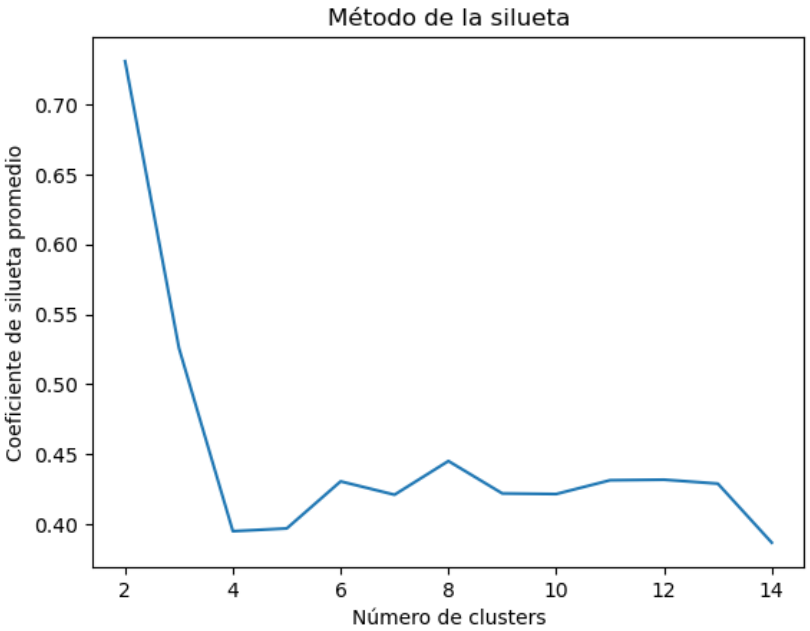


Figura 11: Gráfico método de la silueta.
Fuente: Elaboración propia.

Después de haber evaluado los indicadores mencionados, se ha tomado la decisión de establecer el número óptimo de clusters en k=2. En consecuencia, los datos han sido divididos en empresas competitivas y no competitivas, lo cual se ajusta a la lógica teórica y proporciona una descripción práctica del conjunto de observaciones.

La figura 12 ilustra los resultados obtenidos luego de haber aplicado el método de clustering kmeans con un valor de k=2. En este caso, el puntaje obtenido es de 0.7308, lo cual representa una mejora de al menos 0.21 puntos en comparación con los scores obtenidos al utilizar otros números de clusters.

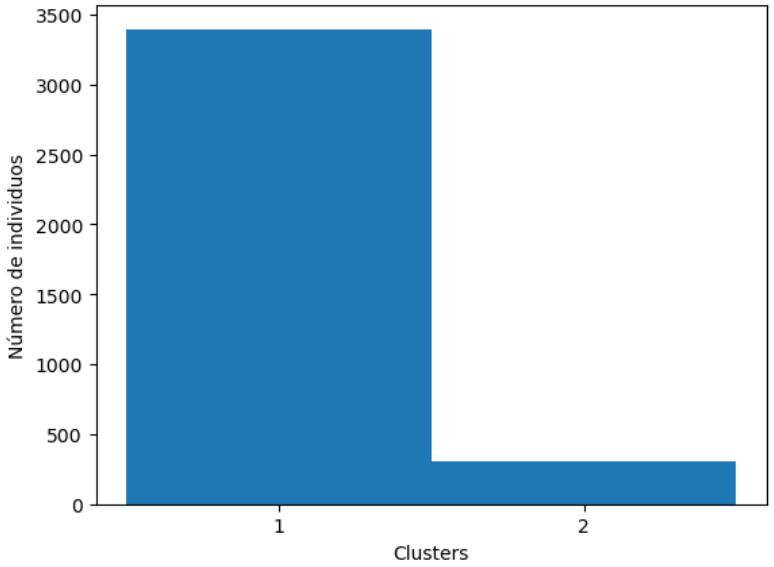


Figura 12: Histograma resultante al método kmeans.
Fuente: Elaboración propia.

