UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Análisis de la rentabilidad sobre los activos y la huella digital de las bodegas de vino en España

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Vizuete Romero, Jorge

Tutor/a: Debón Aucejo, Ana María

Cotutor/a externo: Álvaro Meca, Luis Alejandro

CURSO ACADÉMICO: 2023/2024

UNIVERSIDAD POLITÉCNICA DE VALENCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Análisis de la rentabilidad sobre los activos y la huella digital de las bodegas de vino en España

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Vizuete Romero, Jorge

Tutor/a: Debón Aucejo, Ana María

Cotutor/a: Álvaro Meca, Luis Alejandro

CURSO ACADÉMICO: 2023/2024

Agradecimientos

Quiero expresar mi agradecimiento a Ana y Alex por su acompañamiento técnico y sus valiosas recomendaciones a lo largo del desarrollo de este trabajo.

A mis padres, José Antonio y Mª Pilar, y a mi hermano, Raúl, quienes siempre me han apoyado incondicionalmente y han sido el pilar fundamental en mi educación.

Finalmente, agradezco a mi amiga y compañera de viaje por ayudarme a creer un poco más en mí mismo y a ver los cambios como ilusiones en forma de objetivos.

Resumen

El sector vitivinícola en España presenta una gran importancia sobre la economía y la cultura del país. Debido a su elevada competitividad en el mercado y las nuevas tendencias empresariales para su diferenciación, como el enoturismo, el marketing y el desarrollo de la digitalización empresarial. El objetivo general de este trabajo es analizar la relación de la huella digital y la competitividad (medida como el rendimiento financiero sobre los activos, ROA), en bodegas de vino en España utilizando técnicas multivariantes y de minería de datos.

Para el estudio se utilizaron indicadores de competitividad (ROA), obtenidos del sistema de Análisis de Balances ibéricos (SABI) e indicadores de huella digital extraídos de las páginas web de las bodegas y de sus redes sociales. La muestra consta de 1229 bodegas españolas y 107 variables (ROA, 102 variables de contenido web “keywords” y 5 variables de presencia en redes sociales “hrefwords”).

En primer lugar, se realizó un trabajo de preprocesamiento de los datos con tratamiento de valores atípicos e imputación de valores faltantes y posteriormente se implementaron diversas técnicas de aprendizaje no supervisado para explorar las relaciones del conjunto de variables y observaciones, así como para encontrar grupos o clusters de empresas.

Además, se utilizó el indicador de competitividad (ROA) como target o variable respuesta y los indicadores de huella digital (keywords y Hrefwords) como variables explicativas, para la implementación de diversos modelos de regresión en diferentes lenguajes de programación (Python y R), para comparar su capacidad predictora y extraer información relevante acerca de la importancia de las variables.

Entre las principales conclusiones, cabe destacar que los modelos de regresión para predecir el ROA con los indicadores de huella digital no presentaron resultados satisfactorios, pero se pudieron comparar los modelos y sacar informaciones sobre la importancia de las variables, destacando la relación entre innovación, marketing e investigación con competitividad, productividad y eficiencia empresarial, y la presencia en las redes y la relación entre bodega o marca y calidad del vino.

Palabras clave: ROA, Huella digital, bodega, vino, competitividad, regresión.

ABSTRACT

The wine sector in Spain is of great importance for the country's economy and culture. Due to its high competitiveness in the market and the new business trends for its differentiation, such as wine tourism, marketing and the development of business digitalisation. The aim of this paper is to analyse the relationship between digital footprint and competitiveness (measured as financial return on assets, ROA), in Spanish wineries using multivariate and data mining techniques.

The study used competitiveness indicators (ROA), obtained from the Iberian Balance Sheet Analysis system (SABI) and digital footprint indicators extracted from the wineries' websites and social networks. The sample initially consisted of 4183 observations (Spanish wineries) and 228 variables (ROA and digital footprint indicators) and was finally reduced to a total of 1229 observations and 107 variables (102 variables of web content "keywords" and 5 variables of presence in social networks "hrefwords").

First, data pre-processing work was carried out with outlier treatment and missing value imputation, and then different unsupervised learning techniques were implemented to explore the relationships of the set of variables and observations, as well as to find groups or clusters of companies.

In addition, the competitiveness indicator (ROA) was used as a target or response variable and the digital footprint indicators (keywords and Hrefwords) as explanatory variables, for the implementation of various regression models in different programming languages (Python and R), to compare their predictive capacity and extract relevant information about the importance of the variables.

Among the main conclusions, it should be noted that the regression models to predict ROA with the digital footprint indicators did not provide satisfactory results, but it was possible to extract information about the importance of the digital footprint indicators, highlighting on the one hand, the relationship between innovation, marketing and research with competitiveness, productivity and business efficiency, and on the other hand, the presence in the networks and the relationship between the winery or brand and the quality of the wine.

Keywords: ROA, Digital Fingerprint, winery, wine, competitiveness, regression

Índice general

Índice de figuras

Índice de tablas

1. Introducción

El sector vitivinícola en España tiene una gran importancia sobre la economía, la sociedad y la cultura del país. Siguiendo los datos publicados por la Federación Española del Vino (2022), España dispone de 930.080 Hectáreas de viñedo en 2022, lo que supone un 13% de la superficie total a nivel mundial, con una producción media anual de 36,4 millones de hectólitros de vino se posiciona como tercer productor y con un total de 4347 bodegas exportadoras que distribuyen sus vinos a nivel internacional en 189 países, siendo los principales mercados: Alemania, Estados Unidos, Reino Unido y Francia. Esto presenta una gran influencia en la economía aportando 20.330 millones de euros de valor añadido, lo que representa el 1,9% del PIB español, posicionándose como el primer exportador mundial en volumen con 2.153 millones de toneladas en 2022, equivalentes a 3.423 millones de euros.

La competitividad empresarial es la capacidad de aumentar la productividad (sean bienes o servicios) con menos recursos, generando la mayor satisfacción de los consumidores al menor costo posible de producción. Una mayor productividad redunda en una mayor capacidad de producción a igualdad de costos (Robbins & Sobral, 2009).

Por la gran competitividad del sector, las empresas centran sus esfuerzos de diferenciación en función de factores, entre los que destacan el turismo del vino o enoturismo, en el que se fomenta la cultura alimentaria, el marketing o desarrollo estratégico, con el valor de la marca y la denominación de origen como factores clave en el desarrollo empresarial.

Siguiendo el estudio de la evolución del turismo, se ha desarrollado una nueva tendencia hacia nuevos destinos de interior culturales o rurales mediante la creación de rutas turísticas, entre las que destaca la del vino. De este modo, el enoturismo ha nacido junto con la gastronomía local, promoviendo el desarrollo económico de las regiones vitivinícolas, siendo unos de los motivos para viajar.

El turismo del vino presenta experiencia sensorial al visitante, independientemente de sus conocimientos de enología puede experimentar el placer del vino, del olor y la visión de una copa de vino y de las bodegas, o abrir una botella de buen vino (López et al., 2008).

El patrimonio cultural se considera hoy un recurso turístico de gran potencial. Actualmente, aspectos de la cultura inmaterial como la alimentación se han incorporado al patrimonio

cultural, por lo que la patrimonialización de la “cultura alimentaria” (paisajes productivos, alimentos, platos, vinos y bebidas, rutas, industrias) ocurre ahora en el marco del turismo y de sus beneficios para el desarrollo local (Medina, 2017).

Además, en el actual contexto internacional de creciente globalización e inestabilidad de los mercados, en general, las organizaciones están obligadas a afrontar la redefinición de su orientación estratégica. La formulación de un conjunto claro de prioridades estratégicas se reconoce como un aspecto importante de la gestión eficaz de las organizaciones (Gómez et al., 2013).

De este modo, el estudio de las estrategias comerciales y de marketing es una cuestión clave, para las organizaciones bodegueras, uno de los activos más importantes es la marca, porque proporciona ventajas competitivas.

El valor de marca (VM) es una medida de referencia, ya que una enseña fuerte con un VM positivo tiene numerosas ventajas (altos márgenes, extensión de marca, efectividad de la comunicación y mayor preferencia de compra).

En los mercados agroalimentarios, las denominaciones de origen (DO), son las garantías oficiales más destacadas para la tipicidad de un producto. Las DO no solo representan el origen geográfico del producto, sino tradición y calidad del destino turístico. Son un factor de diferenciación que se transforma en una poderosa herramienta de comercialización. Las ventajas que reportan son diversas: seguridad jurídica, diferenciación de producto, garantía y promoción de las exportaciones. Además, constituyen un elemento identificador no solo de los productos, sino de la región vinícola. Por tanto, las DO se convierten en un elemento clave para el éxito de las zonas turísticas, ya que pueden contribuir a la creación del Valor de marca del destino (Gómez & Molina, 2012).

Por otro lado, en el sector vitivinícola, cabe destacar la importancia de la incorporación de tecnologías digitales como respuesta a la transformación digital actual en el mundo de los negocios. La introducción de tecnologías digitales permite aumentar la eficiencia y mejorar los procesos, adaptándose a los cambios del mercado, a la necesidad y a la demanda de los clientes (Alonso, 2023).

Por esta evolución o tendencia global de digitalización empresarial, el concepto del valor del producto históricamente asociado al prestigio de una marca ha ampliado mucho su capacidad de medición. Actualmente se analiza el valor del producto de otras formas, como por ejemplo a partir de las redes sociales, que han generado indicadores de opinión para

medir factores tan significativos como la satisfacción del cliente o mediante indicadores de popularidad, como la búsqueda de un producto y su intención de compra.

De este modo, las empresas y bodegas vitivinícolas han aumentado significativamente su presencia web, mediante plataformas para la venta online, sitios webs corporativos y redes sociales, incrementando una huella digital de forma considerable.

La huella digital de las empresas puede ser detectada y medida de diversas formas. Si se estudia y se analiza debidamente, es posible monitorizar en tiempo real una gran cantidad de variables e indicadores económicos y de diferentes categorías, siendo de gran valor para las organizaciones permitiendo anticiparse ante sus competidores y demostrar qué estrategias deben permanecer o implementarse en el marco de la transformación digital y tecnologías emergentes (Blázquez et al., 2018).

Por todo ello, teniendo en cuenta la importancia del sector vitivinícola en España, el enoturismo como unas de las tendencias de turismo gastronómico y rural, el desarrollo de las nuevas tecnologías para el desarrollo estratégico y de la competitividad empresarial en el sector bodeguero, se plantea la siguiente investigación, en la que se propone analizar la rentabilidad sobre los activos y la huella digital de las bodegas de vino en España.

1.1 Objetivos

El objetivo general de este trabajo es analizar la relación de la huella digital y la competitividad (medida como el rendimiento financiero sobre los activos, ROA), en bodegas de vino de España utilizando técnicas multivariantes y de minería de datos.

Para entender y poder analizar este objetivo principal, en primer lugar, se definen en el marco teórico, los diferentes conceptos a estudiar:

- Concepto de ROA (Rentabilidad sobre los activos).
- Concepto de huella digital

Posteriormente, se realiza un estudio bibliográfico introductorio mediante la revisión de trabajos previos en los que se relaciona la Rentabilidad sobre los activos (ROA), con la huella digital.

Finalmente, con el objetivo general de analizar la relación de la huella digital con la Rentabilidad empresarial sobre los activos (ROA) en las bodegas de vino de España

utilizando técnicas multivariantes y de minería de datos, se plantean los siguientes objetivos específicos:

- Construir una base de datos que contenga información de la Rentabilidad sobre los Activos (ROA) y la huella digital de las bodegas de vino de España.
- Explorar y analizar estadísticamente la relación de la huella digital con el ROA en dichas bodegas mediante modelos de regresión en diferentes lenguajes de programación (Python y R), para comparar su capacidad predictora y extraer información relevante sobre la importancia de las variables.

2. Marco Teórico

2.1 Rentabilidad económica (ROA)

La rentabilidad económica de una empresa como un objetivo económico a corto plazo que las empresas deben alcanzar, relacionado con la obtención de un beneficio necesario para el buen desarrollo de la empresa y puede ser evaluada en referencia a las ventas, a los activos, al capital o al valor accionario.

El ROA, es considerado como un indicador básico para juzgar la eficiencia en la gestión empresarial, pues es precisamente el comportamiento de los activos, con independencia de su financiación, el que determina con carácter general que una empresa sea o no rentable en términos económicos (De La Hoz Suárez et al., 2008).

Según estudios anteriores en torno a la búsqueda de un aumento de la competitividad y de la eficiencia empresarial con el objetivo de obtener rentabilidad, se encuentran diferentes indicadores económico-financieros, los de productividad y los de precio/costos (Martinez, 2022).

Los indicadores económicos, se clasifican según el cálculo de ratios obtenidos a partir de los valores encontrados en las cuentas anuales de la empresa. Entre ellos, se observan:

• Ratio de liquidez: indica la capacidad de la empresa para hacer frente a los pagos a corto plazo.

• Ratio de endeudamiento: indica la proporción de financiación ajena respecto a su patrimonio neto.

• Rentabilidad financiera (ROE): se encarga de medir el resultado generado por la empresa en relación con la inversión de los propietarios.

• Rentabilidad económica (ROA): permite conocer la evolución y los factores que inciden en la productividad del activo.

El ROA (Return on Assets - Rentabilidad sobre los Activos), es un indicador de rentabilidad con el que se refleja la capacidad que tiene una organización de generar beneficios con los recursos utilizados, es decir, indica la relación entre rentabilidad de una organización y el conjunto de sus activos. Cuanto mayor sea el valor del ROA, es para los inversionistas que la empresa genera más dinero con menor inversión (Lara & Torres, 2015).

Además, en estudios dirigidos al sector industrial se comprobó que el ROA depende en gran medida de los índices de productividad y el posicionamiento estratégico, que a la larga contribuyen a la competitividad empresarial. Se analizaron los componentes del ROA, resaltando la relación con el aumento de las ganancias, para lo cual se requiere gestión de los ingresos y los costes, que proveen mejora en el margen de ventas, de la mano de la gestión en rotación de activos operacionales (Tobón et al.,2022).

Así, para identificar los indicadores que aportan a la competitividad, destacó la importancia de la rentabilidad medida a través del ROA, como un indicador significativo de la competitividad en términos económicos.

2.2. Huella digital

La huella digital surge porque muchas actividades humanas dejan su rastro en sistemas de información digitales que pueden emplearse para generar información y conocimiento, entre otros ámbitos, a través de la producción estadística oficial (Fernández, 2018).

La huella digital en las empresas es una medida de la presencia digital y actividad en línea de la compañía. Se compone de información que se genera y se deja en línea durante la utilización de dispositivos digitales y sus interacciones (Van Dijck, 2014).

Se trata de datos de carácter masivo, que aportan información de distinta naturaleza a los convencionales y que, por tanto, requieren herramientas específicas para su tratamiento. Los humanos dejamos un rastro digital, de forma voluntaria o involuntaria, cuando realizamos actividades.

Entre las actividades con las que los seres humanos generamos huella digital están incluidas algunas tan básicas como el pago con tarjeta de crédito o el uso de los teléfonos móviles y debemos tener en cuenta que, estamos rodeados de dispositivos y sensores que permiten la monitorización de nuestra actividad.

No podemos hablar del concepto de huella digital sin mencionar Big Data, que es un concepto con mayor popularidad en los últimos años y que hace referencia a la producción de cantidades ingentes de datos (Puebla, 2018).

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.

Las empresas pueden compartir información y lanzar campañas publicitarias en diferentes plataformas digitales a través de las que se extienda la huella digital. Las páginas web permiten a las empresas la oportunidad de obtener información valiosa sobre los consumidores para desarrollar estrategias de marketing más efectivas y una interacción más cercana con los clientes. El control y la gestión de la huella digital en línea es importante para que las empresas puedan garantizar una imagen positiva (Soto, 2023).

De este modo, para medir la huella digital de una empresa, se pueden utilizar numerosas variables como el seguimiento de las redes sociales, el tráfico del sitio web, permitiendo ser rastreada para ser analizada de manera apropiada y pudiendo monitorizar en tiempo real una gran cantidad de variables económicas e indicadores de diferentes tipos (Blázquez, 2020).

3. Metodología

En el desarrollo de este trabajo se usaron diferentes modelos y métodos que se explican a continuación.

La manipulación y limpieza de los datos durante el pre-procesamiento, modelado y análisis de los datos se realizó mediante hojas de cálculo de Microsoft Excel y fragmentos de código o módulos en el lenguaje de programación Python y R y sus diferentes librerías.

3.1 Descripción de la base de datos

La base de datos inicial incluye 4183 observaciones de bodegas vitivinícolas de España, registradas en la plataforma SABI (Sistema de Análisis de Balances Ibéricos), desarrollada por Bureau Van Dijk (2017) y accesible desde la UPV. Para la selección, se aplicaron dos filtros: país (España) y actividad de la empresa (CNAE 1102: "Elaboración de vinos"). Estos filtros aseguran que la muestra se enfoque exclusivamente en bodegas españolas dedicadas a la producción de vino.

En una primera exploración, se identificaron 2353 bodegas sin dirección web. Tras una búsqueda en Google, se recuperaron 766 direcciones, lo que dejó un total de 2596 bodegas con web, eliminando 1584 sin esta información.

Además, se compone de 228 variables, agrupadas en dos bloques: indicadores de competitividad (ROA) e indicadores de huella digital (entre las que destacan las variables de contenido web y las de presencia en las redes), que se describen a continuación.

3.1.1 Indicador de competitividad: ROA

El indicador de competitividad está representado por la variable económica obtenida de la base de datos “SABI”, tomando como referencia el fundamento teórico de las dimensiones de competitividad, criterio experto y trabajos como el de Rodríguez (2022), se utilizaron las siguientes variables:

- **Retorno sobre los Activos (ROA):** Informa sobre la rentabilidad que genera la empresa mediante sus activos.

$$ROA = \text{Beneficio neto}/\text{Activos totales}$$

De esta manera, a mayor ROA, mayores beneficios con menores activos, lo ideal para maximizar el beneficio empresarial. Se expresa en porcentaje y corresponde a la media registrada en el periodo 2012-2021.

3.1.2 Indicadores de huella digital

Este grupo de variables se refiere a los contenidos y funciones disponibles en los sitios web corporativas. Se ha accedido a la aplicación Web-based Economic Indicators, dicha herramienta perteneciente a la UPV ha accedido a las páginas web de las diferentes empresas y ha recopilado los indicadores solicitados.

Diferentes estudios como el de Blázquez (2020), han investigado los indicadores de huella digital de acuerdo con la naturaleza de las variables, como sigue a continuación:

- **Variables de presencia web:**

Estás variables facilitan información web y de su propia descarga (Website o URL, Date, Sizehtml y Sizetext):

- Website/Url: es la dirección url de las bodegas.
- Date: es la fecha en la que se hizo la descarga del sitio.
- Sizehtml y Sizetext: variables que miden el tamaño en bytes de la descarga (del código HTML y del texto, respectivamente). Sirve no solo para ver el tamaño de la web, sino para saber si ha habido algún problema en la propia página web o en el proceso de descarga cuando hay tamaños anormalmente bajos.

- **Variables de contenido web:**

Cada valor obtenido en primer lugar indica la frecuencia de aparición de la palabra. Tomando como referencia el fundamento teórico de las dimensiones de competitividad, siguiendo el criterio y trabajos como el de Castro (2023), solo nos interesa saber la presencia o ausencia de la palabra, por lo que se procedió a transformarlas en binarias. Así, se generaron 203 variables dicotómicas de palabras clave (102 del nivel de concordancia exacta y 101 del nivel de Raíz de la palabra).

La lista de palabras se ha completado mediante la detección de palabras clave a dos niveles:

- Concordancia exacta (columnas que empiezan por keywords): Con un total de 102 columnas, considerando la exactitud de la palabra, es decir, si la palabra clave es productividad, la búsqueda en la aplicación ha sido “productividad”.

- Raíz de la palabra (columnas que empiezan por kwstems_es): Con un total de 101 columnas, considerando la raíz de esta, siendo, por ejemplo, si la palabra clave es productividad, la búsqueda en la aplicación fue por "produc".

- **Variables relacionadas con presencia en Redes Sociales:**

La presencia en redes sociales se ha detectado si se encontraba el nombre de la red social en algún enlace (columnas que empiezan por hrefwords). Un total de 5 variables que indican la presencia o ausencia en las diferentes redes sociales (Facebook, Instagram, Linkedin, Pinterest y Twitter). Del mismo modo en principio el valor indica la frecuencia de aparición, siendo transformadas a dicotómicas, con valor 1 en caso de que la bodega analizada se encuentre presente en la red social y 0 en caso contrario.

3.2. Pre-procesamiento de los datos

El pre-procesamiento y preparación de los datos consistió en explorar la naturaleza de las variables, detectar valores atípicos, imputar datos faltantes y eliminar variables con elevada multicolinealidad o desviación estándar igual al 0.

Para preparar la base de datos, se utilizaron diferentes procedimientos, cada uno de los cuales corresponden a los diferentes indicadores a estudiar:

- **Indicador de competitividad o rentabilidad económica (ROA):**

A continuación, para la preparación de la variable ROA de cara a una eficiente aplicación de los diferentes modelos y métodos que requieren de matrices completas para su correcta ejecución, se procedió a la imputación de los datos faltantes.

En primer lugar, se realizó un análisis de las diferentes columnas o variables “ROA”, para determinar el % de datos faltantes en cada uno de sus años y se decidió eliminar la variable perteneciente al ROA para el año 2022 al observar un 45% de datos faltantes para este año.

Posteriormente, se calcularon el nº de filas/bodegas con datos nulos para un año o más obteniendo 1427 bodegas en total. Por otro lado, se calcularon las bodegas con más de dos años o datos faltantes (bodegas con más de 2 años faltantes, más de un 20% de datos), obteniendo 1050 para luego eliminarlas.

Una vez eliminadas las observaciones con más del 20% de datos nulos, se obtuvo el archivo con menos de un 20% de datos nulos, pasando de 2596 a 1546 bodegas de vino.

De este modo, la base de datos definitiva previa a la evaluación de los posibles outliers, se redujo a un total de 1546 bodegas y 10 años (ROA anual desde 2012 a 2021).

Puesto que el objetivo era utilizar la Media del ROA para su posterior análisis como variable dependiente, y todavía se disponían datos faltantes (por debajo del 20% en todos sus individuos), se propuso imputarlos mediante la técnica de imputación MICE (Multiple Imputation by chained Equations).

- **Indicadores de Huella digital:**

A través del acceso a la aplicación, se accedió a las páginas web de las diferentes empresas y se recopilaron los indicadores solicitados mediante la técnica de webscrapping. Como cada valor indica la frecuencia de aparición de la palabra y lo que nos interesaba es saber es si cada palabra aparecía o no, se transformó a binarias adjudicando el valor 0 en los casos de frecuencia 0 y 1 en los casos de frecuencia superior a 0.

3.2.1 NIPALS PCA: Tratamiento de valores atípicos (ROA)

En base a estudios como el de Kourti & MacGregor (1996), se procedió a la detección multivariante de los outliers correspondientes. Para ello y puesto que el algoritmo NIPALS del PCA fue desarrollado de forma apropiada para la gestión de datos faltantes, se realizó la evaluación de valores atípicos utilizando los gráficos de detección de atípicos SPE y T2 de Hotelling, delimitando las observaciones que no sobrepasen el triple de los límites de confianza situados en el percentil 99% y estableciendo los umbrales para determinar los valores que estén a más de 3 desviaciones estándar de la media, siendo que:

- El Error de Predicción al Cuadrado (SPE), permite detectar valores atípicos mediante la siguiente expresión:

$$SPE_i = e_i^t e_i$$

donde e_i representa el vector de residuos de la i-ésima fila de la matriz de residuos.