Caracterización de los clusters

Los resultados del análisis clustering muestran que las 3705 empresas se agrupan en dos clusters, perteneciendo 3396 al primer clúster y 309 al segundo. En relación al número de observaciones que pertenecen a cada clúster se puede afirmar que las clases están desbalanceadas. En la siguiente tabla se pueden observar los valores medios de cada una de las variables estudiadas en función del clúster perteneciente:

Tabla 5: Valores medios de indicadores de competitividad según clúster.

k	1	2
Rentabilidad económica (%)	3.885	4.813
Número de empleados (miles de €)	21.032	131.273
Inmovilizado intangible	66.647	1077.486

Fuente: Elaboración propia.

A pesar de que se observa un desequilibrio en la distribución de las clases, se evidencian diferencias significativas en términos de competitividad, particularmente en las variables que corresponden al número de empleados. Los resultados del análisis del test de Kruskal-Wallis muestran que el valor de p es menor a 0.001, lo que indica que se rechaza la hipótesis nula que establece que no existen diferencias estadísticamente significativas entre los dos grupos.

Es importante destacar la relación existente entre las variables analizadas en los clusters. En el clúster 1 se observa como los valores medios de las variables son menores en comparación al clúster 2. Esto sugiere que existe una relación entre las variables, donde el aumento en la rentabilidad económica podría estar relacionado con el aumento en el número de empleados y en el valor de inmovilizado intangible.

Si nos basamos únicamente en la rentabilidad económica como indicador de competitividad, podría deducirse que las empresas en el clúster 2 son más competitivas que las del clúster 1, dado que poseen una rentabilidad económica media superior. Sin embargo, es importante tener en cuenta que los datos presentan desbalance y la dispersión de los datos en el clúster 1 podría estar afectando la media de rentabilidad económica, lo que podría subestimar la competitividad de algunas empresas en dicho clúster. Por lo tanto, se considerará que las empresas más competitivas se encuentran en el clúster 2, mientras que las menos competitivas pertenecen al clúster 1.

Por tanto, el análisis de clustering resulta una herramienta valiosa para comprender mejor la relación entre estas variables y poder desarrollar estrategias que permitan mejorar la rentabilidad y competitividad de las empresas.

6.2.2 - Evaluación de los modelos de clasificación

En esta sección se describe el proceso utilizado para todos los modelos y los resultados obtenidos. Antes de aplicar los modelos, los datos se dividen en conjuntos de entrenamiento (80%) y prueba (20%). Dado que los datos están desequilibrados, se utiliza el algoritmo SMOTE en los datos de entrenamiento para igualar el número de observaciones en ambas clases y asegurar que el modelo no esté sesgado hacia la clase mayoritaria. Una vez entrenado el modelo, se evalúa su desempeño utilizando los datos de prueba. El objetivo es predecir si una empresa es competitiva o no en función del valor de sus vectores de embedding obtenidos a partir de su sitio web.

Se han desarrollado cinco modelos de clasificación: regresión logística, máquinas de vectores de soporte, árboles de decisión, random forest y redes neuronales. A continuación, se presentan los resultados de cada uno de ellos.

En este proyecto, se utilizan algoritmos para clasificar de forma correcta la mayor cantidad posible de clases, por lo que se considera importante obtener el accuracy. El método Random Forest muestra el valor más alto con un 89.34% de aciertos, seguido por SVM con 77.87%, árbol de decisión con 77.06%, regresión logística con 75.71% y red neuronal con 72.38%.

Además, para complementar esta información se utilizan las matrices de confusión, las cuales comparan los valores predichos en los datos de validación con los valores reales.

En la tabla 6 se observan los resultados obtenidos para cada uno de los modelos.

Tabla 6: Matrices de confusión para 5 métodos aplicados.

Regresión Lineal		
Predicho	Observado	
	Clúster 0	Clúster 1
Clúster 0	513	146
Clúster 1	34	48

SVM		
Predicho	Observado	
	Clúster 0	Clúster 1
Clúster 0	530	129
Clúster1	35	47

Árbol de Decisión		
Predicho	Observado	
	Clúster 0	Clúster 1
Clúster 0	548	111
Clúster 1	59	23

Random Forest		
Predicho	Observado	
	Clúster 0	Clúster 1
Clúster 0	658	1
Clúster 1	78	4

Red Neuronal		
Predicho	Observado	
	Clúster 0	Clúster 1
Clúster 0	489	170
Clúster 1	34	48

Fuente: Elaboración propia.

En la tabla 7 se presentan los valores de sensibilidad y especificidad. En el contexto de los datos, la sensibilidad se refiere a la capacidad del modelo para detectar empresas no competitivas que realmente no lo son, mientras que la especificidad se calcula como el número de empresas competitivas que son clasificadas como tales por el modelo.

En general, los datos muestran valores moderados para la sensibilidad, lo que indica que el modelo está clasificando erróneamente algunas empresas competitivas como no competitivas. Sin embargo, este error no es tan grave como clasificar empresas no competitivas como competitivas, ya que esto tiene un costo de error mucho mayor. Por otro lado, los resultados de la especificidad son mejores, lo que indica que el modelo tiene un menor error al clasificar empresas competitivas.

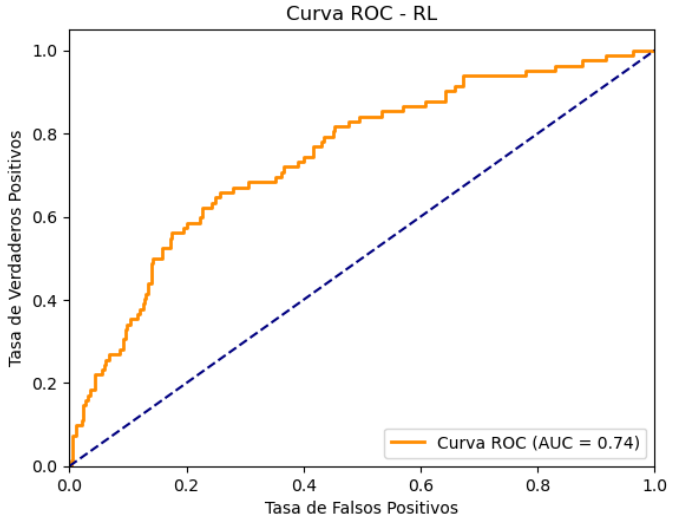
Teniendo en cuenta ambas variables, los mejores resultados se obtienen con los modelos de regresión logística, máquinas de vectores soporte y redes neuronales. Es importante destacar que Random Forest, que obtuvo la mejor precisión (accuracy), tiene una especificidad perfecta (1), lo que indica que clasifica correctamente la clase minoritaria de empresas competitivas, aunque su sensibilidad es la más baja entre los modelos evaluados.

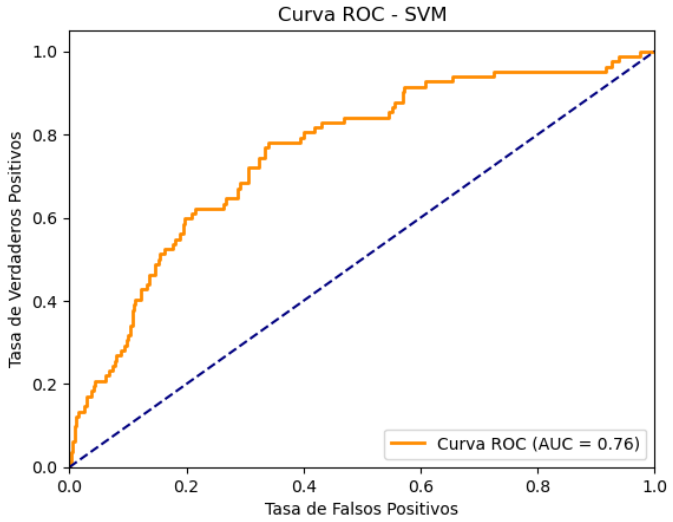
Tabla 7: Valores de sensibilidad y especificidad para los diferentes métodos.