El gráfico SPE comprueba si la distancia (variación de ruido) de una observación al hiperplano latente está dentro de los límites de control. Los valores del gráfico SPE que exceden los límites de control están relacionados a una ruptura de la estructura de correlación del modelo (Ferrer, 2009).

- El T2 Hotelling es una medida de la distancia desde la proyección de una observación al centro del modelo. La suma de los cuadrados tipificados (dividido por la raíz cuadrada de los λa para que su varianza sea 1) componen el T2:

$$T_i^2 = \frac{\Sigma t_{ia}^2}{S_{ta}^2}$$

Para su evaluación, se llevaron a cabo diferentes comprobaciones, mediante el uso de los diferentes gráficos (Biplot de componentes principales y los gráficos de dispersión SPE y T2 Hotelling).

El biplot proporciona una representación visual de cómo las variables originales se relacionan entre sí y con las observaciones en el espacio de las componentes principales, lo que puede ayudar en la interpretación de la estructura subyacente de los datos y la identificación de patrones o relaciones relevantes.

De este modo, al observar la presencia de outliers tras la primera iteración, se ejecutó un módulo en el que en primer lugar se calcularon SPE Y T2 Hotelling y se definieron los umbrales para determinar los valores a más de 3 desviaciones estándar de la media. Después, mediante un proceso iterativo, se encontraron y eliminaron 318 valores (bodegas) atípicos, reduciéndose la base de datos de 1546 bodegas a 1228.

Posteriormente se volvió a observar el gráfico de las componentes principales y los gráficos de dispersión SPE y T2 Hotelling, observando el resto de las observaciones dentro de los límites de control establecidos y la distribución final de los datos mediante el Histograma.

La implementación del algoritmo Nipals (PCA), se realizó mediante la clase PCA de Python, mientras que para el cálculo de SPE y T2 Hotelling se utilizaron las funciones “dot” y “sum” de la librería Numpy.

3.2.2 MICE: Imputación de datos faltantes

Una vez completado el tratamiento de valores atípicos, se procedió a imputar los datos faltantes con la librería de Python “Fancyimpute”, que ofrece varios modelos robustos de ML para la imputación de valores perdidos. Las clases “SimpleImputer” e “IterativeImputer”, más popularmente conocida como MICE (Multiple Imputation by Chained Equations) de la librería Sklearn, pueden utilizarse para la imputación (Gupta & Sedamkar, 2020).

El algoritmo MICE utiliza un método de imputación múltiple basado en ecuaciones encadenadas. Construye un modelo de imputación para la primera variable que requiere el método, utilizando una regresión múltiple con dicha variable como dependiente y el resto como explicativas; esto lo repite para cada variable con datos faltantes y lo hace secuencial aleatoria durante i iteraciones para estimar los resultados con mejor ajuste.

Una vez finalizado el tratamiento de outliers y la imputación de datos faltantes, se calculó la media del ROA para cada bodega en el dataframe con los datos imputados, para evitar así la obtención de valores sesgados comprobada en un principio tras realizar la imputación MICE sin el tratamiento de los outliers presentes.

Por último, previamente al ajuste de los diferentes modelos, se realizó la eliminación de las variables explicativas con desviación estándar igual a cero, un total de 42. Obteniendo un conjunto de datos con un total de 1228 bodegas y 166 variables explicativas (81 variables de contenido web “Keywords” y 80 “kstems” y 5 variables de presencia en redes sociales “hrefwords”). Además, se comprobó la multicolinealidad entre las variables de contenido web “keywords” (nivel de concordancia exacta) y “kwstems” (nivel de raíz de palabra) y se decidió realizar los modelos utilizando sólo las 81 variables de nivel de concordancia exacta, obteniendo la base de datos definitiva con 86 variables (81 keywords y 5 href) con la que se ajustaron los diferentes modelos.

3.3. Aprendizaje no supervisado

3.3.1 Análisis de componentes principales (PCA): Análisis exploratorio de las variables explicativas de huella digital “keywords” (PCA)

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. Pero además nos permite analizar las relaciones entre las variables.

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) y busca encontrar

una dirección de proyección p1 tal que se maximice la variabilidad de la variable latente t, cumpliéndose que las A variables latentes sean incorrelacionadas, ortogonales y combinación lineal de las variables originales (Esbensen, 2009). Es posible reconstruir la matriz X original de datos a partir de los scores tA, los loadings pA y la matriz de residuos E:

$$X = t_1 p_1' + \ldots + t_A p_A' + E$$

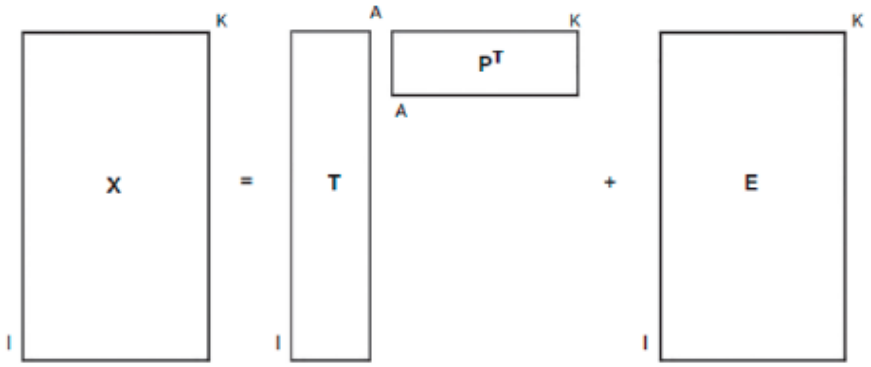


Figura 1: Descomposición de la matriz X

La Figura 1 representa la descomposición de la matriz de datos inicial X en los scores (t), los loadings (p) y la matriz de residuos E, que contiene toda la información que no se encuentra contenida en las componentes principales extraídas.

Puesto que, en relación con las variables de huella digital, se realizaron diferentes comparaciones en los modelos de regresión lineal para evaluar posibles problemas de multicolinealidad, se observó que los *kwstems* proporcionan prácticamente la misma información que los *keywords*, pues las raíces de las palabras coinciden, y se optó por no considerar los *kwstems* debido a las posibles complicaciones de colinealidad que podrían surgir al incluir ambos tipos de variables. El modelo PCA fue construido con el principal objetivo de explorar el comportamiento de las variables explicativas de huella digital "keywords" evaluadas dentro del sub-espacio de variables latentes.

3.3.2 Análisis de clasificación: Clustering

El análisis clustering o de conglomerados, es una técnica de aprendizaje no supervisado de clasificación que se ocupa de agrupar las observaciones de tal forma que las

observaciones en el mismo grupo de clasificación (cluster) son más similares entre sí que con las observaciones de otros grupos o clusters.

Con estas técnicas se espera pues encontrar patrones de agrupación de las observaciones a diferencia del PCA que agrupa variables.En esta investigación se han utilizado dos técnicas de clustering: Clustering difuso (C-means) y Clustering particional (k-means).

Cabe destacar que el análisis clustering ha sido realizado únicamente con las variables explicativas de huella digital “keywords”, para obtener grupos o clusters de empresas con características similares en este bloque de variables, es decir, para determinar si las bodegas se pueden agrupar en base a palabras incluidas en sus webs.

En los análisis del clustering particional (k-means), cada individuo pertenecerá a un solo clúster. Mientras que, el algoritmo del clustering difuso (C-means) se caracteriza por el hecho de que los individuos pueden pertenecer a más de un grupo. Esta asociación a cada grupo se mide por un nivel de pertenencia (partición difusa), el cual indicará la fuerza de relación entre el individuo y un grupo en particular.

Para implementar la técnica C-means, se utilizó el paquete de R llamado “e1071”, mientras que para el K-means, la librería utilizada fue “Factoextra”.

3.3.2.1 Evaluación, Selección y Validación de la tendencia y del número de clusters:

Puesto que los algoritmos no realizan la determinación del número de grupos o clusters, hay que definirlo para inicializar el algoritmo y entrenar el modelo. Para ello se utilizan ciertos indicadores para la evaluación, selección y validación del número de K a utilizar. En nuestro caso, utilizaremos uno de ellos, el método del codo, el cual veremos a continuación.

Antes de aplicar un método de clustering a los datos es conveniente evaluar la tendencia de agrupamiento o cluster para ver si hay indicios de que realmente existe algún tipo de agrupación entre ellos. Este proceso se conoce como “assessing clúster tendency” y puede llevarse a cabo mediante test estadísticos (Hopkins statistics) o de forma visual (Visual assessment of clúster tendency) (De la Fuente, 2011), los cuales son explicados a continuación y han sido implementados en este estudio para valorar la tendencia al agrupamiento.

Estadístico de Hopkins

En cuanto al estadístico Hopkins, cabe destacar que permite evaluar la tendencia de clustering de un conjunto de datos mediante el cálculo de la probabilidad de que dichos datos procedan de una distribución uniforme, es decir, estudia la distribución espacial aleatoria de las observaciones (Larico, 2021).

Por ejemplo, sea D un conjunto de datos reales. La estadística de Hopkins se puede calcular como sigue:

a) Seleccionar una muestra de n puntos ($p_1,\ldots,p_n$) de D.

b) Para cada pi ∈ D encontrar su vecino más cercano p_j y calcular

$$dist(q_{i,}q_j) = x_i$$

c) Generar un conjunto de datos simulados de una distribución uniforme con n puntos ($q_1,\ldots,q_n$) y con la misma dispersión que los datos originales D.

d) Para cada q_i encontrar su vecino más cercano q_j en D y calcular

$$y_i = dist(q_{i,}q_j)$$

e) Calcular el estadístico mediante la siguiente fórmula:

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

Donde $\sum_{i=1}^{n} y_i$ representa la suma de distancias entre cada par de datos de un conjunto de datos simulados provenientes de una distribución uniforme, y $\sum_{i=1}^{n} x_i$ representa la suma de distancias entre cada par de datos del conjunto original de datos.

La hipótesis nula y alternativa está definidas como lo siguiente:

H_0: D está distribuido de forma uniforme (es decir: no hay agrupaciones significativas):

H_1: D no está distribuido de forma uniforme (es decir: el conjunto de datos contiene agrupaciones significativas)

Según Hopkins (1954), se puede realizar la prueba estadística de Hopkins de forma iterativa, utilizando 0.5 como límite para rechazar la hipótesis alternativa. Es decir, si H <0.5, entonces es poco probable que D tenga conglomerados estadísticamente significativos.

Mientras que, si el valor de H está cerca de 1, entonces podemos rechazar la hipótesis nula y concluir que el conjunto de datos D es significativamente agrupable.

Visual assessment of cluster tendency (VAT)

El método VAT permite evaluar visualmente si los datos muestran indicios de algún tipo de agrupación.

En este procedimiento, en primer lugar, se calcula una matriz de distancias euclídeas entre todos los pares de observaciones.

Posteriormente, se reordena la matriz de distancias de forma que las observaciones similares están situadas cerca unas de otras (ordered dissimilarity matrix).

Finalmente se representa gráficamente la matriz de distancias ordenada, empleando un gradiente de color para el valor de las distancias. Si existen agrupaciones subyacentes en los datos se forma un patrón de bloques cuadrados.

Tanto el estadístico de Hopkins, como el método VAT se han calcularon mediante la ejecución de las funciones “Hopkins_statistics” y “VAT” en diferentes fragmentos de código Python.

Tras evaluar la tendencia al agrupamiento, determinar el número óptimo de clústeres es uno de los pasos más complicados al implementar métodos de análisis clúster. No existe una forma única de averiguar el número adecuado de clústeres, pero existen estrategias principales como los métodos de Elbow o método del codo (Niño, 2020), el cual hemos implementado en este trabajo para la determinación del número de clusters, mediante la utilización.

Método del codo:

El método del codo (Elbow method) utiliza el valor de WCSS (suma de la distancia al cuadrado de cada una de las observaciones a su centroide).

Obtenidos los valores de WCSS, se representan linealmente según un rango alto de valores K. En el gráfico se aprecia un cambio brusco en la línea continua que une cada punto. A ese cambio brusco en la gráfica se le denomina “codo”. El punto en el eje y donde se observa el codo es el número de clusters K (Lopez, 2019)

Para la implementación del método del codo se utilizaron las clases “KMeans” y “KElbowVisualizer” de Python y la función get_clust_tendency del paquete Factoextra de R, donde el visualizador K-Elbow selecciona el número óptimo de conglomerados para la agrupación K-means.

3.3.2.2 Clustering difuso (C-Means)

La minería de datos permite manejar y clasificar grandes cantidades de datos, una de las tareas típicas de la minería de datos es el clustering o agrupamiento, Fuzzy C-Means (FCM) es una técnica difusa de minería de datos para el clustering que se basa en el algoritmo clásico C-Means. Fuzzy C-Means asigna a cada dato un grado de pertenencia dentro de cada cluster y, como consecuencia, un dato puede pertenecer parcialmente a más de un grupo. A diferencia del algoritmo K-means clásico que trabaja con una partición nítida, FCM realiza una partición suave del conjunto de datos, en la que los datos pertenecen en cierto nivel a todos los clusters, es decir, a diferencia del k-means, que asigna cada observación exactamente a un grupo, el fuzzy-clustering permite cierto grado de incertidumbre en la asignación de observaciones a los clusters, por lo que un punto puede pertenecer a más de un cluster.

Del mismo modo que el K-means, C-Means trabaja con aquellos objetos que pueden ser representados en un espacio n-dimensional con una medida de distancia definida, a fin de minimizar la siguiente función objetivo:

$$SSE(M,C) = \Sigma_{i=1}^{n} \Sigma_{i=1}^{k} m_{ij}^{\phi} {d_{ij}}^2$$

Donde SSE(M,C) es la suma del cuadrado de los errores dentro de las clases, m es la matriz (n × k) de pertenencia a los grupos y cumple que mij = 1 si el elemento i pertenece totalmente al clúster j y mij = 0, en caso contrario; C es la matriz (k × p) de centro de las clases, donde p es el número de componentes del espacio, ϕ es el grado de imprecisión de la solución (del inglés fuzziness exponent) y dij 2 es el cuadrado de la distancia entre el elemento i y el centro representativo del clúster j.

La técnica fuzzy clustering fue originalmente introducida como una mejora sobre los métodos de clustering (Bezdek et al. 1981).

El algoritmo más usado para esta técnica es el fuzzy C-Means (Pal et al,. 1996).

3.3.2.3 Clustering particional (K-Means)

El algoritmo K-Means es uno de los métodos de clasificación o agrupamiento más populares en Minería de datos. Su objetivo es encontrar una partición de datos descritos por valores numéricos, que consiste en grupos representados por un centro. Para la determinación del número de grupos a crear por el algoritmo, el usuario tiene que definir

previamente el valor del parámetro k, el cual indica cuántos grupos se formarán en la partición resultante del conjunto de los datos. El proceso de este algoritmo se basa en la iteración hasta encontrar la mejor partición de los datos, afinando la posición de los centros representativos en el espacio de objeto, es decir, encontrar a los individuos que integren a los individuos más parecidos a ellos (Franco, et al,. 2021).

Con el fin de calcular la distancia que existe entre dos objetos distintos y así identificar a los objetos más parecidos para crear los grupos, el algoritmo k- means puede emplear funciones de distancia. La distancia euclidiana es una de las más empleadas en la literatura y preferida en el proceso de este algoritmo. La distancia euclidea mide la distancia más corta entre dos puntos y se calcula usando mediante la siguiente ecuación:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2}$$

Donde se indica que los objetos están descritos por i valores de atributos descriptores y que todos son considerados para poder obtener la distancia entre X y Y. Entre más pequeño el valor de distancia más cercanos son los objetos comparados y por ende más similares.

El algoritmo K-means intenta optimizar la suma de cuadrados dentro de cada cluster. Busca minimizar la siguiente función objetivo:

$$SSE = \Sigma_{i=1}^{k} \Sigma_{x \epsilon ci}^{d(x,c_i)^2}$$

Donde SSE es la suma del cuadrado de los errores, Ci es el i-ésimo clúster de la partición, k es el número de clústeres y d (x,Ci) es la medida de disimilitud o distancia entre el elemento x y el clúster Ci:

A continuación, en la Figura 2 vemos el funcionamiento del algoritmo K-means:

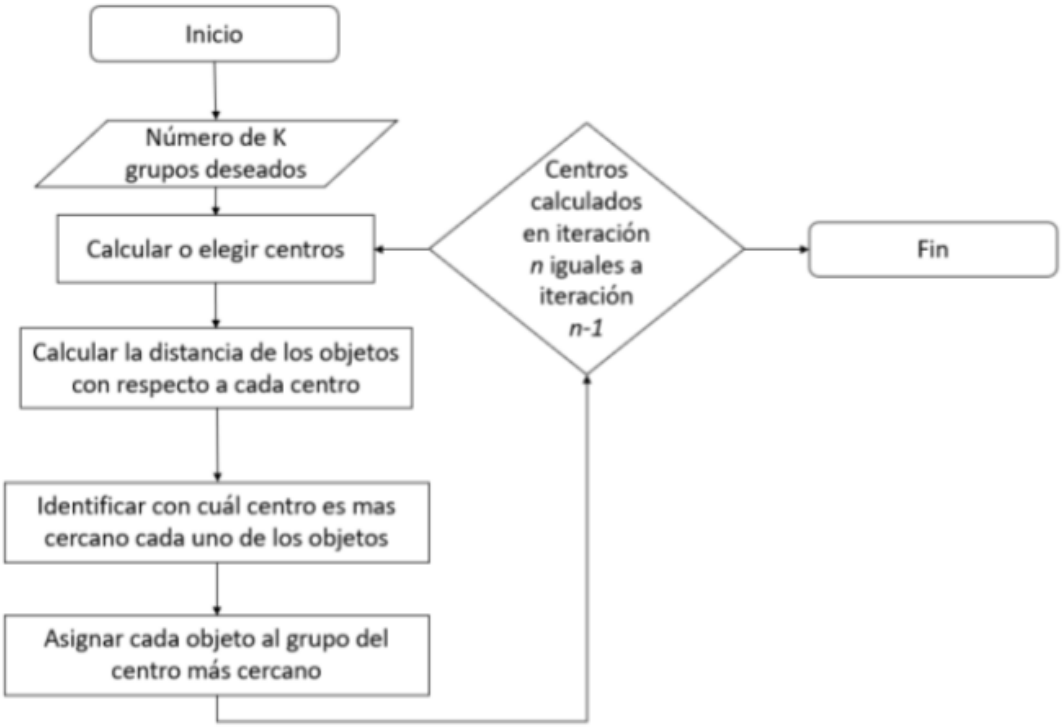


Figura 2: Diagrama funcionamiento algoritmo K-means

La ejecución de ambos métodos de clustering se realizó mediante la utilización de los paquetes “e1071” y “cluster” en R.

3.4 Aprendizaje supervisado

El aprendizaje supervisado se da cuando se tiene variables explicativas (x) y una variable dependiente (Y) a explicar, y se utiliza un algoritmo o modelo para aprender la función de mapeo (las normas) de las variables explicativas y la dependiente: Y = f (X). El modelo va realizando predicciones de manera iterativa con los datos que tiene. El aprendizaje termina cuando el modelo consigue obtener unos resultados aceptables.

A diferencia del aprendizaje no supervisado, el aprendizaje supervisado se da cuando se conoce a priori el objetivo, es decir, cuando para los individuos se conoce alguna variable respuesta asociada a ellos. El objetivo es aproximar la función de mapeo mediante un “entrenamiento”, para crear un modelo que ajusta los datos, de manera que cuando se tengan nuevos datos de entrada (x) se pueda predecir la variable dependiente (Y) para esos nuevos datos.

Los problemas de aprendizaje supervisado pueden agruparse en problemas de regresión y clasificación. Cuando la variable respuesta es numérica, se utilizan modelos de regresión y cuando es categórica los de clasificación. El aprendizaje termina cuando el algoritmo consigue unos resultados aceptables (Serra, 2020).

Algunos ejemplos populares de algoritmos supervisados de aprendizaje automático son:

- Regresión lineal para problemas de regresión.
- Bosque aleatorio (Random Forest) para problemas de clasificación y regresión.
- Soporte de máquinas de vectores (Support Vector Machine) para problemas de clasificación y regresión.

Durante el desarrollo del presente Trabajo Final del Máster, se aplicaron diferentes análisis univariantes y bivariantes. Luego se describirán los análisis presentados en la sección de resultados.

Por otro lado, en cuanto a las variables utilizadas para las diferentes técnicas de análisis:

Variables cuantitativas:

- **Variable dependiente o respuesta (y):**

a) Rentabilidad económica (ROA) en bodegas de vino de España:

En esta variable se esperaba obtener la rentabilidad de las bodegas vitivinícolas españolas dentro de la muestra. Saber si tuvieron resultados positivos o negativos, entre qué rangos de resultados se encontraban, entender si tuvieron un comportamiento más homogéneo o heterogéneo en cuanto a la distribución de los datos, o si eran simétricos o asimétricos.

Esto podría darnos una idea de si realmente estas empresas están llevando o no una correcta gestión y si los accionistas estaban recibiendo rendimientos.

Con la rentabilidad, podría darnos una breve vista sobre qué bodegas tienen mayores beneficios para ser más competitivas.

El método empleado en este estudio, en primer lugar, permitió un análisis exhaustivo de la variable dependiente planteada en este caso (“Roa Mean”) utilizando tanto enfoques estadísticos como visuales.

- **Variables independientes o explicativas (x):**

b) Variable de contenido web

c) Variables de presencia en redes sociales

Como hemos comentado en el apartado 3.1 (descripción de la base de datos), estas variables se transformaron a binarias, con valor 1 si la palabra está presente en la página web y 0 en caso contrario para las de contenido web y con valor 1 si la bodega analizada está presente en la red social y 0 en caso contrario, en las de presencia en redes sociales.

En nuestro caso, dado un conjunto de observaciones donde se conoce el valor de la variable respuesta o dependiente, el objetivo es construir un modelo para predecir nuevos casos.

Puesto que se propone predecir una variable respuesta continua, con un valor real y numérico o cuantitativo (ROA = Beneficio Neto/Total Activos), libre al completar en el preprocesamiento de los datos la limpieza y eliminación de datos faltantes, y se disponen de una serie de variables independientes o explicativas de las cuales conocemos el dato para las diferentes observaciones o individuos (Bodegas de Vino de España), se ha propuesto realizar diferentes modelos de regresión, para su comparación:

- Regresión lineal
- Arbol de regresión
- Bosque aleatorio (Random Forest)
- Máquinas de soporte vectorial (Support Vector Machines: SVM)

Para la realización de los diferentes modelos, además, se ha propuesto su confección tanto en lenguaje de programación Python, con la ejecución de los diferentes módulos o fragmentos de código en Visual Studio Code (entorno de desarrollo integrado “IDE” gratuito y de código abierto desarrollado por Microsoft) como en el lenguaje de programación R, mediante el desarrollo de los diferentes fragmentos de código en RStudio (IDE que proporciona una interfaz amigable que facilita la escritura, ejecución y depuración de código R). El código implementado con sus diferentes librerías se puede encontrar en Github (ver anexo).

3.4.1 Modelos de regresión

3.4.1.1 Regresión Lineal Múltiple

En la Regresión Lineal Simple se estudia la influencia que ejerce una variable explicativa (x) en una variable dependiente (y), mientras que, con la Regresión Lineal Múltiple, se estudian más de una variable explicativa (Abuín, 2007).