Modelo	Sensibilidad	Especificidad
Regresión Logística	0.59	0.78
SVM	0.57	0.80
Árbol de decisión	0.28	0.83
Random Forest	0.05	1.00
Red Neuronal	0.59	0.74

Fuente: Elaboración propia

De una forma más visual se muestran a continuación los resultados de las curvas ROC con su respectivo valor AUC:





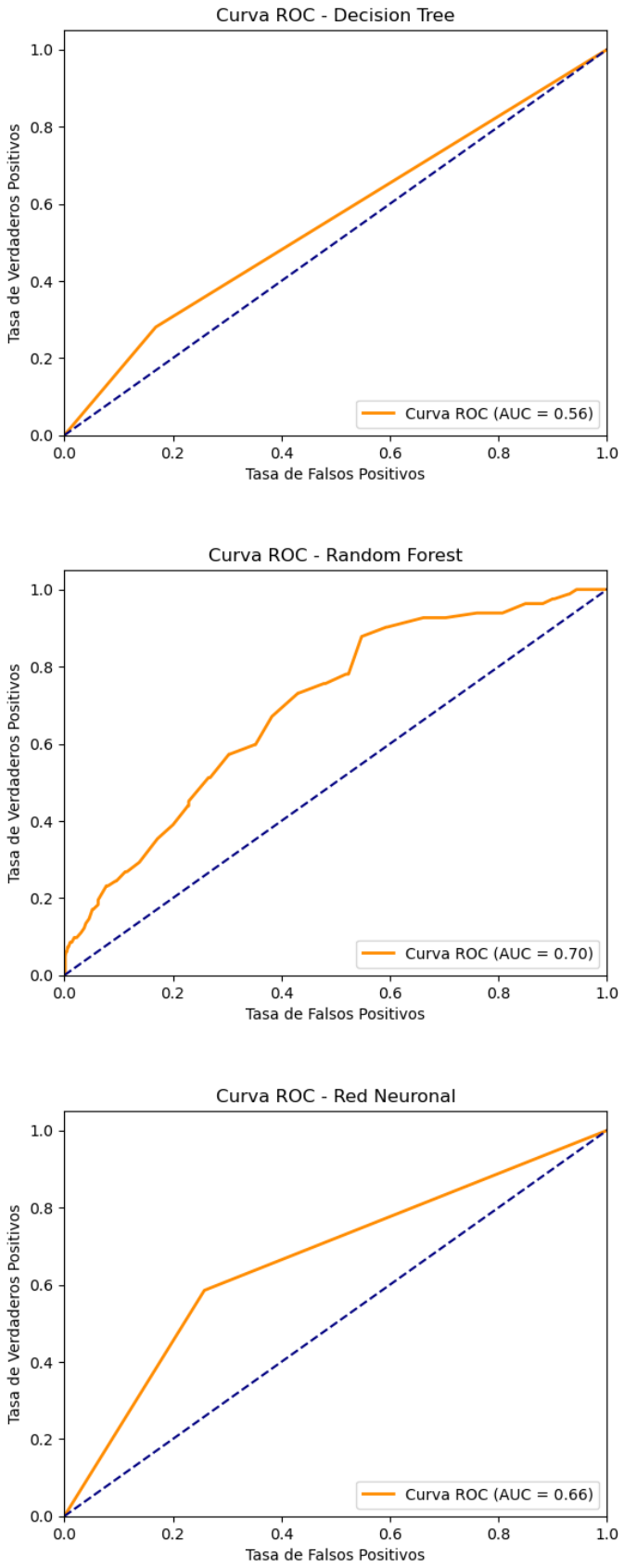


Figura 13: Gráficos curva ROC y el valor AUC para el método de Regresión Logística, SVM, Árbol de Decisión, Random Forest y Red Neuronal.

Fuente: Elaboración propia.

Tras el análisis de los gráficos y la consideración de otras métricas previamente descritas, se concluye que los modelos de SVM y Regresión Lineal presentan el mejor rendimiento. A pesar de que el modelo Random Forest muestra el mejor índice de precisión, se debe tener en cuenta que también incurre en un elevado costo de error al clasificar erróneamente un gran número de empresas no competitivas. Aunque el costo de clasificar empresas competitivas como no competitivas sea menor que el costo opuesto, se considera que este tipo de error también afecta negativamente la calidad del modelo.