De este modo, se obtiene la ventaja de utilizar más información en la elaboración del modelo y, por tanto, realizar estimaciones más precisas. Así al utilizar más de una variable explicativa, surgen algunas diferencias frente a la Regresión Lineal Simple.

Siguiendo estudios como el de Frutos (2021), igual que en la Regresión Lineal Simple, se plantea que la variable dependiente (y) es una combinación lineal de las variables explicativas (x) y una perturbación aleatoria (u):

$$yi = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

El modelo anterior también se puede representar de forma matricial como:

$$Y = X\beta + U$$

Es importante resaltar que la primera columna de la matriz X está compuesta de unos, que multiplican a β_0 (término independiente).

El modelo expuesto anteriormente representa a la población y sus parámetros son desconocidos, con lo que es necesario extraer una muestra con la que poder estimarlos:

$$\hat{Y} = X\hat{\beta}$$

Este es el modelo estimado y es el modelo que se obtendrá en este trabajo para llevar a cabo la predicción deseada. En este caso no se trata de una recta, sino de un hiperplano, puesto que cuenta con varias variables explicativas.

Al igual que con Regresión Lineal Simple, la diferencia entre el valor real (Y) y el valor estimado (Ŷ) viene definida por la estimación del error o residuo, de forma que:

$$\widehat{U} = Y - \hat{Y}$$

De esto se deduce que la variable Y, además de por el modelo expuesto anteriormente, también viene dada como*:*

$$Y = X\hat{\beta} + \widehat{U}$$

• Estimación de los parámetros: Método de mínimos cuadrados

Al igual que con Regresión Lineal Simple, se pretende calcular un hiperplano de regresión (en vez de una recta de regresión) tal que se minimice la suma de los cuadrados de los residuos:

$$\Sigma(\hat{u}_i)^2 = \hat{U}^T * \hat{U} = (Y - X\beta)^T * (Y - X\beta)$$

Siguiendo una línea similar a la seguida con Regresión Lineal Simple, si se minimiza la expresión anterior derivándola respecto de βˆ, se puede llegar a la expresión:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Por otro lado, se aplicó uno de los métodos de selección escalonada, que básicamente son tres:

a) eliminación progresiva (forward)

b) eliminación regresiva (backward)

c) el método por pasos (stepwise)

Este último es la combinación de los dos primeros más una modificación, coloquialmente se le llama el procedimiento de eliminación por pasos (Navarro, 2009).

Siguiendo el principio de parsimonia y con el objetivo de la modelización estadística, de encontrar la combinación de variables más simple que tenga mayor poder explicativo, se utilizó el método stepwise basado en criterios de información teóricos (AIC).

El AIC (Akaike Information Criterion) enfatiza la bondad del ajuste del modelo, su idea clave es

la de penalizar un exceso de parámetros ajustado.

Se trata de un estimador muestral de $E[lnf(X|\theta)]$, esperanza de la log-verosimilitud, que viene dado por la expresión general $AIC(k)=-2ln\mathcal{L}[\theta(k)]+2k$, donde $\mathcal{L}[\theta(k)]$, es la función de verosimilitud de las observaciones, $\theta(k)$ es el estimador máximo verosímil del vector de parámetros θ y k es el número de parámetros independientes en el modelo.

El primer término del AIC puede interpretarse como una medida de la bondad de ajuste, mientras que el segundo es una penalización, que crece conforme aumenta el número de parámetros, según el principio de parsimonia (González et al., 2016).

Con el tiempo se ha demostrado que la regresión lineal múltiple es una herramienta que permite predecir ciertos acontecimientos que suceden en la realidad, pero que dicho modelo requiere cumplir ciertas hipótesis asociadas al modelo, para que tenga sentido y podamos utilizarlo (Hernández, 2019).

Las hipótesis asociadas al modelo de regresión múltiple son:

1. **Normalidad de la variable dependiente**
2. **Homocedasticidad**
3. **Autocorrelación**
4. **Multicolinealidad**

Trabajar con datos distribuidos normalmente, "con una forma de campana", es muy relevante, ya que, si los datos no están distribuidos normalmente, es inviable llegar a conclusiones con altos niveles de fiabilidad o hacer buenas interpretaciones a través de métodos paramétricos. Según estudios como el de Razali y Wah (2011), los tres procedimientos más comunes para evaluar la normalidad del error de una muestra aleatoria con un determinado tamaño de muestra son los siguientes:

- Métodos numéricos (Coeficiente de asimetría y Curtosis)
- Métodos gráficos (Histogramas y Diagrama de caja)
- Pruebas de normalidad, entre las que destacan:
    - Shapiro-Wilk
    - Kolgomorov

En este estudio se realizaron los 3 métodos más comunes para evaluar la normalidad y realizar una breve interpretación de los resultados obtenidos. Tras la obtención de los resultados descriptivos y con el apoyo de los métodos gráficos, se observó que el valor mínimo -26.4961, se encontraba muy alejado del resto, pero no fue eliminado como outlier en el tratamiento anterior, por lo que se realizaron los análisis incluyéndolo y excluyéndolo para valorar su importancia y determinar su posible eliminación. Así que, se decidió eliminarlo al observar ligeras mejoras en las medidas de bondad de ajuste.

Posteriormente, se procedió a la elaboración del modelo predictivo de Regresion Lineal Multiple mediante el entrenamiento del modelo y para ello se utilizó la clase LinearRegression de la librería Scikit-Learn mediante el uso del método Fit en Python y la función lm () en R de la librería Caret, a las cuales se les pasa los datos del conjunto de

entrenamiento para calcular los coeficientes estimados mediante mínimos cuadrados y con estos se elabora el hiperplano estimado del modelo. Estos coeficientes, junto con otros resultados y métricas, se pueden obtener de la función summary() en R, mientras que en Python se obtienen con el atributo “coef_” de la clase LinearRegression.

Además, en ambos lenguajes se realizaron las pruebas de las diferentes hipótesis asociadas al modelo de regresión lineal multiple, mediante el uso de las librerías SciPY y Statsmodel con sus diferentes funciones en Python, Shapiro(), que pone a prueba la hipótesis nula de que los datos proceden de una distribución normal, Kstest(), que realiza la prueba de Kolmogorov – Smirnov, durbin_watson() que plantea la hipótesis de la independencia de los errores para evaluar la autocorrelacion, y variance_inflation_factor() que calcula el factor de inflación de la varianza para medir la multicolinealidad entre las variables predictoras del modelo.

Mientras que, en R, del mismo modo, Shapiro.test(), ks.test(), car:vif() y finalmente, el comando step () que selecciona automáticamente el modelo mejor según el menor valor del estadístico criterio de información de Akaike (AIC) que busca un equilibrio entre el modelo que mejor ajusta y a la vez es el más simple, siguiendo el principio de parsimonia.

3.4.1.2 Arbol de regresión

Los árboles de regresión “classification and regression trees (CART)”, son una alternativa de regresión y se considera un método de fácil interpretación de sus resultados, siendo su principal ventaja. Un CART establece sucesivas particiones del conjunto datos de manera que los subconjuntos resultantes sean lo más homogéneos posible.

La variable para predecir es de naturaleza numérica y realiza un conjunto de condiciones organizadas en una estructura jerárquica. Está formado por nodos de decisión que realizan una pregunta sobre el valor de un atributo o particiones de una variable continua. Las diferentes respuestas que se pueden obtener generan otros nodos hijos. De este modo el árbol se construye de forma que se crea un modelo que pronostica valores de una variable dependiente basada en valores de variables explicativas. Cada nuevo caso se predice siguiendo el camino del árbol y la predicción es la media de las observaciones de dicho nodo (Therneau et al., 2019).

Siguiendo algunos de los estudios planteados de los árboles de clasificación y regresión CART (Breiman et al., 1984), se presentó un gran interés científico en la utilización de este

método, dada su fácil implementación en todo tipo de problemas y su fácil interpretación de resultados.

Existen distintos métodos para desarrollar un árbol, pero en general se busca maximizar la capacidad predictiva determinando sobre qué atributos o variables hay que preguntar y en qué secuencia.

El método “classification and regression trees (CART)” se realiza siguiendo el siguiente proceso:

- Creación o construcción del árbol máximo.
- Realización de la poda del árbol.
- Selección del árbol óptimo por validación cruzada.

Para la implementación del árbol de regresión se utilizó la clase DecisionTreeRegressor del módulo Tree de la librería Scikit-Learn mediante el uso del método Fit en Python y la función rpart() de la librería rpart en R.

3.4.1.3 Bosques aleatorios (Ramdom Forest)

El bosque aleatorio (RF) ofrece una combinación única de precisión de predicción e interpretabilidad del modelo entre los métodos de aprendizaje automático más conocidos.

Además, Random Forest tiene la ventaja de que también permite estudiar la importancia de las variables. Utiliza los índices de “Mean Decrease Accuracy (MDA)”, el cual expresa cuánta precisión pierde el modelo al excluir cada variable y el indice de “Gini”, que mide cómo cada variable contribuye a la homogeneidad de los nodos y hojas en el bosque aleatorio resultante.

Cuanto mayor sea el valor de “MDA” o la puntuación de “Gini”, mayor será la importancia de la variable explicativa correspondiente en el modelo (Breiman, 2001).

De tal manera que, si una variable tiene una alta importancia, significa que es un fuerte predictor de la variable objetivo y puede usarse para dar peso el proceso de selección de variables (Archer & Kimes 2008).

Para la realización del modelo Random Forest en Python se utilizó la clase RandomForestRegressor de la librería Scikit-Learn mediante el uso del módulo ensemble. Mientras que en R se utilizó el paquete randomForest, que permite aplicar el método y proporciona además de la clasificación de los objetos, información adicional de interés

como la medida de la importancia de las variables predictoras y una medida de la estructura interna de los datos la proximidad de unos a otros (Liaw & Wiener, 2002).

3.4.1.4 Máquinas de soporte vectorial (SVM)

Una máquina de soporte vectorial "SVM" (Support Vector Machines) son un conjunto de algoritmos de aprendizaje supervisado usados para la clasificación de datos. Se basa en la idea de encontrar un hiperplano que divida mejor un conjunto de datos en dos clases, procurando separar de forma óptima los puntos de una clase de la otra.

Los vectores de soporte son los puntos de datos más cercanos al hiperplano (support vector).

El SVM es una técnica de "caja negra" que solo funciona como predictora, pero no permite estudiar la importancia de las variables dentro del funcionamiento del modelo.

Para la implementación del modelo SVM en Python se utilizó la clase SVR del módulo SVM de la librería Scikit-Learn. Mientras que, en R, se utilizó la función svm() de la librería e1071.

3.5 Medidas de bondad de ajuste para la evaluación de modelos de regresión

En los modelos de regresión, el mejor modelo es el que minimiza la distancia entre observados y predichos. Para su evaluación, existen varias medidas de bondad de ajuste, las cuales se pueden clasificar como sigue:

Absolutas: MSE, RMSE, MAE

Error Cuadrático Medio (MSE):

$$MSE = \frac{1}{N}\,\Sigma_{i=1}^{N}\,(\hat{y}_i - y_i)^2$$

El MSE calcula el promedio de los cuadrados de las diferencias entre las predicciones y las observaciones. Cuanto menor sea el MSE, mejor será el ajuste del modelo. Se calcula mediante la siguiente fórmula:

Raíz del Error Cuadrático Medio (RMSE):

$$MSE = \sqrt{MSE}$$

El RMSE es simplemente la raíz cuadrada del MSE. Proporciona una medida del error en la misma unidad que la variable dependiente.

Error Absoluto Medio (MAE):

$$MAE = \frac{1}{N}\Sigma_{i=1}^{N} |y_i - y_i|$$

El MAE calcula el promedio de las diferencias absolutas entre las predicciones y las observaciones. Al igual que el MSE, cuanto menor sea el MAE, mejor será el ajuste del modelo.

Relativas: NMSE, NMAE, MAPE

Error Cuadrático Normalizado (NMSE):

$$NMSE = \frac{\frac{1}{N}\Sigma_{i=1}^{N}(\hat{y}_i - y_i)^2}{\frac{1}{N}\Sigma_{i=1}^{N}(\hat{y} - y_i)^2}$$

El NMSE normaliza el MSE dividiendo por la varianza de las observaciones. Esto proporciona una medida de la precisión del modelo en relación con la variabilidad inherente de los datos.

Error Absoluto Normalizado (NMAE):

$$NMAE = \frac{\frac{1}{N}\Sigma_{i=1}^{N}|\hat{y}_i - y_i|}{\frac{1}{N}\Sigma_{i=1}^{N}|\bar{y} - y_i|}$$

El NMAE normaliza el MAE dividiendo por la desviación media absoluta de las observaciones.

Error Porcentual Absoluto Medio (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| x\ 100$$

El MAPE calcula el promedio de los errores porcentuales absolutos entre las predicciones y las observaciones. Se expresa como un porcentaje y proporciona una medida de la precisión relativa del modelo.

3.6 Validación de modelos de regresión

3.6.1 Hold out

Existen diversos métodos para validar los modelos de regresión, como son la comparación de los parámetros ajustados con los obtenidos mediante modelos físicos teóricos o con simulaciones, utilizar nuevos conjuntos de datos conocidos para comparar con los ajustados a los datos originales o el uso de técnicas de validación cruzada (CV), con K-pliegues. En el método CV los datos se dividen aleatoriamente en k grupos, donde un grupo se utiliza para el conjunto de validación del modelo (test), y con los otros k-1 restantes se construye un modelo de entrenamiento (train), que se utiliza para predecir el resultado de los datos de validación, iterando este proceso k veces.

El método hold-out es el más sencillo de los distintos métodos de validación. Como hemos comentado anteriormente, este separa el conjunto de datos disponibles en dos subconjuntos, uno utilizado para entrenar el modelo y otro para realizar el test de validación. De este modo, se construye un modelo únicamente con los datos de entrenamiento. Con el modelo creado, se generan predicciones del conjunto de datos reservados para realizar la validación que se comparan con su valor real. Los estadísticos obtenidos con los datos del subconjunto de validación son los que nos dan la validez del modelo empleado en términos de error (Perez et al., 2015).

3.6.2 Hold Out Repetido

Una aplicación alternativa de este método consiste en repetir el proceso hold-out, tomando distintos conjuntos de datos de entrenamiento (aleatorios) un determinado número de veces, de manera que se calculan los estadísticos de la regresión a partir de la media de los valores en cada una de las repeticiones.

El método de validación “Hold-out repetido”, se realiza dividiendo aleatoriamente los datos en un conjunto de entrenamiento y un conjunto de prueba, y entrenando el modelo en el conjunto de entrenamiento. Después, se evalúa el rendimiento del modelo en el conjunto de prueba. En este caso, este proceso se repite varias veces, utilizando diferentes divisiones aleatorias de los datos cada vez.

Una de las grandes ventajas del hold out repetido es que facilita una estimación del rendimiento del modelo en datos no utilizados para la construcción del modelo, de modo

que obtenemos un criterio de evaluación más óptimo que el hold out, ya que permite promediar las medidas de bondad de ajuste del total de iteraciones realizadas en el entrenamiento y en la validación.

La proporción para la segmentación de los datos fue de un 75% para entrenamiento y un 25% para validación. Las iteraciones usadas fueron 100 para cada modelo de regresión.

Como hemos comentado anteriormente, utilizamos los resultados obtenidos de las medidas de bondad de ajuste para la evaluación de los diferentes modelos. En el hold out obtenemos los resultados de las métricas en una tabla y comparamos los valores de los resultados de cada modelo, mientras que en la técnica de hold out repetido, calculamos la media de los resultados del total de iteraciones.

De este modo, para el hold out repetido, resulta más práctico la visualización gráfica de la media de alguna de las medidas de bondad de ajuste y su comparación entre los valores obtenidos para cada uno de los modelos.

3.6.3 Test de ANOVA

La técnica estadística por excelencia para comparar más de dos poblaciones (en este caso los resultados de los diferentes modelos) es la técnica de análisis de la varianza (ANOVA).

Por lo tanto, para evaluar si existen diferencias significativas entre medias se realizó un ANOVA, ya que este método permite comparar las diferencias entre las medias de las medidas de bondad de ajuste de más de dos modelos y determinar si estas diferencias son estadísticamente significativas.

Sólo cuando el resultado de un ANOVA es significativo, sugiriendo que hay diferencias entre los grupos, es lícito averiguar dónde reside la o las diferencias. Esto se realiza con la aplicación de una de diversas pruebas, también pruebas de hipótesis, que se conocen como procedimientos de comparación múltiple o test post hoc.

Dichos procedimientos se hacen después que un ANOVA ha dado un valor de F significativo y los siguientes se usan para grupos de igual tamaño, destacando el método de Bonferroni, el de comparación múltiple de Dunn, el de Scheffé, el procedimiento de Newman-Keuls, el procedimiento de Dunnet y la prueba HSD de Tukey, la cual hemos implementado en nuestro trabajo al obtener el resultado del ANOVA significativo, mediante el método “TukeyHSD” en R.

Prueba HSD de Tukey (diferencia honestamente significativa) permite la comparación entre todos los pares de medias. Sería el procedimiento más potente y exacto para usar en estas circunstancias y permite el cálculo de intervalos de confianza. Se dice que la prueba Tukey es exacto en el ajuste de α a 5% mientras que el Bonferroni es aproximado, en el ajuste de α a 5% o menos (Boqué & Maroto, 2004).

4. Resultados

4.1 Exploración de datos faltantes en las variables anuales ROA:

En primer lugar, se realizó un análisis de las diferentes columnas o variables “ROA”, para determinar el % de datos faltantes en cada uno de sus años, como sigue en la tabla 1:

AÑO ROA	% DATOS FALTANTES
2022	45
2021	27
2020	21
2019	23
2018	23
2017	19,5
2016	17
2015	15,5
2014	15,5
2013	14
2012	12

Tabla 1: Exploración de datos faltantes del ROA en sus diferentes años. Elaboración propia.

Puesto que se observó un 45% de datos faltantes para el año 2022, se decidió eliminar la variable perteneciente al ROA para este año.

4.2 Exploración de datos con valores atípicos (“Outliers”) en las variables anuales ROA:

Una vez eliminados aquellas variables con un alto % de datos faltantes, previamente a la imputación MICE y el cálculo de la media del ROA, se evaluó la existencia de outliers en las diferentes variables ROA para cada uno de los años de estudio, que pudieran influir en

la media. Como sigue en la Tabla 2, donde se aprecian valores muy elevados en algunos años (2019 y 2017, por ejemplo):

Variable	Min	P25	Median	Mean	P75	Max
ROA 2021	-90.453	-0.780	1.177	1.800	4.331	99.983
ROA 2020	-93.751	-2.826	0.254	-0.186	2.602	100.743
ROA 2019	-32475855.000	-1.000	1.000	-21500.000	4.000	90.000
ROA 2018	-16454.788	0.000	1.362	-8.827	4.592	91.797
ROA 2017	-13500055.000	0.0000	1.000	-8851.000	5.000	44.000
ROA 2016	-190.050	-0.366	1.037	1.178	4.023	63.622
ROA 2015	-213.835	-0.695	0.948	1.070	3.722	56.545
ROA 2014	-107.241	-1.346	0.602	0.577	2.851	110.566
ROA 2013	-142.750	-1.893	0.445	-0.594	2.156	71.200
ROA 2012	-902.952	-2.473	0.365	-1.922	2.222	52.272

Tabla 2: Análisis descriptivo de las variables ROA anuales. Elaboración propia.

Se visualizó mediante el histograma de cada variable (ROA anual) en la Figura 3, el histograma de la media ROA (Figura 4) y el diagrama de cajas y bigotes de la media ROA (Figura 5), la presencia de datos anómalos, como sigue:

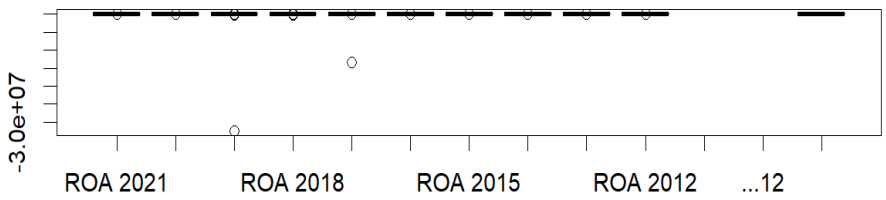


Figura 3:Histograma ROA Anual

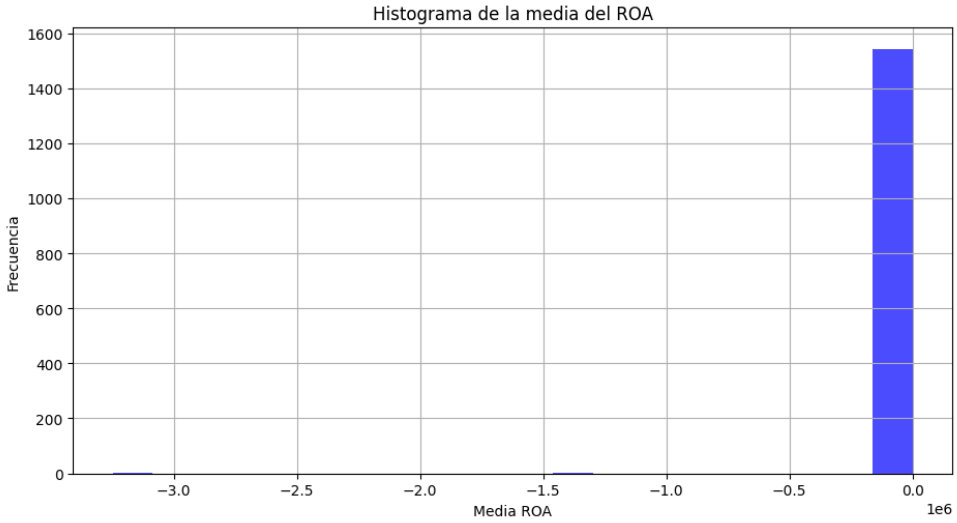


Figura 4:Histograma de la Media del ROA

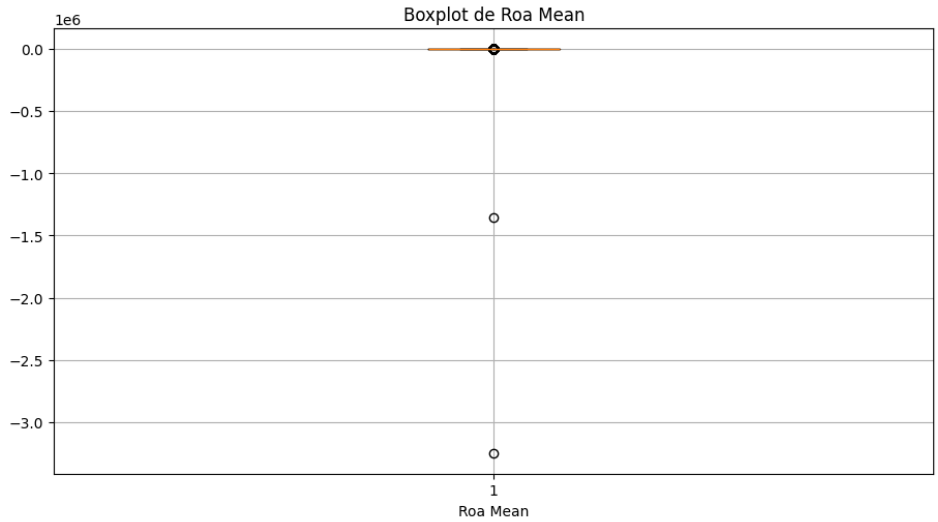
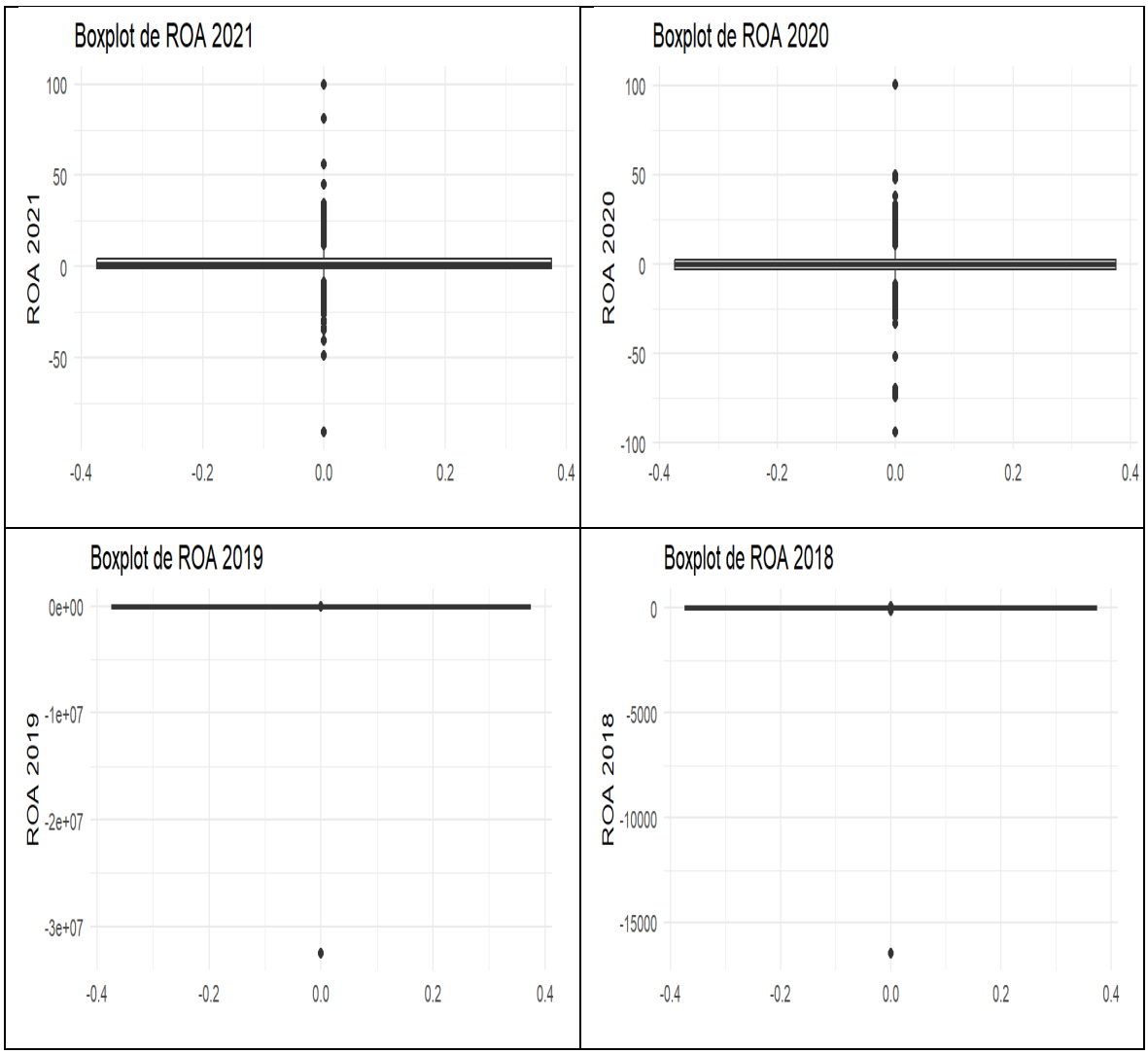


Figura 5: Diagrama de cajas y bigotes Media del ROA

Como se puede apreciar tanto en la tabla como en los gráficos obtenidos del análisis descriptivo de la media (histrograma y boxplot), a pesar de que la mayoría de las bodegas se encuentran concentradas con un ROA cerca de 0, se observan algunas que tienen valores muy dispersos, encontrándose valores máximos de entre 100 y 110 y valores mínimos de entre -90 y -900 y con dos valores mínimos muy llamativos para los ROA 2019, con un mínimo de -32475855.000) y ROA 2017, con un mínimo de -13500055.000. Además, en ambos años se ve claramente afectada la media por dichos valores. Por todo ello, se decidió realizar un análisis de todas las variables, con el objetivo de detectar los

diferentes outliers y eliminarlos para evitar su influencia en la imputación de los datos faltantes.

A continuación, mostramos los gráficos Boxplot de cada una de las variables (Figura 6), donde podemos observar la presencia de los outliers en cada uno de los casos:



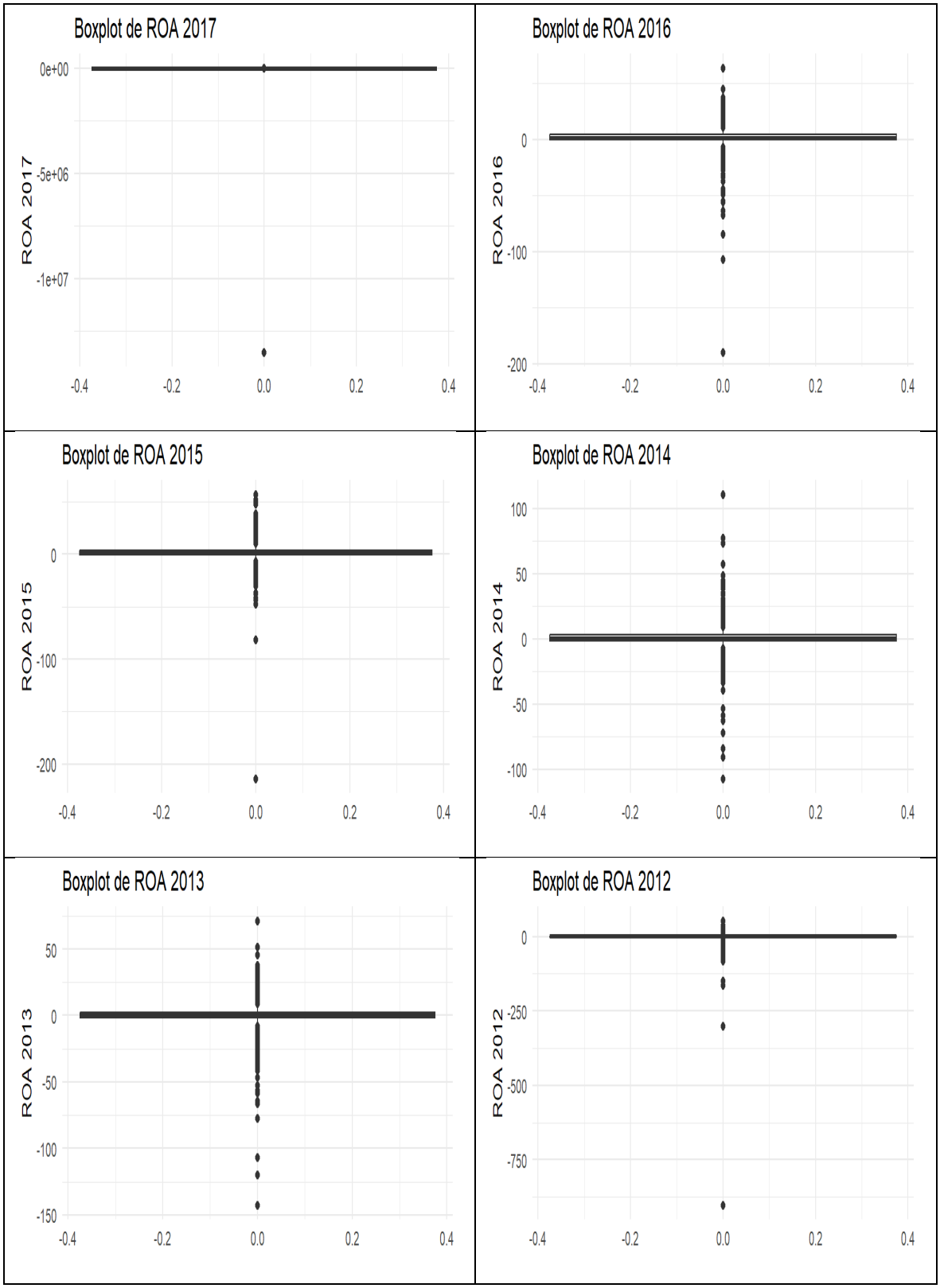


Figura 6: Boxplots de los ROA anuales

De este modo, observamos que los datos no siguen una distribución simétrica con la presencia de outliers en cada uno de los años estudiados, con valores que se alejan de la

media y los cuales afectarían al cálculo de la Media del ROA, una vez aplicada la técnica de imputación MICE.

4.3 Análisis de componentes principales (PCA)

4.3.1 Nipals (PCA): Detección y eliminación de observaciones atípicas

Como se ha indicado en el punto 3.3.1.1, tras la aplicación de la técnica de aprendizaje no supervisado mediante el algoritmo NIPALS PCA para el tratamiento de observaciones atípicas, se han observado los siguientes resultados que se presentan a continuación en el siguiente gráfico (Figura 7: Biplot de componentes principales):

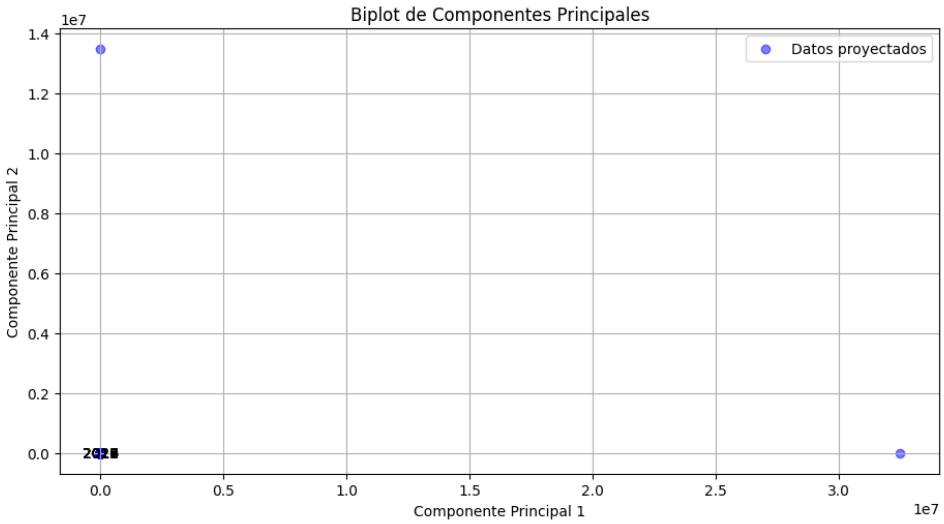


Figura 7: Biplot de componentes principales

Cada punto representa una observación en el espacio de las dos primeras componentes principales.

Se obtiene un valor propio de cada componente principal de: 6.8218955e+11, 1.1788496e+11 y una cantidad de varianza explicada por cada componente principal de: 0.85265752 y 0.14734248.

La distancia entre los puntos y el origen muestra la proyección de cada observación en el espacio de las componentes principales. Las observaciones más alejadas del origen tienen

valores más altos en los componentes principales correspondientes, por lo que con 2 componentes principales se alcanza la varianza de los datos, siendo el componente principal 1 el más destacado de forma significativa con un total del 85 % de la variabilidad explicada de los datos, encontrándose casi la totalidad de las observaciones en valores aproximados a 0. Además, se observan 2 puntos alejados que corresponden a los ROA 2017 y 2019, en los que se han encontrado observaciones atípicas con valores muy alejados de la media.

A continuación, observamos el gráfico de dispersión SPE para la detección de valores atípicos, con una primera iteración (Figura 8 y 9):

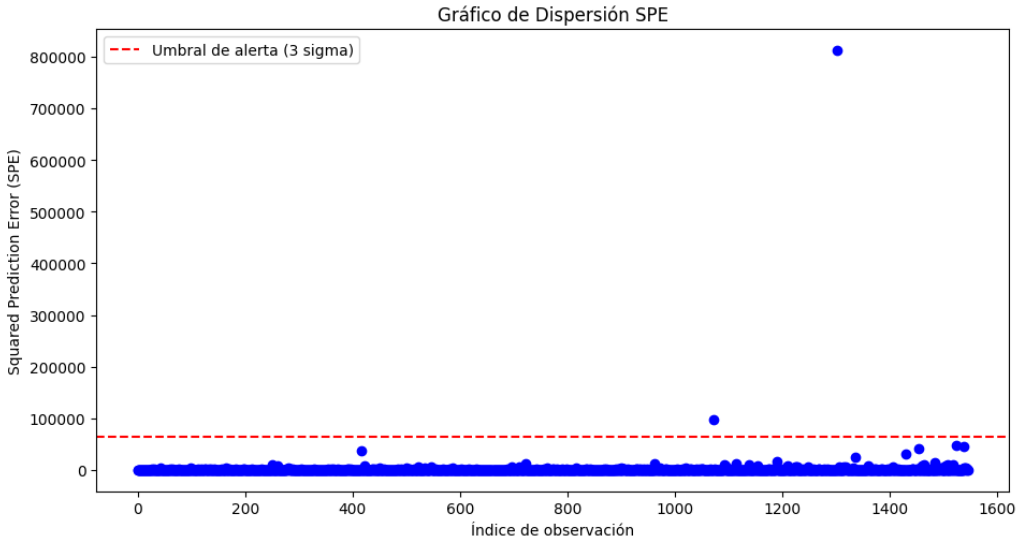


Figura 8: Gráfico de dispersión SPE (Squared prediction error)

Mediante el gráfico de dispersión SPE, observamos valores atípicos en las observaciones 1071 y 1301.

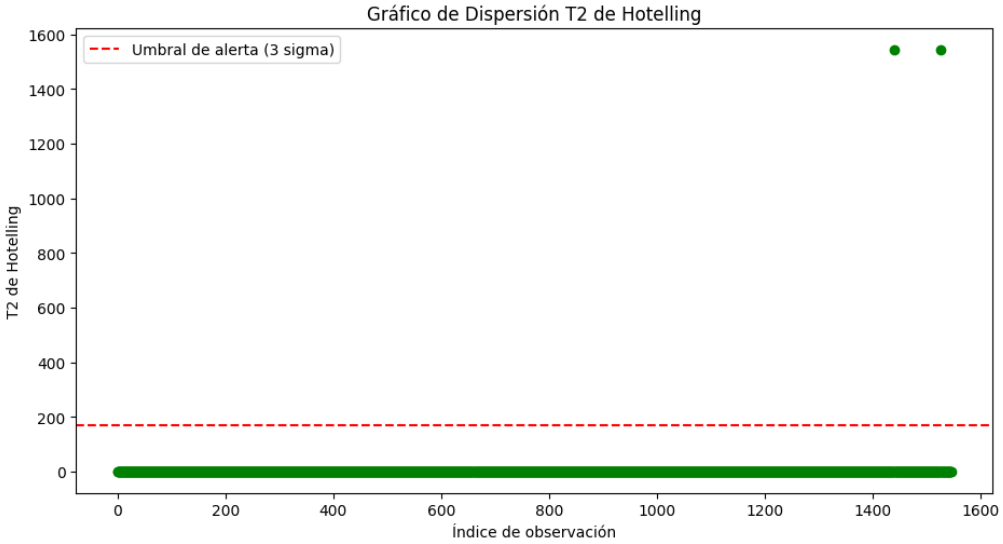


Figura 9: Gráfico de dispersión T2 de Hotelling

Mientras que, con el gráfico T2 de Hotelling, se observan valores atípicos en las observaciones 1439 y 1526.

Así se procedió a su detección y eliminación. Una vez eliminados estos individuos, se volvió a observar los gráficos de componentes principales y de dispersión (Figuras 10, 11 y 12):

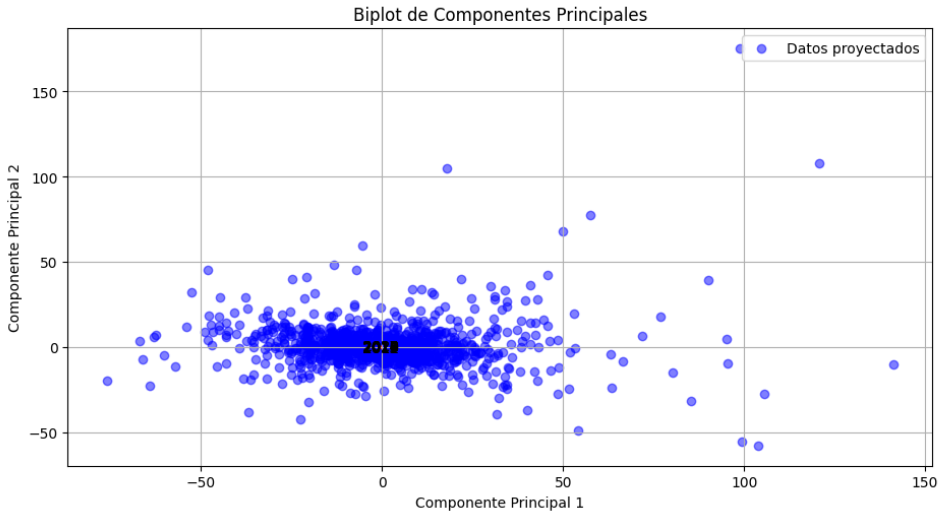


Figura 10: Biplot de componentes principales

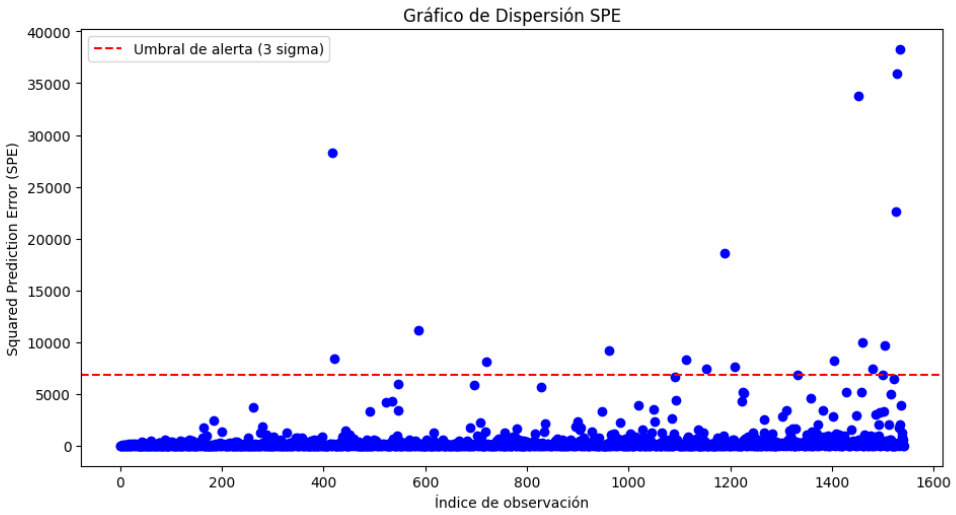


Figura 11: Gráfico de dispersión SPE (Squared prediction error)

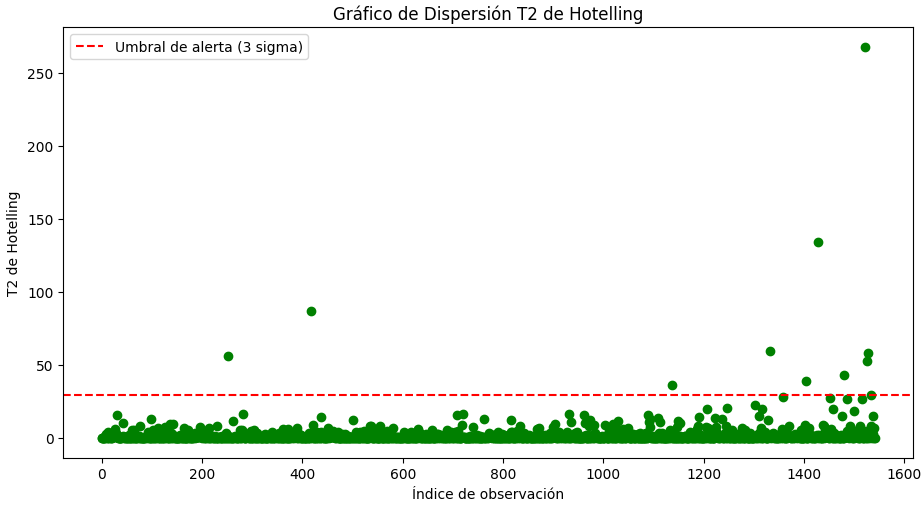


Figura 12: Gráfico de dispersión T2 de Hotelling

Además, se observó de nuevo el histograma y el boxplot de la media ROA (Figuras 13 y 14), donde se observa la distribución.

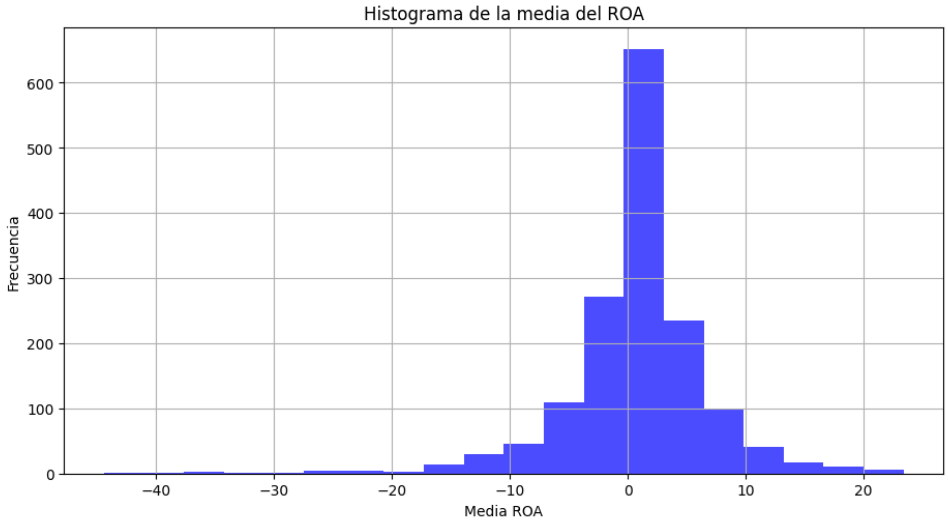


Figura 13: Histograma de la media del ROA

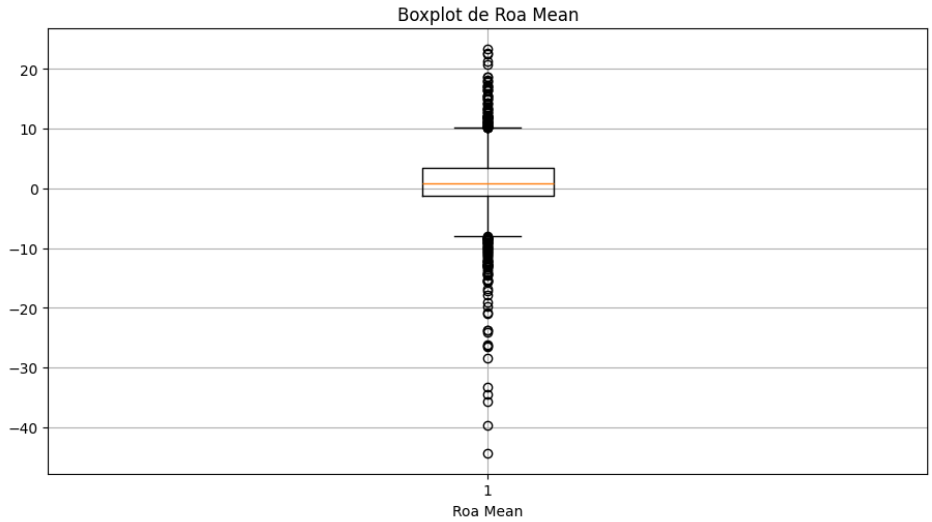


Figura 14: Boxplot de la media del ROA

De este modo, al observar la presencia de outliers tras la primera iteración, se ejecutó un módulo en el que en primer lugar se calcularon SPE Y T2 Hotelling y se definieron los umbrales para determinar los valores a más de 3 desviaciones estándar de la media.