6.2.3 - Predicción rentabilidad económica

Finalmente, se realizó un modelo de regresión lineal con el fin de predecir la rentabilidad económica de la empresa a partir de los valores del vector de embedding. Se obtuvieron 1536 coeficientes correspondientes a cada una de las variables independientes del vector para calcular el valor de la variable dependiente.

En la figura 14 se puede observar la diferencia de los valores predichos frente a los valores reales. Las medidas de error utilizadas en la evaluación de modelos de regresión son la desviación absoluta media y el error cuadrático medio de la raíz

MAD se refiere a la media de las diferencias absolutas entre las predicciones del modelo y los valores reales. Un valor de MAD de 4.06099 significa que, en promedio, las predicciones del modelo difieren de los valores reales en una cantidad de 4.06099 puntos porcentuales.

RMSE, por otro lado, es una medida de cuánto varían las predicciones del modelo en relación con los valores reales. Un valor de RMSE de 5.5870 significa que la raíz cuadrada de la media de los errores cuadrados entre las predicciones del modelo y los valores reales es 5.5870 en la escala de la variable de respuesta. En otras palabras, el modelo tiene una variación de error de 5.5870 unidades en promedio.

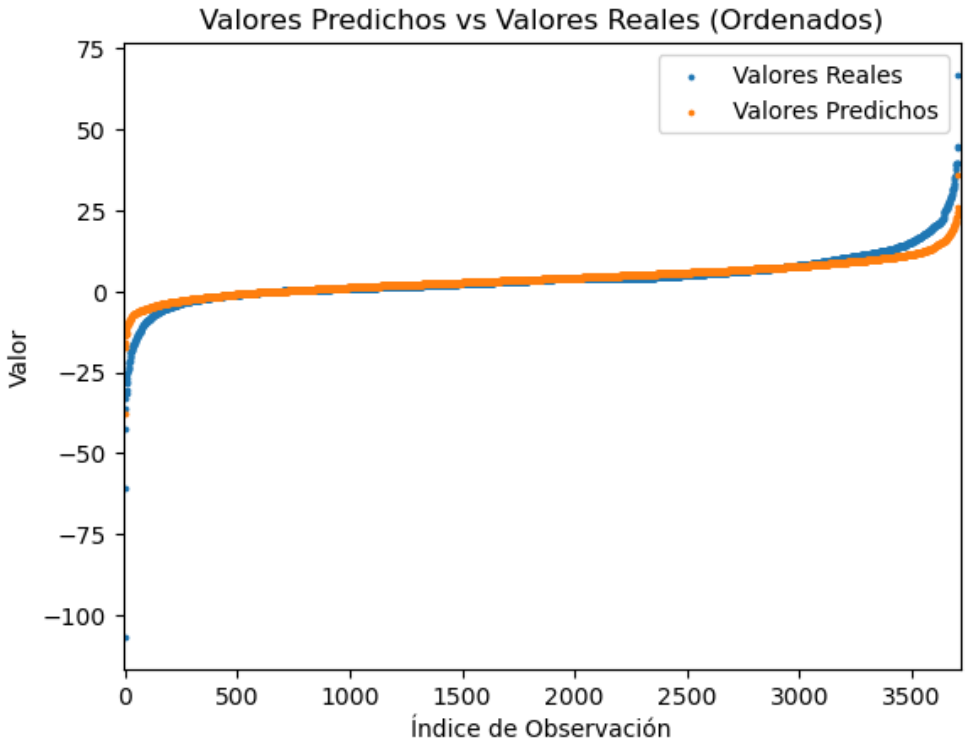


Figura 14: Valores reales frente a los predichos de rentabilidad económica.

Fuente: Elaboración propia.