Posteriormente se inicializó un proceso iterativo para encontrar los puntos atípicos basados en los umbrales definidos y se volvió a observar el gráfico de las componentes principales, observando el resto de las observaciones dentro de los límites de control establecidos. (Figura 15):

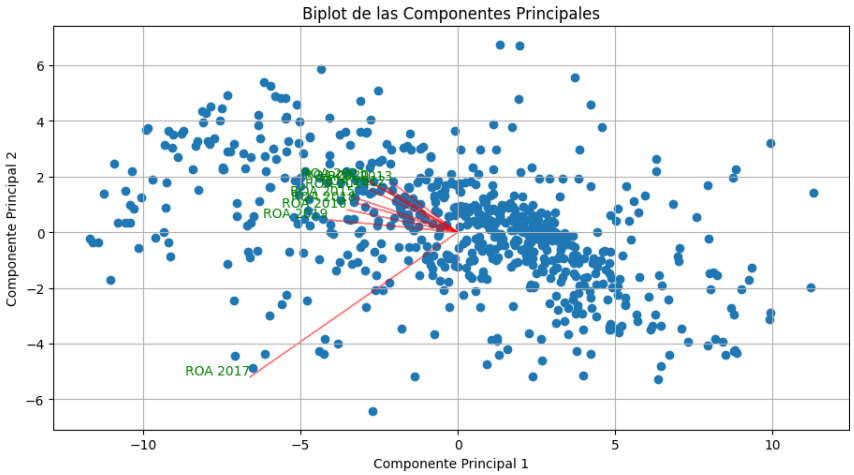


Figura 15: Biplot PCA de los datos después del tratamiento de outliers

Por último, se observaron los gráficos de dispersión, donde se observan ahora sí los puntos dentro de los límites y sin valores atípicos (Figuras 16 y 17):

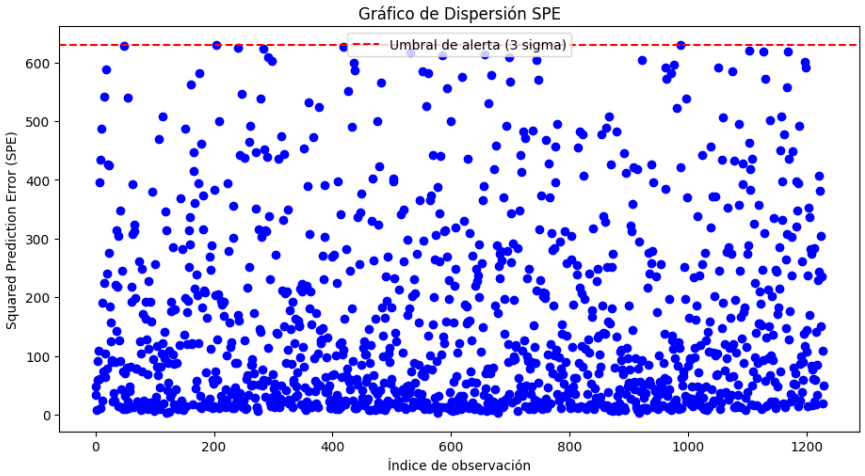


Figura 16: Gráfico de dispersión SPE (Squared prediction error)

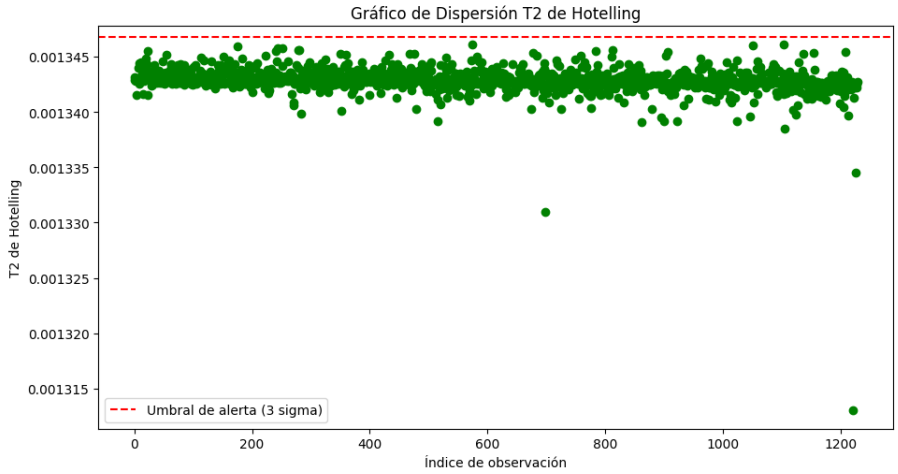


Figura 17: Gráfico de dispersión T2 de Hotelling

Posteriormente, se observó mediante el histograma (Figura 18), la distribución con una mayor semejanza a la forma de la distribución normal, verificando así, el resultado de la aplicación del proceso iterativo para la eliminación de los valores atípicos.

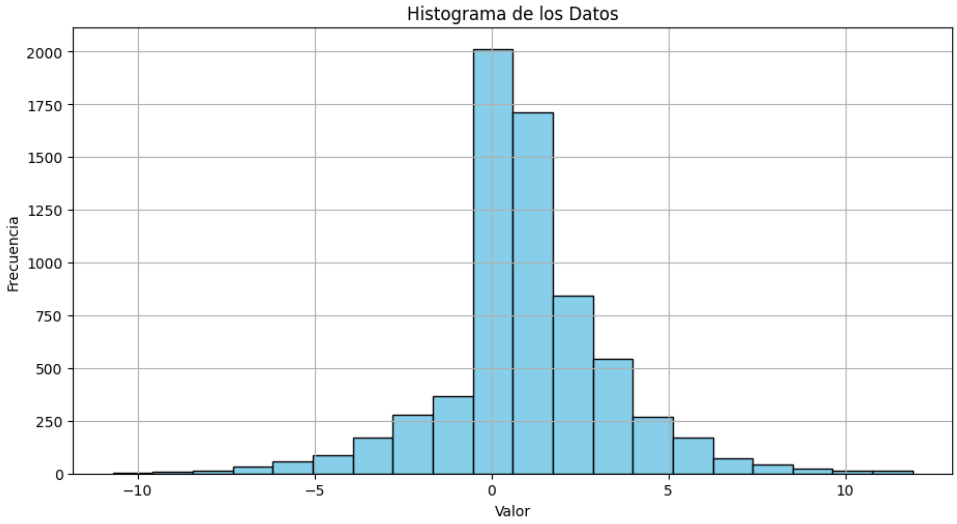


Figura 18: Histograma de los datos después del tratamiento de outliers

4.3.2 Análisis exploratorio de las variables explicativas de huella digital "keywords"

Una vez finalizado el tratamiento de outliers para las variables Roa (anual), dejando los datos con valores ajustados para no alterar los resultados en el tratamiento de datos

faltantes mediante la imputación MICE, se procedió al análisis exploratorio de las variables explicativas de huella digital “Keywords”.

Para ello, en primer lugar, se ha aplicado un PCA para comprender la relación latente entre las variables, tal y como se presenta en las Figuras 19 y 20.

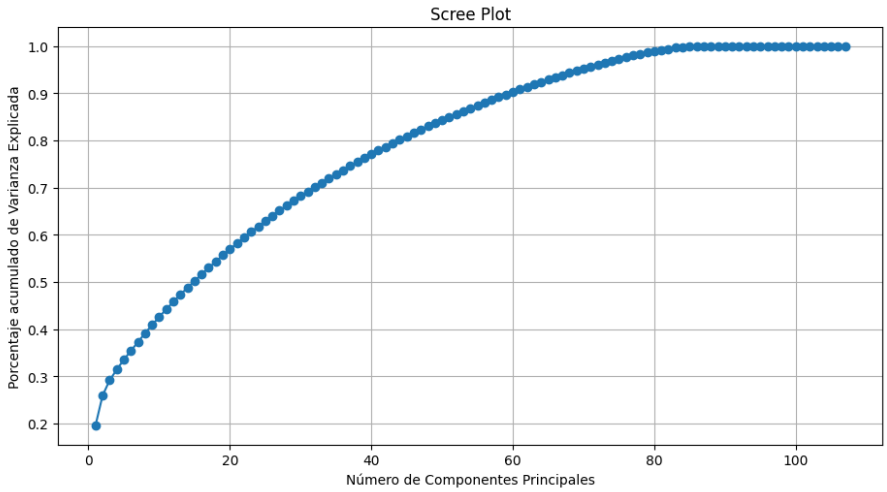


Figura 19: Scree Plot (Python) - PCA Variables Keywords

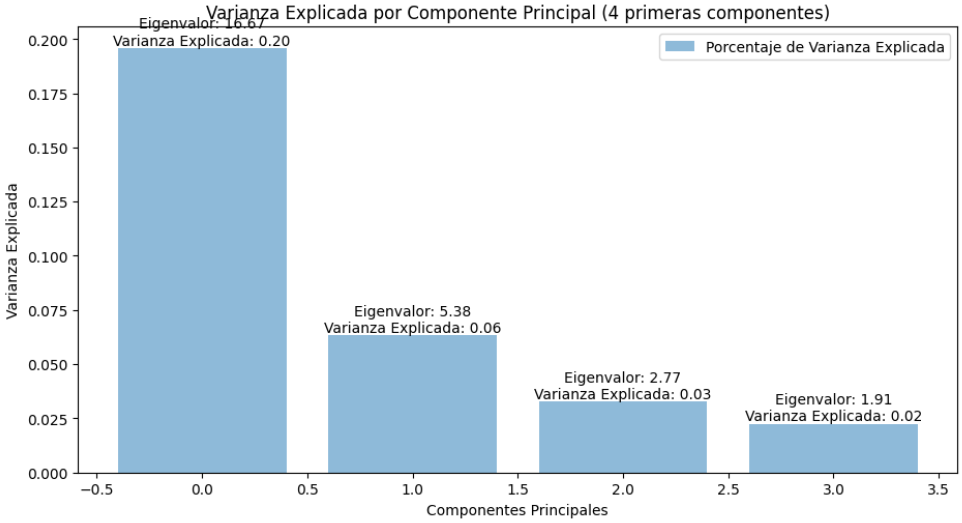


Figura 20: Varianza explicada PCA

En este caso realizado con Python, tal y como se puede observar en ambas figuras anteriores, las dos primeras componentes explican un 26% de la variabilidad de los datos, lo que indica que, las dos primeras componentes principales pueden estar capturando aspectos importantes de la estructura subyacente de los datos. Aunque no capturan toda la variabilidad, la cantidad que explican es considerable.

Cabe considerar, que la tercera componente explica un 3,3% aproximadamente, siendo una tercera dimensión que recoge ligeramente más información que el resto de las dimensiones formadas a partir de la cuarta componente.

Mientras que, mediante R, los resultados obtenidos son muy similares, mostrando un 25,9% el porcentaje de variabilidad explicada por la suma de las dos primeras componentes, como se presenta a continuación en la Figura 21:

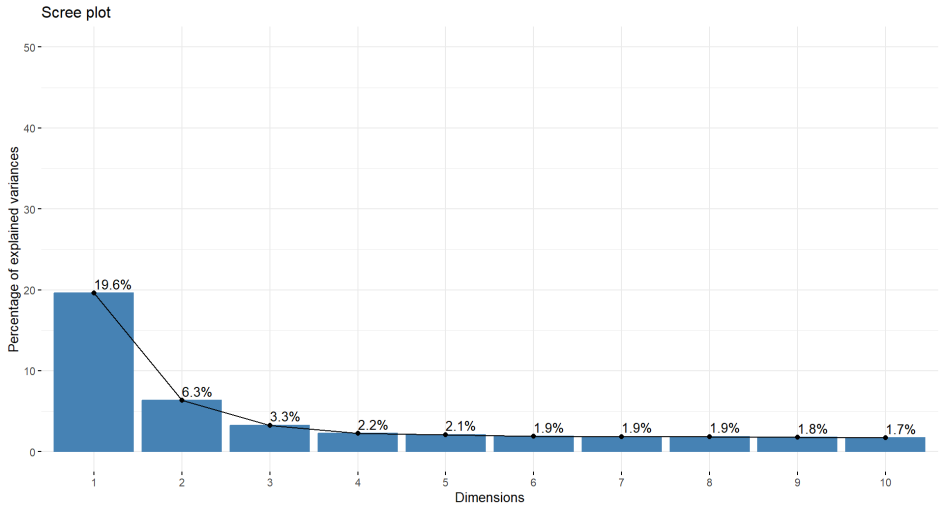


Figura 21: Scree Plot (R) - PCA Variables Keywords

Esta ligera diferencia se debe a los métodos de redondeo utilizados en los cálculos de cada algoritmo.

4.3.2.1 Relación entre variables

A continuación, podemos observar a través de los gráficos de Python, tanto cómo se identifican dos direcciones de variabilidad mediante el gráfico biplot de las componentes principales (Figura 22), como dónde se agrupan conjuntos de variables con correlaciones positivas mediante en el gráfico de Loadings P1-P2 (Figura 23), a través de la primera y segunda componente principal.

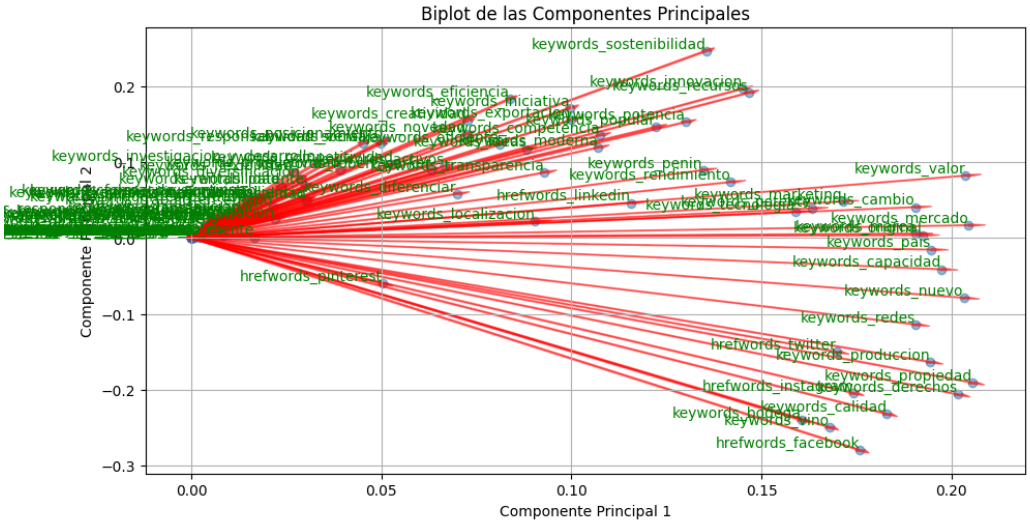


Figura 22: Biplot de las componentes principales

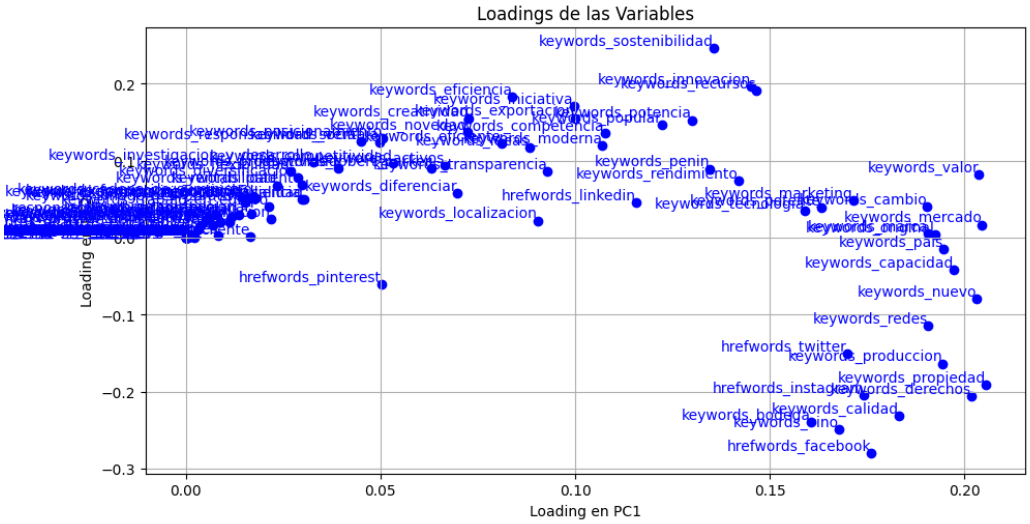


Figura 23: Gráfico de Loadings P1-P2

Las magnitudes de los loadings representan la cantidad de variación en los datos originales que captura cada componente principal, siendo las magnitudes más altas las que capturan una mayor proporción de la varianza de los datos originales.

Tal y como se observa, muchas de las variables se explican en la primera componente principal (variables de la derecha), entre las que destacan aquellas variables de presencia en redes sociales “hrefwords” y algunas “keywords” relevantes como “_valor”, “_marca”, “_vino”, “_bodega”, “_calidad”, “_propiedad”, “_derechos”.

Por otro lado, parece que se diferencian las variables “keywords” como, por ejemplo, “_productividad”, “competitividad”, “_eficiencia”, “_innovación”, “_novedad”, “_marketing”, “_posicionamiento”, “_investigacion_y_desarrollo”, “_sistemas_de_información” o “_flexibilidad”, entre otras. De las que podríamos suponer que guardan una correlación al ser palabras que de alguna manera tienen que ver con el desarrollo, la competitividad y la productividad de la empresa en relación con el marketing, ya que son prácticamente ortogonales con respecto a las anteriores, para ello ampliamos el gráfico de loading plots en la dirección de las variables que captura la pc2, como vemos en la Figura 24:

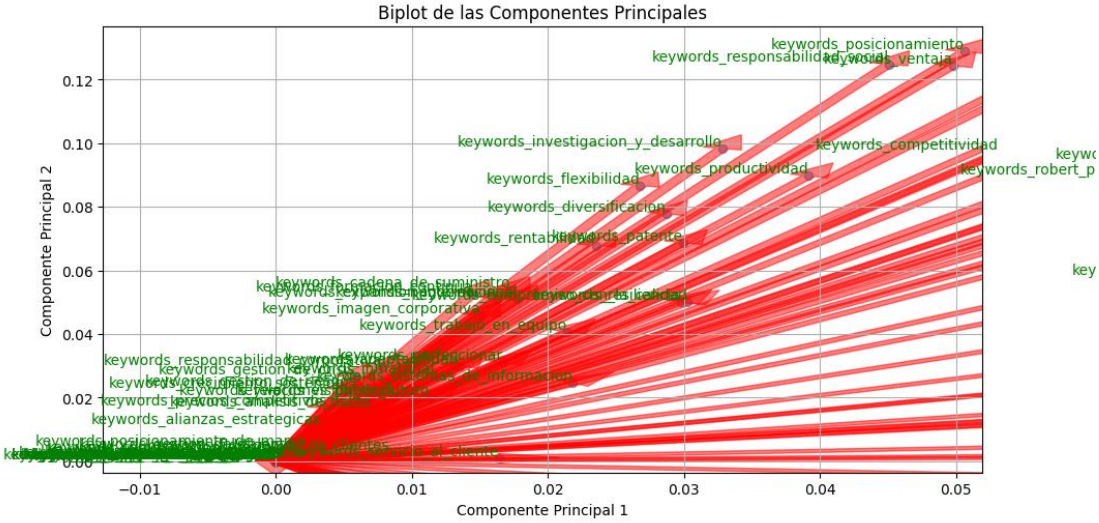


Figura 24: Biplot de PC2

Además, para comprender la relación de esas variables en las dos primeras componentes, también nos apoyamos en el gráfico de loadings de R, que también nos muestra las contribuciones de las variables a cada componente principal, tal y como vemos en la Figura 25 y 26:

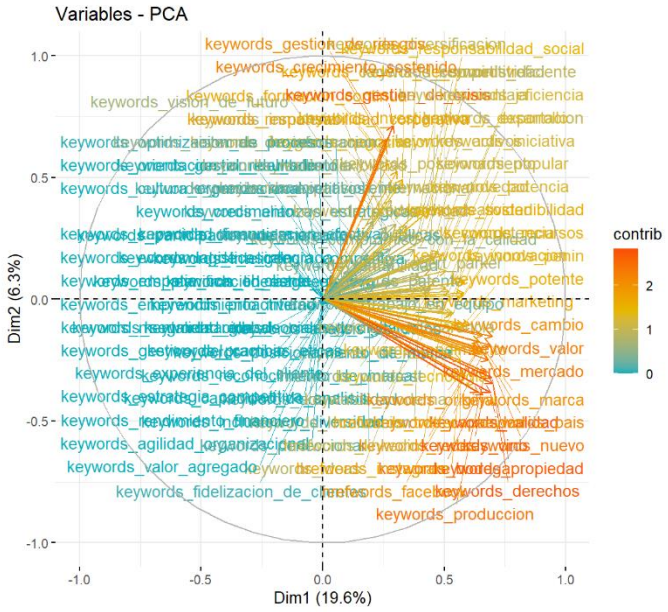


Figura 25: Gráfico de contribuciones a PC

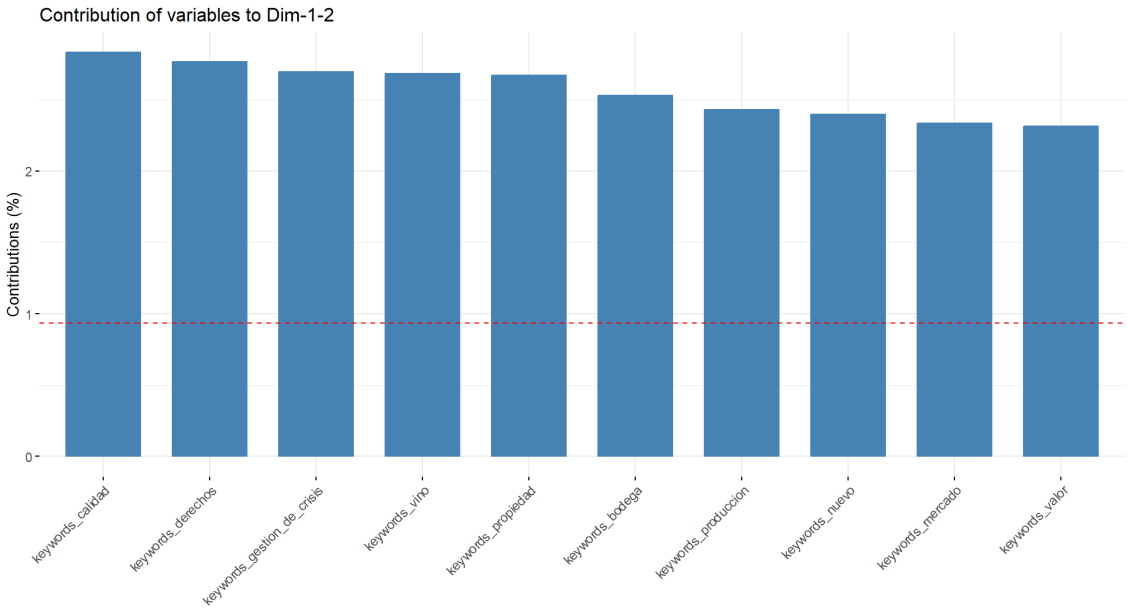


Figura 26: Gráfico de contribuciones a PC

De este modo, verificamos de alguna manera lo observado y comentado anteriormente, ya que, entre las variables con mayor contribución a cada uno de los dos componentes, se observan las variables relacionadas con la presencia en redes sociales “hrefwords” de Facebook, Instagram y twitter, así como la “keyword_redes”. Mientras que, nos apoyamos con el gráfico de barras (Figura 26) para identificar rápidamente cuáles son las variables más importantes para las componentes principales seleccionadas entre las que destacan

el tipo de palabra como las keyword “calidad”, “derechos”, “gestión_de_crisis”, “vino”, “propiedad”, “bodega”, “producción”, “nuevo”, “mercado y “valor”, con un claro patrón a la contribución de la primera componente principal, observándose el resto y gran mayoría de variables keyword, con una mayor contribución hacia la segunda componente principal.
Esto nos ayuda a concluir la importancia de la presencia en las redes, la importancia o relación entre la bodega o marca y la calidad del vino, y se relaciona la innovación, el marketing o la investigación con la competitividad, la productividad y la eficiencia, lo cual resulta lógico, al ser factores relevantes en el desarrollo y el buen funcionamiento empresarial.