7 - Conclusiones

Para culminar el proyecto, se presentan las conclusiones obtenidas, habiendo cumplido con el objetivo de clasificar empresas como competitivas o no, a partir de su presencia online, a través de vectores de embeddings. El objetivo principal de todas las empresas es, sin duda, lograr la competitividad, y a través de nuestro clasificador se ha logrado diferenciar las que lo son de las que no lo son. Tras el análisis exploratorio, con el fin de obtener información relevante para entender los datos que se están estudiando, se obtuvo que había cinco grupos diferentes de empresas dentro de la muestra. Las diferencias entre ellos se hallaba en el contenido de su página web ya que se podía ver como cada sector podría referirse a empresas que se enfocan en la venta de aceites, en los ibéricos, en control de calidad…

Además de cumplir con el objetivo general, también se lograron resolver los objetivos específicos del proyecto. Se recopiló toda la información necesaria sobre la competitividad y se obtuvieron las variables de la presencia online que se requerían para el análisis. Se realizó el preprocesamiento del texto de la web, eliminando información irrelevante y convirtiendo el texto en variables numéricas para el análisis. Se aplicaron las técnicas de aprendizaje automático adecuadas para construir el modelo que permitió predecir los indicadores de competitividad. Se halló un clúster más grande que corresponde a las empresas menos competitivas. Y otro clúster más pequeño correspondiente a empresas competitivas y amplitud de innovación, con valores más altos en el número de empleados, la rentabilidad económica y el inmovilizado intangible, haciendo que destaquen sobre el resto. Por último, se evaluó con éxito la calidad de los métodos utilizados, determinando la existencia o no de una relación significativa entre el texto de la página web de las empresas y sus indicadores de competitividad. El código con el cual se ha llevado a cabo el proyecto está disponible en el siguiente *repositorio* de github.

La ventaja de utilizar minería de datos para la obtención del texto de la página web supone trabajar con datos a tiempo real. A través de estos clasificadores, las empresas pueden conocer su situación y, en el caso de no ser competitivas, identificar sus fallos y corregirlos para mejorar.

Por otro lado, es importante mencionar que el costo de etiquetar erróneamente una empresa como competitiva o no competitiva puede ser significativo. Si una empresa es considerada no competitiva cuando en realidad sí lo es, puede perder oportunidades de negocio y caer detrás de la competencia. En cambio, si una empresa es considerada competitiva cuando en realidad no lo es, puede llevar a la inversión en estrategias de marketing y publicidad que no generan los resultados deseados y pueden ser un gasto innecesario. Por lo tanto, es esencial que los

resultados del análisis de huella digital se validen con otros indicadores de competitividad empresarial antes de tomar decisiones importantes.
Dicho esto, una mejora en la página web de una empresa puede ser un factor importante para mejorar su competitividad en línea. Una página web bien diseñada, fácil de usar y con contenido relevante y actualizado puede ayudar a atraer a más visitantes, aumentar la retención de clientes existentes y mejorar la imagen de la empresa. También puede mejorar la posición de la empresa en los resultados de búsqueda en línea, lo que puede aumentar la visibilidad y la confianza de los consumidores. Si una empresa no es competitiva, es posible que su contenido no esté a la altura de las expectativas de sus clientes. Al mejorar el texto de su página web, las empresas pueden enfocarse en destacar los beneficios y las características de sus productos o servicios de una manera más clara y efectiva, lo que puede aumentar su atractivo para los consumidores y mejorar su competitividad. Por lo tanto, una empresa que no es competitiva no puede llegar a serlo solo mejorando su página web, pero una mejora en la página web puede ser un factor importante para mejorar la competitividad en línea y atraer a más clientes.

Cabe mencionar que el estado puede utilizar los resultados obtenidos de este proyecto para monitorizar la economía y tomar medidas adecuadas en el caso de detectar que las empresas están volviéndose menos competitivas. Si se detecta que sectores específicos de la economía están experimentando una disminución en la competitividad, el estado puede centrarse en la evaluación de factores como la innovación, la calidad de los productos o servicios, la eficiencia operativa y la capacidad de adaptación a los cambios del mercado. Además, puede evaluar el impacto de las políticas en la competitividad empresarial y realizar ajustes si es necesario. También se pueden implementar programas de apoyo dirigidos a empresas que están experimentando dificultades competitivas. Estos programas pueden incluir asesoramiento empresarial, acceso a financiamiento, incentivos para la innovación y la adopción de nuevas tecnologías, entre otros.