4.3.2.2 Relación entre observaciones

A continuación, en la Figura 27, mostramos el comportamiento de las observaciones dentro del espacio latente. El gráfico de scores T1-T2 es una herramienta poderosa para visualizar y comprender la estructura subyacente de los datos después de aplicar PCA.

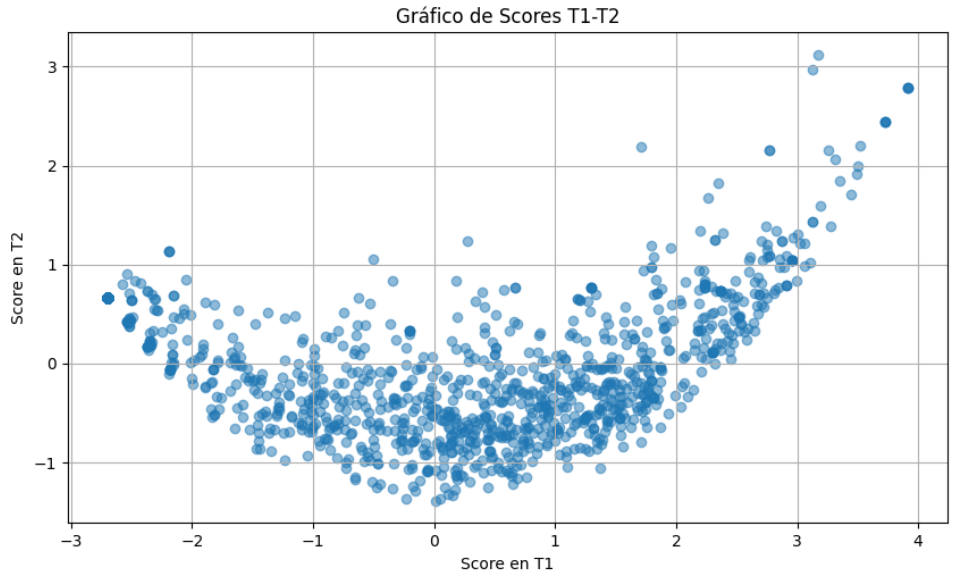


Figura 27: Gráfico de scores T1-T2

Además, se ha implementado la visualización en 3D utilizando las tres primeras componentes principales para ver si la estructura observada cambia o se clarifica al agregar una dimensión adicional (Figura 28):

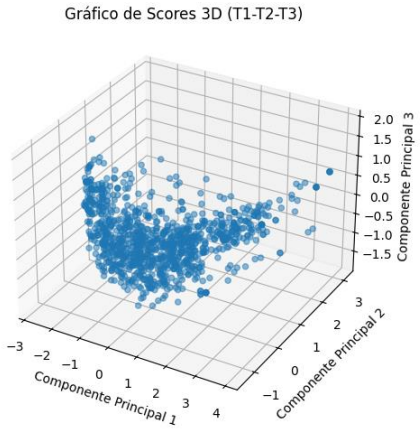


Figura 28: Gráfico de scores 3D

Como vemos, la estructura observada anteriormente mantiene su forma, observando una cola o tendencia de algunos puntos que se separan hacia la formación de la tercera componente, siendo la minoría, por lo que resulta coherente con el menor % de variabilidad explicada en dicha componente (3,3%).

Por último, en la Figura 29, coloreamos por una de las variables del bloque de huella digital, la cual ha sido mostrada anteriormente con un mayor peso en el gráfico de contribuciones ("keyword_calidad").

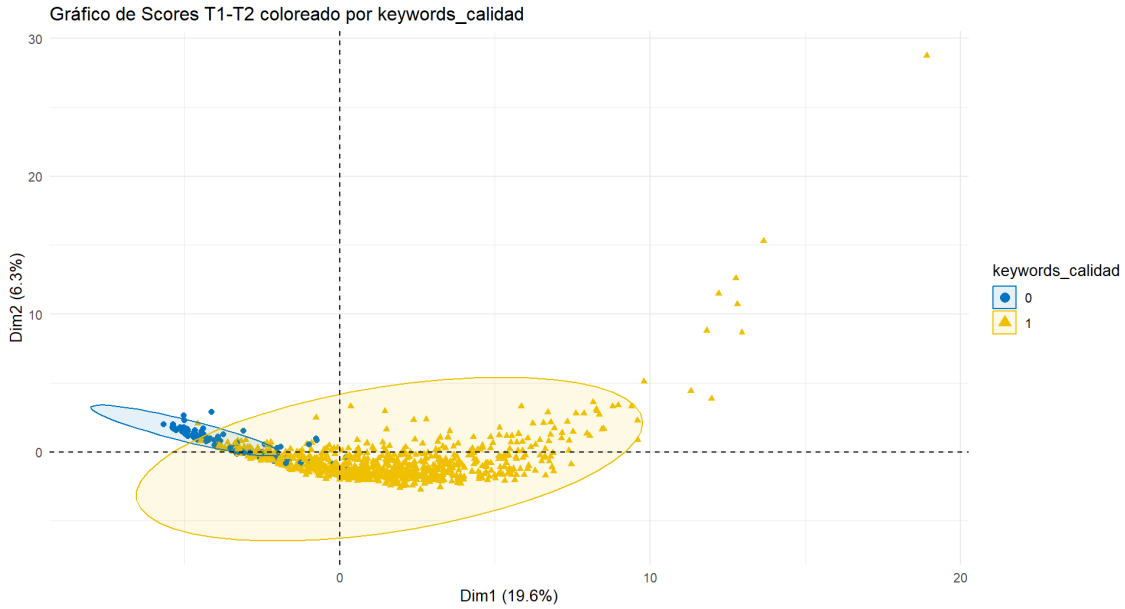


Figura 29: Gráfico de scores (t1-t2) en el espacio de las variables latentes coloreado por variable keyword calidad

De este modo, al colorear las observaciones por la variable "keyword_calidad" se observa que las mismas siguen la dirección de variabilidad observada anteriormente en el gráfico

de loadings (p1-p2), y presentan algunas diferencias en cuanto a sus magnitudes o niveles (colores distintos), que siguen un patrón, por lo que sería un indicador de que en el conjunto de bodegas de vino existen grupos con factores en común.

4.4 Análisis de clasificación Clustering

4.4.1 Estadístico de Hopkins

Inicialmente se presentan los resultados del estadístico de Hopkins para ver si los datos tienen tendencia a agruparse, obteniendo el siguiente resultado:

Hopkins statistic: 0.745

Por lo que, al ser un valor HO>0,5, siendo que el valor de H se acerca a 1, podemos rechazar la hipótesis nula y concluir el conjunto de datos presenta una tendencia moderada a agruparse.

4.4.2 VAT (Visual assessment of cluster tendency)

Posteriormente, evaluamos visualmente indicios de algún tipo de agrupación en los datos mediante el método “VAT”, donde se representa gráficamente la matriz de distancias ordenada (Figura 30), empleando un gradiente de color para el valor de las distancias.

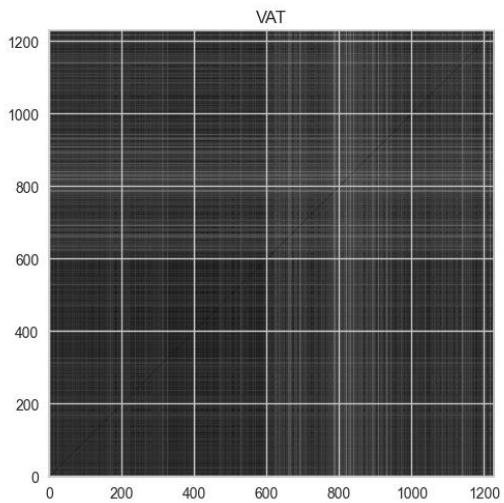


Figura 30: Matriz de distancias (VAT)

Observamos así, la existencia de agrupaciones subyacentes en los datos, donde se forma un patrón de bloques cuadrados.

4.4.3 Método del codo

Finalmente, tras valorar el resultado del estadístico de Hopkins y observar una cierta tendencia a la agrupación, en la Figura 31 se muestra el gráfico del método del codo para la evaluación y selección del número de clusters (K):

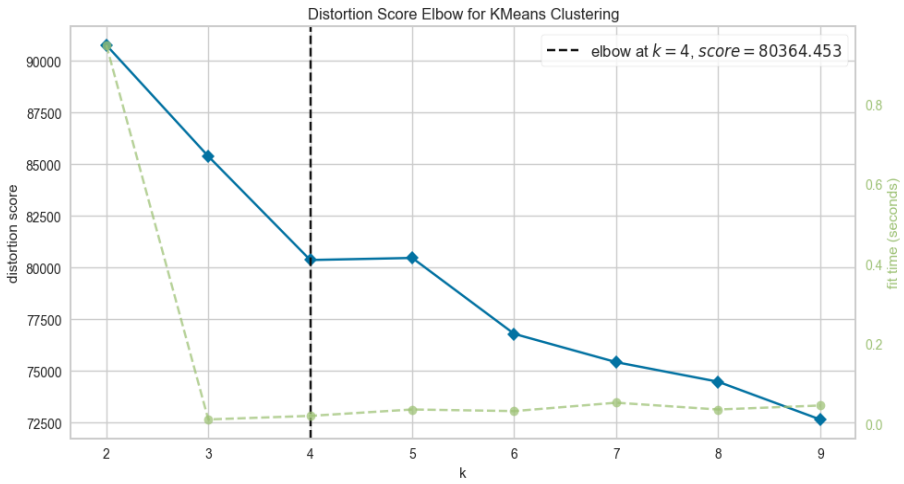


Figura 31: Gráfico Método del codo con Python

Como podemos apreciar en el gráfico, el eje Y “distribution score” muestra típicamente la suma de cuadrados dentro del grupo (Within-Cluster Sum of Squares, WSS), para cada clusters (K) en el eje X. Donde el número óptimo es donde en el gráfico se estabiliza WCSS, es decir cuando observamos el cambio brusco o “codo” en la línea continua que une los puntos (cuando K=4).

Observamos así, en ambos gráficos, que el número óptimo de clusters sería k=4, donde se estabiliza la curva.

4.4.4 Caracterización de los clusters:

4.4.4.1 Clustering difuso (C-Means)

A continuación, se exponen en la Figura 32 los resultados obtenidos tras aplicar el método de Clustering difuso:

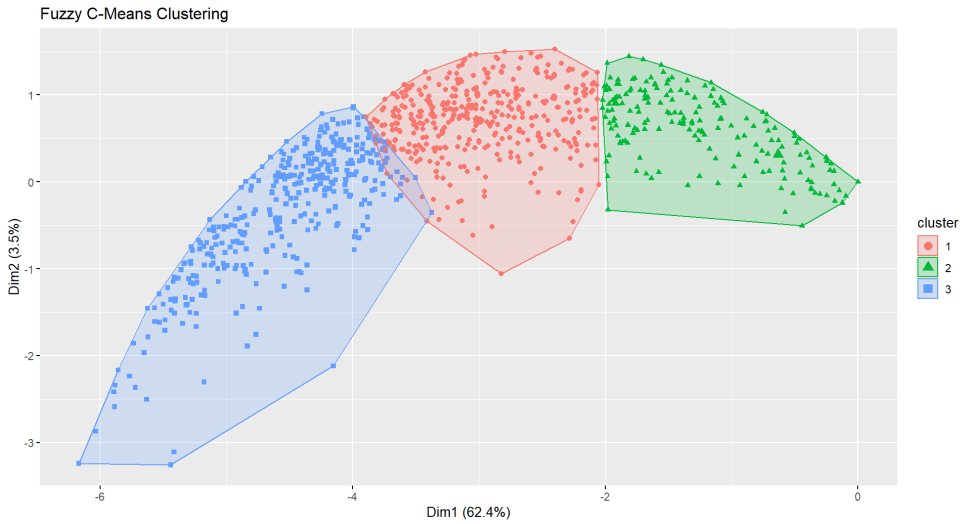


Figura 32: Cluster plot (C-Means Clustering)

Mediante la partición del clustering difuso (C-Means), que asigna a cada dato un grado de pertenencia dentro de cada cluster y siendo que un dato puede pertenecer parcialmente a más de un grupo, observamos que algunos individuos comparten los grupos 1 y 3, aunque se diferencian bien 3 grupos o clusters entre los datos.

El Cluster 1 contiene 405 observaciones y las variables más representativas que destacan en este grupo son, keywords_calidad, keywords_vino, keywords_bodega, keywords_derechos, keywords_propiedad, keywords_produccion, hrefwords_facebook, keywords_sostenibilidad, keywords_redes. Por lo que, este cluster parece estar caracterizado por términos relacionados con la calidad, vino, bodega, derechos y propiedad. Mientras que, la presencia de hrefwords_facebook y keywords_redes, sugiere un enfoque en redes sociales y sostenibilidad.

Por otro lado, el cluster 2 con un total de 395 bodegas, parece compartir varias variables con el Cluster 1, pero tiene un énfasis adicional en términos como “novedad” y “mercado”. Mientras que, las variables keywords_propiedad y keywords_derechos tienen una alta representación, lo que sugiere un enfoque fuerte en aspectos legales o de propiedad.

Por último, en el Cluster 3 formado por 429 bodegas, destacan las variables keywords_valor, keywords_mercado, keywords_cambio, keywords_original, keywords_capacidad, keywords_pais, keywords_marca, keywords_nuevo, keywords_marketing, por lo que este cluster parece estar más orientado hacia términos relacionados con valor, mercado, cambio, originalidad, capacidad, país, marca y marketing, destacando una posible orientación hacia la comercialización y el valor de mercado.

En definitiva, las variables keywords_calidad, keywords_vino, keywords_bodega y keywords_produccion son comunes en Clusters 1 y 2, sugiriendo una similitud en términos de enfoque o características en estos clusters.

4.4.4.2 Clustering particional (K-means)

Mientras que, en las Figuras 33 y 34 se muestran los resultados de las agrupaciones de los datos tras la aplicación del método clásico o Clustering particional (k-means), tanto siguiendo la pauta del método del codo para k=4, como tras realizar el ajuste a k=3, respectivamente:

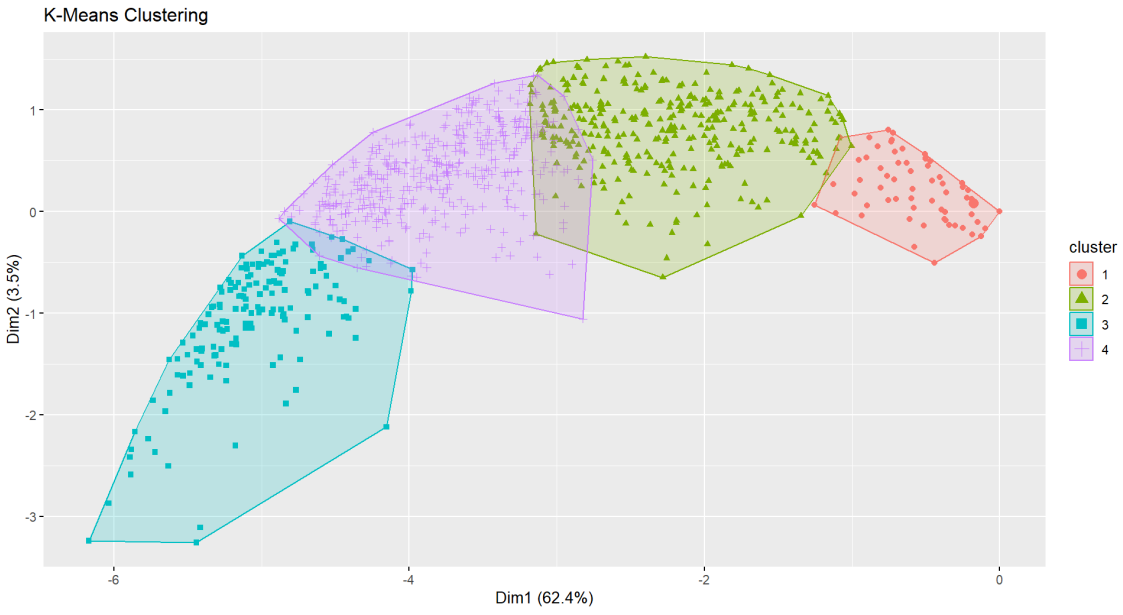


Figura 33: Cluster plot (K-means clustering, k=4)

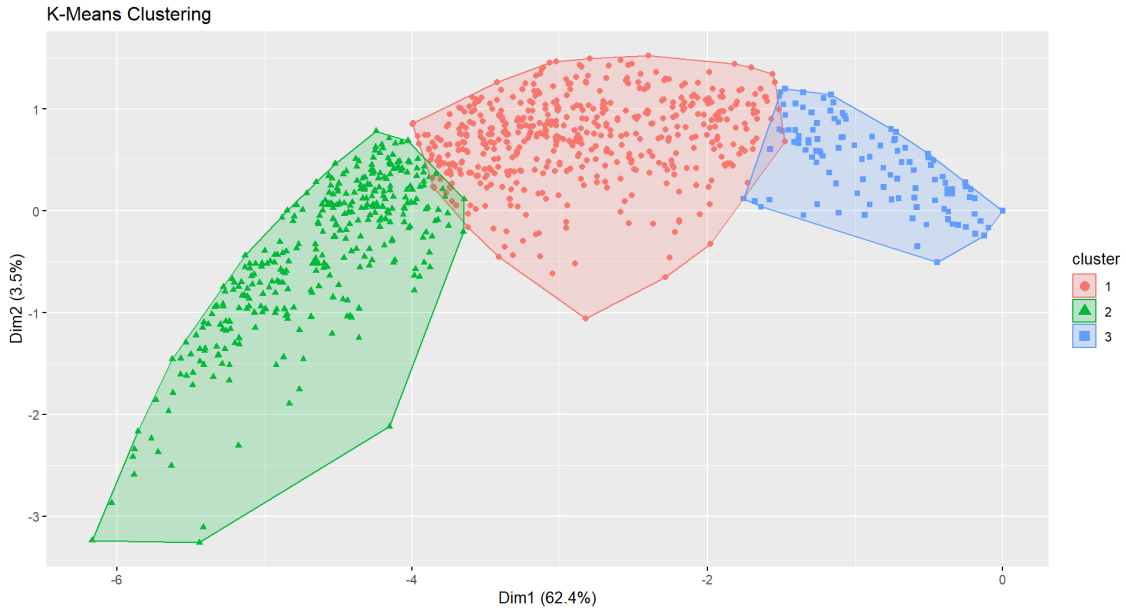


Figura 34: Cluster plot (K-means clustering, k=3)

En el modelo jerárquico con 4 grupos, se aprecia que la mayoría de las observaciones se agrupan en los grupos 2 y 4, llegando a superponerse, por lo que las observaciones en ambos grupos comparten atributos. Por tanto, ajustando el modelo a 3, se observa cómo se soluciona el problema de superposición entre grupos y la separación entre estos es más clara. De esta manera es más sencillo identificar las diferencias entre los distintos grupos y se observan 3 grupos mejor diferenciados.

No obstante, cabe destacar que, debido a que los clusters parecen no estar tan bien separados y poseen diferente tamaño y densidad, los resultados del clustering difuso son mejores, ya que este algoritmo clasifica los grupos difusos en función de un grado de pertenencia de cada observación a cada grupo o cluster, permitiendo que la clasificación de las observaciones en cada cluster esté mejor definida que en el K-means.

4.5 Regresión Múltiple

4.5.1 Análisis exploratorio y evaluación de las hipótesis asociadas al modelo de regresión múltiple

Tal y como hemos comentado en el punto 3.4.1 “Regresión Lineal Múltiple”; posteriormente al tratamiento de los outliers, se realizó un último análisis exploratorio de los datos donde se realizó la evaluación de las diferentes hipótesis asociadas al modelo de regresión

múltiple (normalidad, homocedasticidad, autocorrelación y multicolinealidad). En primer lugar, incluyendo el dato mínimo observado tras la eliminación de los outliers que se encontraba muy alejado del resto de las observaciones, como sigue:

Análisis incluyendo el dato mínimo:

Resultados métodos numéricos:

Mean	std	min	P25	Median	P75	Max	Coef asimetría:	Curtosis
0,986	3,151	26.496	-0.441	0,821	2,695	9,657	-0.540	4.690

Tabla 3: Análisis descriptivo: Elaboración propia

Los valores del coeficiente de asimetría y la curtosis, proporcionan información sobre la forma de la distribución de los datos. Un coeficiente de asimetría negativo indica que la cola de la distribución se extiende más hacia la izquierda que hacia la derecha. Mientras que, una curtosis positiva indica que la distribución tiene colas más pesadas y un pico más alto que una distribución normal. Esto se ve fundamentado por la presencia del dato mínimo con un valor demasiado alto, tal y como se ve a continuación en la distribución de los datos, Para ello, se observó el Histrograma de la media del ROA (Figura 35):

Normalidad:

Métodos Gráficos para evaluar la normalidad:

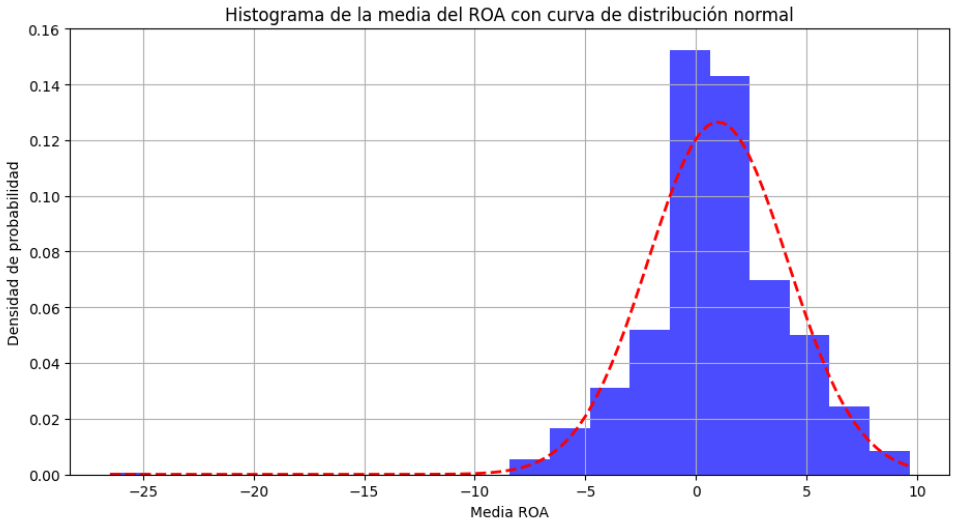


Figura 35: Histograma de la media del ROA (incluyendo el dato mínimo)

Además, se comprobó la presencia del dato mínimo alejado del resto y la distribución de los datos mediante el diagrama de caja ROA (Figura 36) y el gráfico Q-Q (Figura 37):

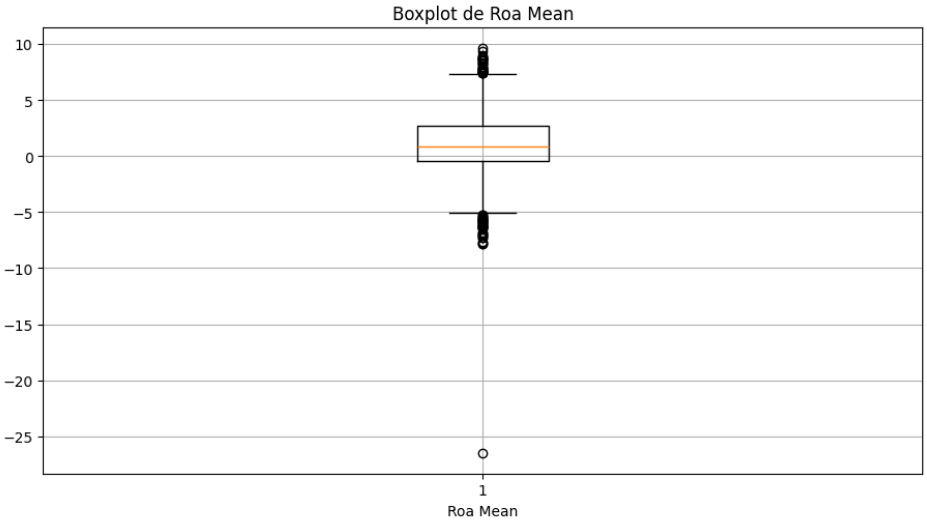


Figura 36: Boxplot de la media del ROA (incluyendo el dato mínimo)

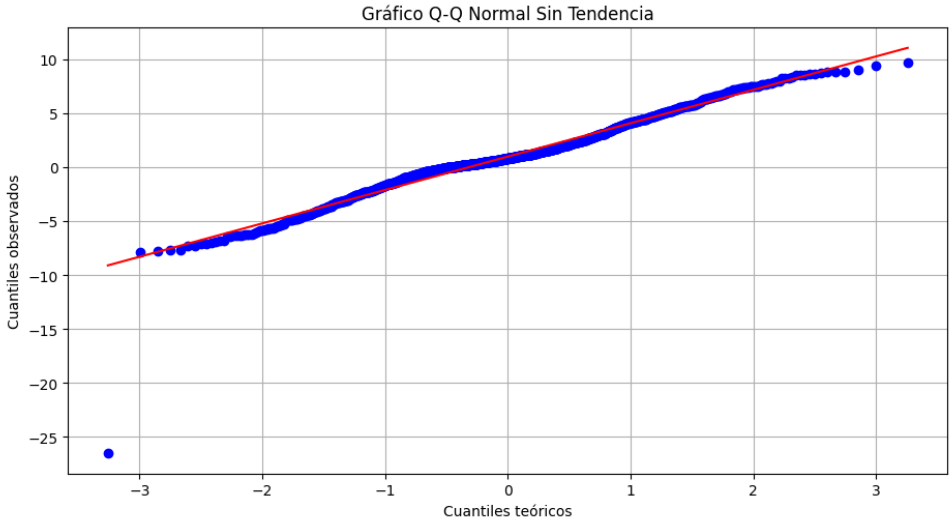


Figura 37: Gráfico Q-Q (incluyendo el dato mínimo)

Por último, se verificó mediante las pruebas de normalidad de forma numérica la hipótesis nula de la normalidad de los datos (Tabla 4 y 5):

Métodos numéricos para la evaluación de la normalidad:

Pruebas de normalidad Shapiro-Wilk y de Kolmogorov-Smirnov:

Resultado de la prueba de Shapiro-Wilk	
Estadístico de prueba	0.963

Tabla 4: Resultado de la prueba de Shaphiro-Wilk incluyendo el dato mínimo

Resultado de la prueba de Kolmogorov-Smirnov	
Estadístico de prueba	0.317
P-valor	4.866e-111

Tabla 5: Resultado de la prueba de Kolmogorov-Smirnov incluyendo el dato mínimo

Tras la visualización de los resultados de las pruebas de normalidad (Shapiro-Wilk y Kolmogorov-Smirnov) indican que los datos no siguen una distribución normal. En ambos casos, el P-valor asociado es extremadamente pequeño (menor que 0.05), lo que sugiere que hay suficiente evidencia para rechazar la hipótesis nula de que los datos provienen de una distribución normal.

Posteriormente, se procedió nuevamente a la exploración de los datos excluyendo el dato mínimo, utilizando los mismos métodos gráficos y numéricos, obteniendo los siguientes resultados:

Análisis excluyendo el dato mínimo:

Resultados métodos numéricos:

Mean	std	Min	P25	P50	P75	Max	Coef asimetría:	Curtosis
1.008	3.053	7.862	-0.438	0.822	2.697	9.657	-0.021	0.362

Tabla 6: Análisis descriptivo: Elaboración propia

Métodos Gráficos para evaluar la normalidad:

Para evaluar la normalidad gráficamente, a continuación, vemos el histograma, el diagrama de caja de la media del ROA y el gráfico Q-Q (Figura 38, Figura 39 y Figura 40 respectivamente).

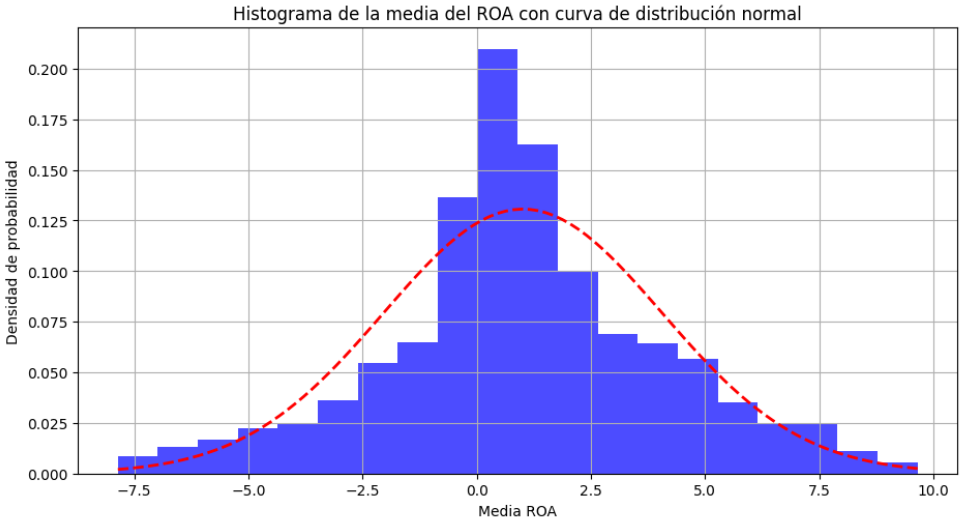


Figura 38: Histograma de la media del ROA (excluyendo el dato mínimo)

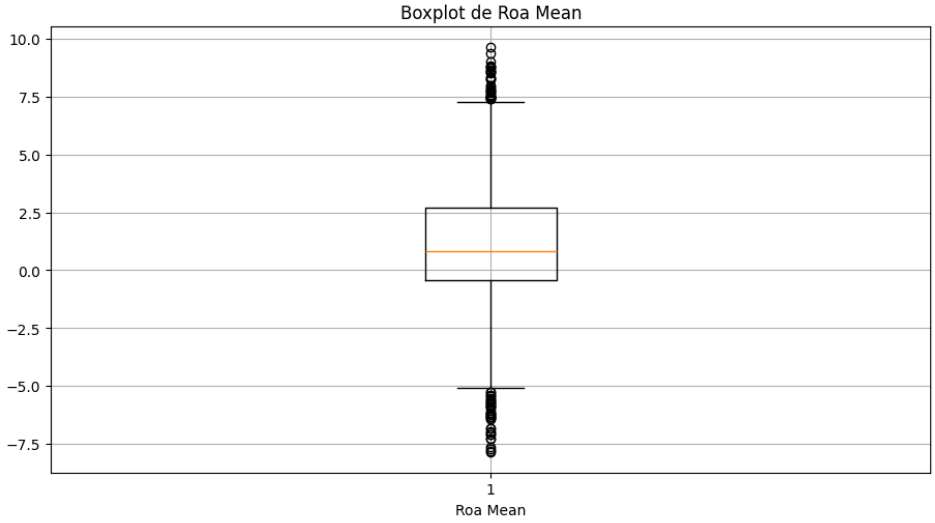


Figura 39: Boxplot de la media del ROA (excluyendo el dato mínimo)

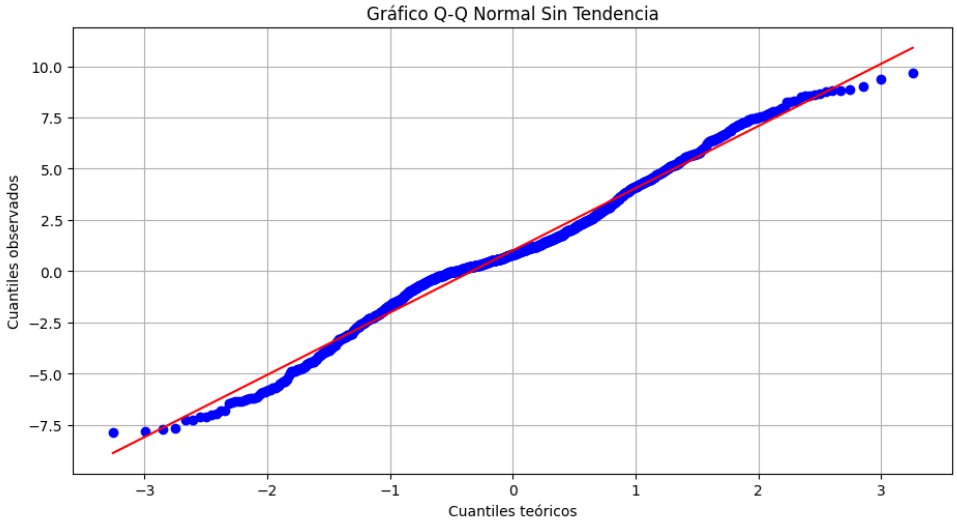


Figura 40: Gráfico Q-Q (excluyendo el dato mínimo)

Del mismo modo, se verificó mediante las pruebas de normalidad de forma numérica la hipótesis nula de la normalidad de los datos (Tabla 7 y 8):

Métodos numéricos para la evaluación de la normalidad:

Pruebas de normalidad Shapiro-Wilk y de Kolmogorov-Smirnov:

Resultado de la prueba de Shapiro-Wilk	
Estadístico de prueba	0.985
P-valor	1.251e-09

Tabla 7: Resultado de la prueba de Shapiro-Wilk excluyendo el dato mínimo

Resultado de la prueba de Kolmogorov-Smirnov	
Estadístico de prueba	0.318
P-valor	3.520e-111

Tabla 8: Resultado de la prueba de Kolmogorov-Smirnov excluyendo el dato mínimo

Como se observó en los gráficos, los datos se encontraron más centrados tras su eliminación, en un rango entre -7.862 y 9.657 y fijando la media en 1.008.

Por un lado, los histogramas revelaron la distribución de frecuencias y los patrones de los datos, observando una distribución muy similar a la normal. Mientras que, los gráficos Q-Q proporcionaron una comparación directa con una distribución normal ideal. Se observaron algunos puntos alejados de la línea diagonal que marca la normalidad de los datos, siendo que lo que se espera es que los residuos estandarizados estén los más cerca posible a la línea diagonal que aparece en el gráfico.

Mientras que, los resultados de ambas pruebas, Shapiro-Wilk y Kolmogorov-Smirnov, indican que los datos no provienen de una distribución normal. Dado que el p-valor es menor que el nivel de significancia (0.05) en ambas pruebas (1.251e-09 para Shapiro-Wilk y 3.520e-111 para Kolmogorov-Smirnov), se rechaza la hipótesis nula en ambos casos, lo que indica que los datos no siguen una distribución normal.

No obstante, se puede comprobar numéricamente mediante la prueba de Shapiro-Wilk, que la distribución se acerca más a la normalidad, una vez eliminado el dato mínimo alejado del resto, pasando de 0,9636 a 0,99856.

Al eliminar el dato mínimo -26,4961, se evidenció una mejora en la distribución de los datos, encontrándose de una forma más centrada, con una apariencia más cercana a la normal, una mejora en el resultado de la prueba de Shapiro-Wilk y apoyándonos en las primeras pruebas realizadas en los modelos de regresión lineal en las que se compararon los modelos incluyendo y excluyendo el dato, observando una ligera mejora en los modelos sin el dato mínimo, decidimos eliminarlo para los posteriores análisis.

Homocedasticidad:

El modelo de regresión lineal múltiple debe cumplir que la varianza de los errores es constante en sus observaciones. La homocedasticidad del modelo de regresión es una cualidad necesaria para que los coeficientes estimados sean eficientes y lineales. En caso contrario se dice que el modelo presenta heterocedasticidad.

Una forma de evaluar la homocedasticidad del modelo es mediante el gráfico de residuos frente a predichos (Figura 41 y Figura 42). En dicho gráfico lo ideal es que los residuos no muestren ningún patrón, ya que si se aprecia que el tamaño de los residuos crece o disminuye en función de los predichos, entonces no se cumple la condición de homocedasticidad:

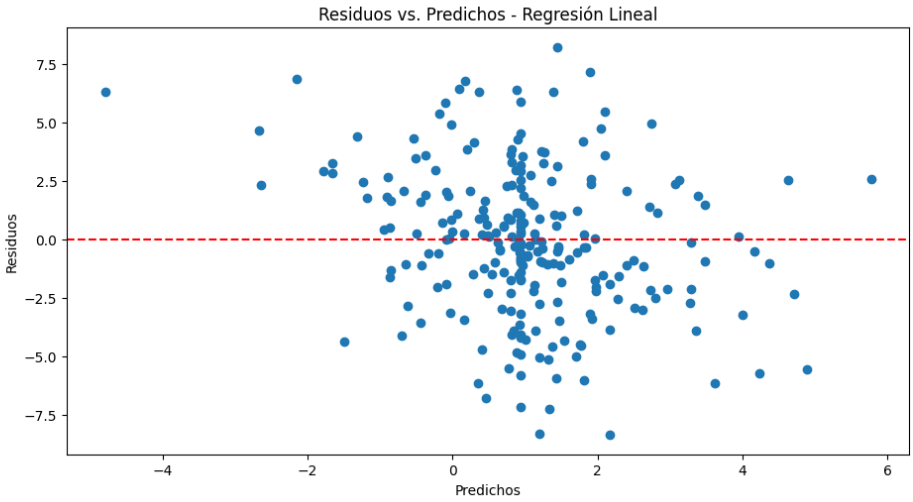


Figura 41: Gráfico Residuos vs Valores predichos con Python

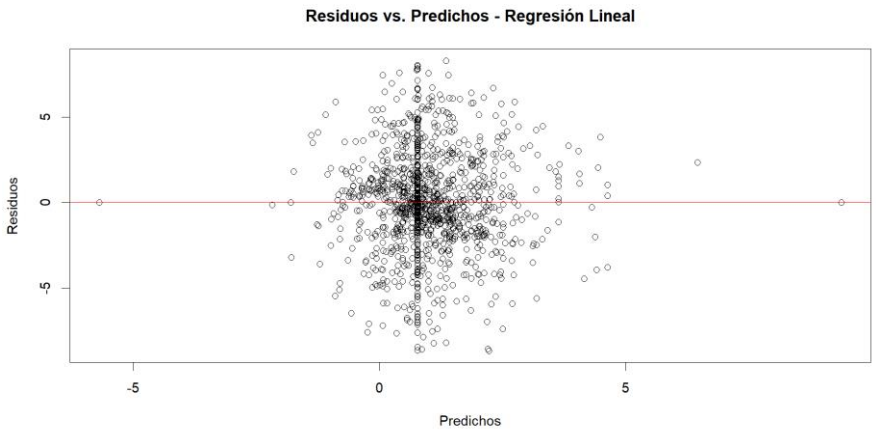


Figura 42: Gráfico Residuos vs Valores predichos con R

En los gráficos se observan los residuos distribuidos de una manera aleatoria y sin formar patrones por lo que, se puede considerar un modelo homocedástico.

Autocorrelación:

El modelo de regresión lineal múltiple debe cumplir que los errores sean independientes, para que los valores no dependan unos de otros.

Para determinar la autocorrelación del módelo se utiliza el Test de Durbin-Watson, que sirve como diagnóstico en este sentido. Este test realiza el siguiente contraste de hipótesis:

$$H_0: \rho = 0 \text{ [ErroresIndependientes]}$$

$$H_1: \rho \neq 0 \text{ [ErroresDependientes]}$$

Resultados del Test de Durbin-Watson	
Estadístico de Durbin-Watson	2.009

Tabla 9: Test de Durbin-Watson

Los residuos no tienen autocorrelación significativa.

Multicolinealidad:

La multicolinealidad de las variables en el modelo se determina si los valores de inflación de variables “VIF” son altos (se consideran altos por encima de 10).

Para resolver dicho problema debemos realizar de nuevo una selección de variables:

vif = 1. / (1. - r_squared_i)

Variable	VIF
Const	1.058127
kwstems_es_sistem_de_informacion	1.000219
keywords_optimizacion_de_procesos	inf
kwstems_es_optimizacion_de_proces	inf

Tabla 10: Evaluación de multicolinealidad con todas las variables dicotómica palabras clave - VIF

Los resultados muestran que las variables keywords_optimizacion_de_procesos y kwstems_es_optimizacion_de_proces tienen un valor de VIF igual a infinito (inf). Esto indica una alta multicolinealidad entre estas variables y otras en el modelo. Esto resulta una obviedad, ya que como comentamos en un principio en el punto 3.1.2, en las variables de contenido web, se generaron un total de 203 variables dicotómicas de palabras clave (102 del nivel de concordancia exacta y 101 del nivel de Raíz de la palabra), de modo que siempre va a coincidir el dato en la búsqueda de una palabra por su raíz con esa misma variable buscada con la “concordancia exacta”.

De este modo, resulta lógico eliminar las 101 variables buscadas en la aplicación por su “raíz de palabra”, para realizar los modelos, obteniendo así los siguientes resultados para el diagnóstico de la multicolinealidad:

Variable	VIF
Const	1.044633

keywords_sistemas_de_informacion	1.000164
keywords_optimizacion_de_procesos	1.000164

Tabla 11: Evaluación de multicolinealidad con las variables del nivel de concordancia exacta – VIF

Observamos así, que en este caso no existe multicolinealidad entre las variables, ya que los valores de la VIF son pequeños.

4.5.2 Regresión múltiple – Python

En la Tabla 12 se presentan los resultados o medidas de bondad de ajuste del modelo de regresión lineal múltiple:

Medidas de Bondad de Ajuste	
Error Cuadrático Medio (MSE)	9.500
Error Absoluto Medio (MAE)	2.375
Raíz del Error Cuadrático Medio (RMSE)	3.082
Normalized Mean Squared Error (NMSE)	1.018
Normalized Mean Absolute Error (NMAE)	0.141
Mean Absolute Percentage Error (MAPE)	206.525

Tabla 12: Medidas de bondad de ajuste modelo de regresión lineal múltiple Python

Para evaluar los diferentes modelos, utilizamos el indicador MSE que mide el promedio de los cuadrados de los errores entre los valores predichos y los valores reales. Un MSE más bajo indica un mejor rendimiento del modelo. En este caso, observamos un valor del MSE igual a 9.5, lo que indica que las predicciones difieren en casi 10 puntos de media con respecto del valor real, siendo una diferencia demasiado grande. Por lo que se considera que no sería un buen modelo para predecir la variable respuesta “Roa mean”.

Mientras que, en la Tabla 13 se muestran los resultados obtenidos por el modelo de regresión:

Resultados	
Residual standard error	3.030
Multiple R-squared	0.016

Adjusted R-squared	0.014
F-statistic	8.238
p-value	0.0002

Tabla 13: Resultados modelo de regresión lineal múltiple

En cuanto a los resultados obtenidos por el modelo de regresión lineal múltiple, observamos que el valor del R^2 es demasiado bajo, por lo que la proporción de la varianza de la variable dependiente explicada por las variables independientes es muy baja. Este valor indica que sólo el 1,65% de la variabilidad de los datos es explicada por el modelo. Sin embargo, vemos que el p valor es muy bajo, por lo que sugiere que al menos una de las variables independientes tiene un efecto significativo sobre la variable dependiente. Así, entendemos que el modelo no sería un buen modelo para predecir la variable dependiente, pero sí podemos obtener información sobre las variables que sí aportan un poder explicativo sobre ella, uno de nuestros objetivos en el trabajo. Por ello, se realizan los modelos de regresión para evaluar y comparar su capacidad predictora y extraer información relevante sobre la importancia de las variables.

Se comprobaron también las diferentes hipótesis asociadas al modelo.

Evaluación de hipótesis asociadas al modelo:

1- Normalidad:

Para evaluar la normalidad gráficamente, a continuación, vemos el histograma, y el gráfico Q-Q (Figura 43, y Figura 44 respectivamente).