Sería valioso explorar en mayor profundidad la medición de la competitividad empresarial y considerar otras variables que podrían estar relacionadas y ayudar a entender mejor el fenómeno. Como posible trabajo futuro, se podría ampliar el estudio para analizar el impacto de la presencia en redes sociales de las empresas en su competitividad. Además, sería importante evaluar la calidad del servicio y la atención al cliente como factores clave de competitividad. Para mantener el trabajo, se podría actualizar el modelo de clasificación periódicamente a medida que cambien los patrones de comportamiento de los consumidores y las tendencias en la industria.

En el transcurso de este proyecto, se han utilizado habilidades adquiridas de manera integral en todas las asignaturas cursadas en el grado, pero es pertinente resaltar aquellas que han resultado más determinantes. Entre ellas, destacan "Modelos descriptivos y predictivos", "Evaluación, despliegue y monitorización de

modelos", "Comportamiento económico y social" y "Lenguaje Natural y recuperación de la información". Estas asignaturas han proporcionado las herramientas necesarias para la obtención y procesamiento de datos a partir del texto web, la construcción y evaluación del modelo de clasificación y la comprensión de los indicadores de competitividad desde una perspectiva económica. Además, la asignatura de proyectos ha sido clave para la gestión y desarrollo efectivo del TFG.

Bibliografía

Aghion, P., & Howitt, P. (2009). The Economics of Growth. MIT Press.

Álvarez, J. R. (2005). Los activos intangibles y su valoración. Actualidad Contable FACES, 8(11), 11-23.

Amazon Web Services. (s.f.). ¿Qué es una red neuronal? | Amazon Web Services. Recuperado 10 de mayo de 2023.

Berlanga Silvestre, V., & Vilà-Baños, R. (2014). Cómo obtener un modelo de Regresión Logística Binaria con SPSS. REIRE, Revista d'Innovació i Recerca en Educació, 105-118.

Blazquez, D., Domenech, J., & Debón, A. (2018). Do corporate websites changes reflect firms survival? Online Information Review, 42(6), 956–970.

Blazquez, D., & Domenech, J. (2018). Web data mining for monitoring business export orientation. Technological and Economic Development of Economy, 24(2), 406–428.

Brealey, R. A., Myers, S. C., & Allen, F. (2017). Principles of corporate finance. McGraw-Hill Education.

Breiman, L. (2001). Random forests. Machine learning, 45:5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Chapman & Hall.

CEA (2021). El SABI, una herramienta indispensable para el inversor. Recuperado el 9 de mayo de 2023, de https://masempresas.cea.es/el-sabi-una-herramienta-indispensable-para-el-inversor/

CEDE. (2009). Activos intangibles en la empresa. Recuperado el 9 de mayo de 2023, de http://www.cede.es/uploads/File/Pdf/Informes/ActivosIntangibles.pdf

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chesbrought, H. (2003). Open Innovation: The New Imperative for Creating and Profiting from Technology. Harvard Business Review Press.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (2012). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics, 25(4), 351-361.

Datasource.ai. (2022). Comprensión de la matriz de confusión y cómo implementarla en Python.

Economia 3 (2021). OpenAI, la empresa creada por Elon Musk y Sam Altman para avanzar en la Inteligencia Artificial. Recuperado el 9 de mayo de 2023, de https://economia3.com/openai-empresa/

Esbensen, K. H. and Geladi, P. (2009). Principal component analy- sis: Concept, geometrical interpretation, mathematical background, algo- rithms, history, practice. In Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, VOLS 1-4, pages A211–A226.

Flynn, B. B., Schroeder, R. G., & Sakakibara, S. (2010). The impact of quality management practices on performance and competitive advantage. Operations Management Research, 3(1-2), 22-42.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Greene, D. J., y Goggins, S. P. (2010). The digital footprint as a cue for contextual suggestion. ACM Transactions on Computer-Human Interaction (TOCHI), 17(1), 1-19.

Gulati, H., Singh, P. (2015). Clustering techniques in data mining: A comparison. In 2015 2nd international conference on computing for sustainable global development (INDIACom), pages 410–415. IEEE.

Han, S., Ryu, S., & Kim, J. (2018). The Effect of Website Quality on Purchase Intention in Mobile Commerce: Trust as a Mediator. Journal of Theoretical and Applied Electronic Commerce Research, 13(2), 37-51.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

Hiros, G. (2022). K-means Clustering. Disponible en https://h1ros.github.io/posts/k-means-clustering/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.

Kotler, P. (2000). Marketing management. Prentice-Hall.

Kotler, P. (2011). Reinventing marketing to manage the environmental imperative. Journal of Marketing, 75(4), 132-135.

Li, J., Luo, S., & Wang, S. (2016). A multiple imputation method for missing data using high-order interactions. Computational Statistics & Data Analysis, 104, 221-233.

Li, X., Chen, S., & He, J. (2019). How Does Website Quality Affect Consumers' Quality Perceptions and Behavioral Intentions? Journal of Interactive Marketing, 45, 91-104.

OpenAI. (2022). What are embeddings? OpenAI Platform.https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

Porter, M. E. (1990). The competitive advantage of nations. Harvard Business Review Press.

Porter, M. E. (1996). ¿Qué es la estrategia?. Harvard Business Review, 74(6), 50-63.

Porter, M. E. (2008). La ventaja competitiva de las naciones. Plaza y Janés.

Project Management Institute. (2017). Guía de los fundamentos para la dirección de proyectos (Guía del PMBOK). Proyectos & Produccion.

Rothaermel, F. T. (2015). Strategic management (2nd ed.). McGraw-Hill Education.

Smith, J. D. (2023). Analyzing the Competitive State of the Productive Fabric through Corporate Websites. Operations Research and Decisions, 50(2), 121-136.

Suárez, E. J. (2014). Tutorial sobre Máquinas de Vectores Soporte (SVM). Madrid: UNED.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. Surveillance & Society, 12(2), 197-208.

We Are Social y Hootsuite. (2021). Digital 2021: Global Overview Report

Anexos

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

Tablas: Objetivos de desarrollo sostenible.

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No procede
1. Fin de la pobreza.				X
2. Hambre cero.				X
3. Salud y bienestar.			X	
4. Educación de calidad.				X
5. Igualdad de género.				X
6. Agua limpia y saneamiento.				X
7. Energía asequible y no contaminante.				X
8. Trabajo decente y crecimiento económico.	X			
9. Industria, innovación e infraestructuras.	X			
10. Reducción de las desigualdades.		X		
11. Ciudades y comunidades				X

sostenibles.				
12. Producción y consumo responsables.		X		
13. Acción por el clima.				X
14. Vida submarina.				X
15. Vida de ecosistemas terrestres.				X
16. Paz, justicia e instituciones sólidas.				X
17. Alianzas para lograr objetivos.		X		

Fuente: Elaboración propia.

Con el fin de abordar desafíos mundiales y promover un desarrollo sostenible en diferentes áreas, la Organización de las Naciones Unidas (ONU) estableció los Objetivos de Desarrollo Sostenible (ODS).

A continuación, se describen las dos ODS que tienen mayor relación con el proyecto:

- ODS 8: Trabajo decente y crecimiento económico: el proyecto está estrechamente relacionado con este objetivo, ya que busca analizar la competitividad de las empresas para poder establecer medidas que ayuden al crecimiento de estas.
- ODS 9: Industria, innovación e infraestructura: el proyecto tiene una fuerte relación con este objetivo , ya que busca utilizar la web de las empresas como una fuente de datos para analizar la competitividad y promover la innovación.

En menor medida, hay otros objetivos que también están relacionados con el proyecto:

- ODS 10: Reducción de las desigualdades: si bien el proyecto no aborda directamente la reducción de desigualdades, pero puede tener una implicaciones indirectas al identificar factores que contribuyen a la competitividad económica.
- ODS 12: Producción y consumo responsables: El proyecto puede tener relación indirecta con este objetivo al analizar la competitividad económica y su relación con prácticas responsables de producción y consumo.
- ODS 17: Alianzas para lograr los objetivos: El proyecto puede promover la colaboración entre diferentes actores para analizar la competitividad económica y contribuir a objetivos económicos más amplios.

Además, alguno de los objetivos puede tener una relación indirecta con el proyecto:

- ODS 3: Salud y bienestar: El proyecto podría contribuir indirectamente a este objetivo, ya que la competitividad económica puede influir en la calidad de vida y en el acceso a servicios de salud.