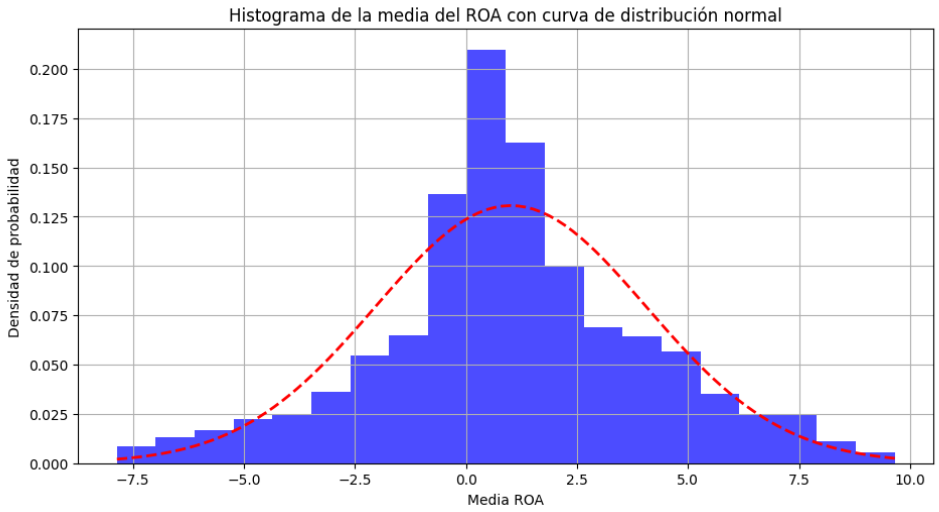


Figura 43: Histograma de la media del ROA

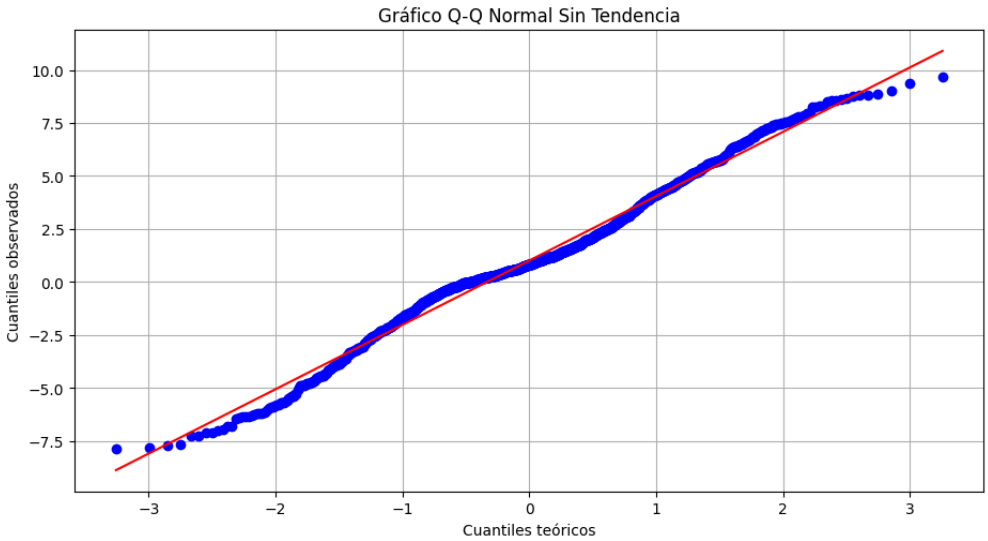


Figura 44: Gráfico Q-Q

Se observa mediante el histograma una distribución muy similar a la normal. Mientras que, del mismo modo, el gráfico Q-Q evidencia una distribución normal, presentando solo puntos alejados de la línea diagonal que marca la normalidad de los datos.

Mientras que en las Tablas 14 y 15 se muestran los resultados obtenidos para la evaluación de la normalidad por métodos numéricos:

Pruebas de normalidad Shapiro-Wilk y de Kolmogorov-Smirnov:

Resultado de la prueba de Shapiro-Wilk	
Estadístico de prueba	0.984
P-valor	6.791e-09

Tabla 14: Resultado de la prueba de Shapiro-Wilk

Resultado de la prueba de Kolmogorov-Smirnov	
Estadístico de prueba	0.206
P-valor	4.402e-37

Tabla 15: Resultado de la prueba de Kolmogorov-Smirnov

Donde los resultados obtenidos sugieren que los residuos no parecen provenir de una distribución normal (se rechaza H0).

2- Homocedasticidad:

Para evaluar la homocedasticidad del modelo se observó el gráfico de residuos frente a predichos (Figura 45):

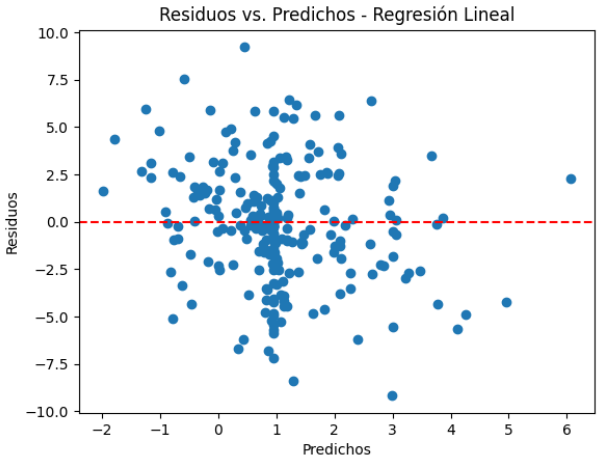


Figura 45: Gráfico Residuos vs Valores predichos

Se considera que se cumple el principio de homocedasticidad, al observarse los residuos distribuidos de forma aleatoria.

3- Autocorrelación:

En la Tabla 16 se muestran los resultados del test para la evaluación de la autocorrelación:

Resultados del Test de Durbin-Watson	
Estadístico de Durbin-Watson	2.001

Tabla 16: Test de Durbin-Watson

El resultado obtenido indica que los residuos no tienen autocorrelación significativa.

4- Multicolinealidad:

Por último, en la Tabla 17 se muestran los resultados del VIF para la evaluación de la multicolinealidad:

Variable	**VIF**
Const	1.044633
keywords_sistemas_de_informacion:	1.000164
keywords_optimizacion_de_procesos:	1.000164

Tabla 17: Resultados VIF – Evaluación multicolinealidad

Observamos así, que en este caso no existe multicolinealidad entre las variables, ya que los valores de la VIF son pequeños.

Finalmente, en la Figura 46 observamos mediante el gráfico de valores observados frente a valores predichos.

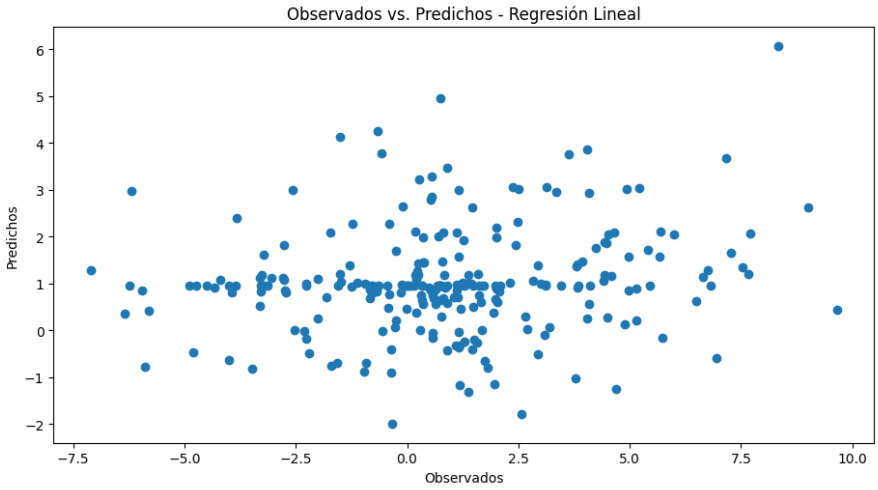


Figura 46: Gráfico valores observados vs valores predichos

En el gráfico de valores observados y valores predichos, observamos los puntos distribuidos de forma aleatoria, mostrando la distancia existente entre el valor real y el valor predicho. Por lo que, como nos indican los resultados de las medidas de bondad de ajuste obtenidas en el modelo, se considera que no es un modelo satisfactorio para predecir la variable respuesta “Roa_mean”.

4.6 Árbol de regresión

4.6.1 Árbol de regresión – Python

En la Tabla 18, observamos los resultados de las medidas de bondad de ajuste tras la implementación del modelo árbol de regresión en Python.

Medidas de bondad de ajuste	
Error Cuadrático Medio (MSE)	10.594

Error Absoluto Medio (MAE)	2.487
Raíz del Error Cuadrático Medio (RMSE)	3.2549
Normalized Mean Squared Error (NMSE)	1.135
Normalized Mean Absolute Error (NMAE)	0.148
Mean Absolute Percentage Error (MAPE)	340.703

Tabla 18: Medidas de bondad de ajuste modelo árbol de regresión - Python

Obteniendo resultados relativamente elevados, lo que nos indica que el modelo presenta errores significativos en sus predicciones.

Por otro lado, en la Figura 47 mostramos Visualización gráfica del árbol de regresión:

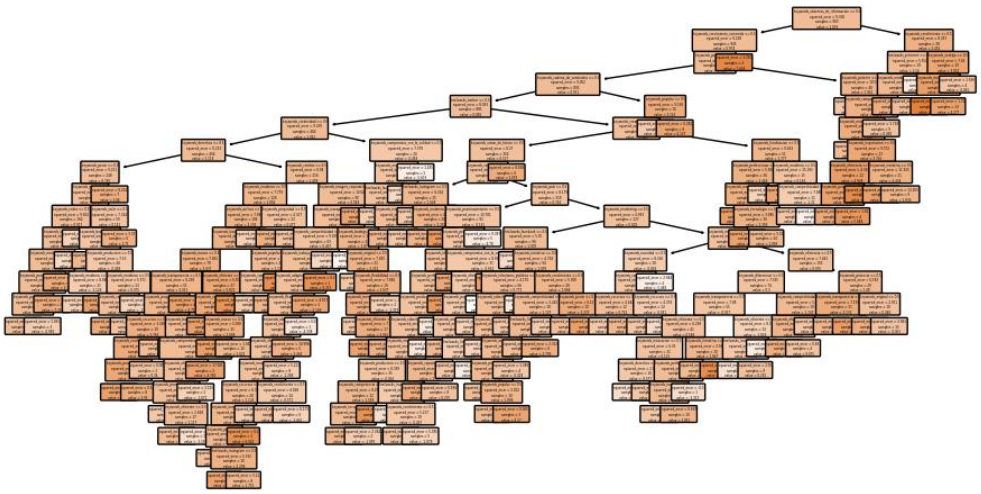


Figura 47: Árbol de regresión con Python

Mientras que, a continuación, se muestra el nivel de importancia de las variables en la Tabla 19:

Feature	**Importance**
keywords_eficiente	0.051635
keywords_moderna	0.048644
keywords_competitividad	0.037713

keywords_popular	0.034959
keywords_transparencia	0.033555
keywords_rendimiento	0.032454
keywords_marca	0.028867
keywords_bodega	0.027598
keywords_creatividad	0.025990
keywords_novedad	0.025788

Tabla 19: Importancia de las variables – Python

Se observa que las variables con mayor importancia son las keywords eficiente, moderna, competitividad, popular, transparencia, rendimiento, marca, bodega, creatividad y novedad. Esto resulta coherente según lo comentado en el punto 4.3.2 en los resultados del PCA, donde veíamos que muchas de estas variables se agrupaban en la formación de las diferentes componentes principales y también coinciden con las variables destacadas en la formación del modelo de regresión lineal múltiple, donde veíamos en el punto 4.5.3, que algunas de las variables más importantes o influyentes sobre la variable dependiente roa_mea coinciden con las destacadas por su nivel de importancia, como por ejemplo las keywords, “eficiente”, “moderna”, “marca”, “bodega”, etc.

4.7 Bosques aleatorios (Ramdom forest)

4.7.1 Bosques aleatorios (Ramdom forest) – Python

A continuación, en la Tabla 20 observamos los resultados de las medidas de bondad de ajuste el modelo de bosques aleatorios tras su implementación en Python.

Medidas de bondad de ajuste bosques aleatorios	
Error Cuadrático Medio (MSE)	8.737
Error Absoluto Medio (MAE)	2.240
Raíz del Error Cuadrático Medio (RMSE)	2.955

Normalized Mean Squared Error (NMSE)	0.936
Normalized Mean Absolute Error (NMAE)	0.133
Mean Absolute Percentage Error (MAPE)	225.834
Coeficiente de Determinación (R^2)	0.063

Tabla 20: Medidas de bondad de ajuste en el modelo Ramdom forest – Python

4.8 Máquinas de soporte vectorial (SVM)

4.8.1 Máquinas de soporte vectorial (SVM) – Python

En la Tabla 21 observamos los resultados de las medidas de bondad de ajuste el modelo SVM tras su implementación en Python.

Medidas de Bondad de ajuste	
Error Cuadrático Medio (MSE)	9.375
Error Absoluto Medio (MAE)	2.355
Raíz del Error Cuadrático Medio (RMSE)	3.061
Normalized Mean Squared Error (NMSE)	1.004
Normalized Mean Absolute Error (NMAE)	0.140
Mean Absolute Percentage Error (MAPE)	176.876

Tabla 21: Medidas de bondad de ajuste en el modelo SVM - Python

A continuación, vemos los gráficos de predicciones frente a valores reales (Figura 48) y residuos vs valores predichos (Figura 49).

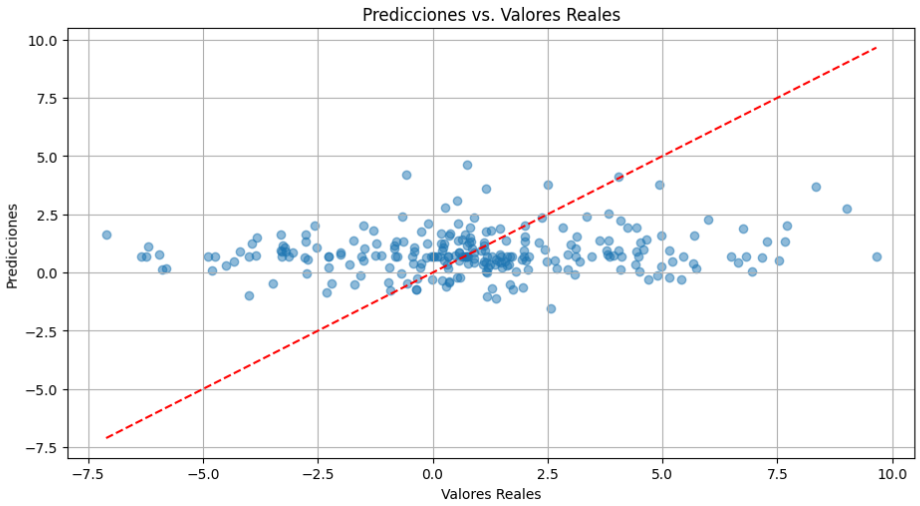


Figura 48: Gráfico de valores predichos vs valores reales con Python

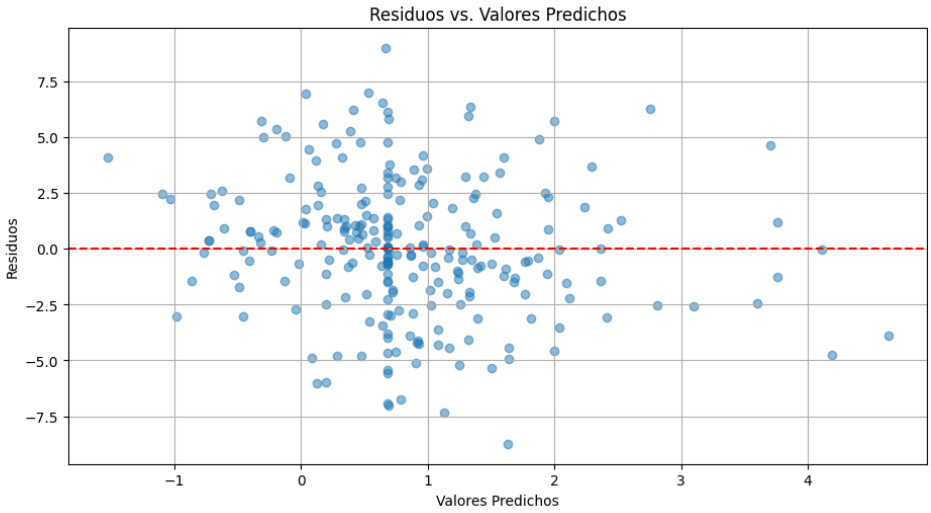


Figura 49: Gráfico de residuos vs valores predichos con Python

El modelo SVM, con un MSE de 9.3750, presenta resultados que indican, tal y como vemos en los gráficos de residuos y de observados frente a predichos, que no es un buen modelo para predecir la variable respuesta “roa mean”, donde se observan los puntos distribuidos de forma aleatoria y alejados de la línea trazada que marca la igualdad entre los valores observados y los valores predichos.

Así, pasamos a la evaluación y comparación de los modelos de regresión implementados en el estudio.

4.9 Evaluación y comparación de modelos de regresión

Como comentamos anteriormente, para la evaluación y comparación de modelos, existen diferentes criterios, siendo el más popular el cálculo del desempeño predictivo. El desempeño predictivo de los modelos de regresión se obtiene comparando las predicciones de los modelos con los valores reales de las variables objetivo, y mediante el cálculo de alguna medida de error promedio de esta comparación: Medidas de bondad de ajuste.

Para ello, el primer paso es la obtención de las predicciones del modelo para el conjunto de casos donde se quiere evaluar. Tanto en Python como en R, se obtienen las predicciones de cualquier modelo utilizando el método “predict()”, el cual recibe un modelo y un conjunto de datos generando las correspondientes predicciones del modelo:

4.9.1 Hold out

Se han desarrollado los cuatro modelos descritos anteriormente (regresión lineal, árbol de regresión, random forest y svm) utilizando la técnica holdout, obteniendo los siguientes resultados representados en la tabla de comparación de resultados (Tabla 22) para las diferentes medidas de bondad de ajuste y en ambos lenguajes de programación:

TABLA COMPARACION DE RESULTADOS – MEDIDAS DE BONDAD DEL AJUSTE

	MODELO	**MSE**	**MAE**	**RMSE**	**NMSE**	**NMAE**	**MAPE**	**VARIABLE RESPUESTA (Y)**
PYTHON	Regresion Lineal	9.5001	2.3756	3.0822	1.014163	0.1416	206.5256	ROA MEAN
R	Regresion Lineal	9.5001	2.3756	3.0822	1.0142	0.9756	206.5257	ROA MEAN
PYTHON	Arbol regresión	10.5948	2,4873	3.2549	1.1356	0.1483	340.7036	ROA MEAN
R	Arbol regresión	9.3485	2.3017	3.0575	0.9980	0.9452	196.9496	ROA MEAN

PYTHON	Random forest	8.7375	2.2407	2.9559	0.9365	0.1336	225.8342	ROA MEAN
R	Random forest	8.6935	2.2506	2.9485	0.9281	0.9242	190.1601	ROA MEAN
PYTHON	SVM	9.3750	2.3558	3.0618	1.0048	0.1405	176.8769	ROA MEAN
R	SVM	9.0222	2.2912	3.0037	0.9631	0.9409	175.2122	ROA MEAN

Tabla 22: Modelos de regresión: Resultados Holdout. Elaboración propia.

Se evalúan los modelos en función del RMSE, pues este indicador evalúa el error de predicción frente a los valores reales. Resulta evidente que los modelos estudiados no presentan una gran capacidad de predicción para la variable respuesta “Roa mean”, presentando valores demasiado elevados en sus medidas de bondad de ajuste. Por otro lado, los modelos de Random forest destacan como los más efectivos y precisos, pues sus RMSE son inferiores a los demás.

4.9.2 Hold out repetido

Puesto que la técnica hold out repetido realiza el cálculo de las medidas de bondad de ajuste en un total de 100 iteraciones, representamos en la Tabla 23 los resultados obtenidos de las medias del total de iteraciones.

MEDIAS MEDIDAS DE BONDAD DEL AJUSTE

MODELO	MAE	RMSE	MSE	NMSE	NMAE
Lm	2.40340	3.197766	10.237633	1.093697	1.051686
cart	2.286450	3.055400	9.343597	0.997494	1.000311
rf	2.312542	3.078065	9.483041	1.012661	1.011802
Svm	2.270759	3.041237	9.256747	0.988358	0.993491

Tabla 23: Modelos de regresión: Resultados Hold out repetido. Elaboración propia.

Del mismo modo que vimos anteriormente en los resultados del Hold out, los resultados obtenidos de los diferentes modelos estudiados no presentan una gran capacidad de predicción para la variable respuesta “Roa mean”.

En la Figura 50 mostramos gráficamente la comparación de los resultados de las medias de la métrica RMSE para cada modelo.

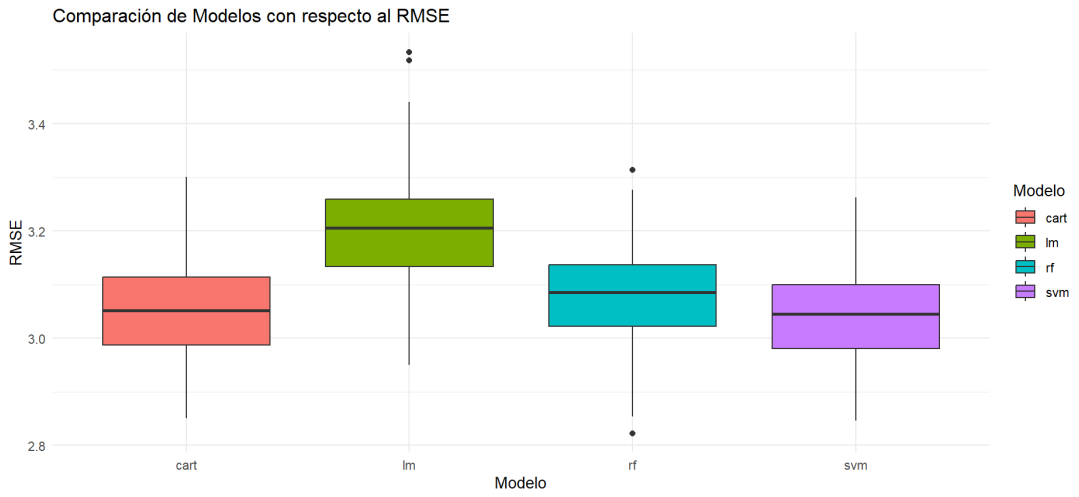


Figura 50: Gráfico comparación de modelos - RMSE

De este modo, observamos tanto numéricamente, como gráficamente, que el error de predicción es muy similar, ya que todos los modelos presentan valores muy similares del RMSE, destacando el modelo de regresión lineal como el peor modelo con el valor de RMSE más alto.

4.9.3 TEST ANOVA

Por otro lado, para la comparación de más de dos modelos aplicamos la técnica ANOVA para verificar si el modelo es un factor que afecta a los resultados, obteniendo los siguientes resultados en la Tabla 24:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
modelo	3	1.5292	0.5097	271.82	<2e-16 ***
bloque	99	3.0668	0.0310	16.52	<2e-16 ***
Residuals	297	0.5569	0.0019		

Signif. codes: 0‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Tabla 24: Test ANOVA

Puesto que los resultados de los modelos son significativamente diferentes, con un nivel de significación del 5% (p-valor = 2×10-16 > 0.05), posteriormente realizamos el contraste de intervalos de Tukey, donde se calculan todas las diferencias para observar cuales difieren de 0 (Tabla 25).

	diff	wr	upr	p adj
lm-cart	0.14236538	0.126542879	0.158187886 0.	0000000
rf-cart	0.02266517	0.006842667	0.038487673	0.0014539
svm-cart	-0.01416366	-0.029986160	0.001658847	0.0974159
rf-lm	-0.11970021	-0.135522716	-0.103877709	0.0000000
svm-lm	-0.15652904	-0.172351543	-0.140706536	0.0000000
svm-rf	-0.03682883	-0.052651330	-0.021006324	0.0000000

Tabla 25: Contraste intervalos TUKEY

Como vemos, existen diferencias significativas entre los modelos, excepto entre el modelo svm y el cart con un p valor ajustado = 0.0974159

A continuación, vemos en la Figura 58 el contraste de intervalos de Tukey, donde se observa cuales difieren de 0, encontrando diferencias en todos, excepto entre el svm y el cart, tal y como nos indicaba en la tabla anterior.

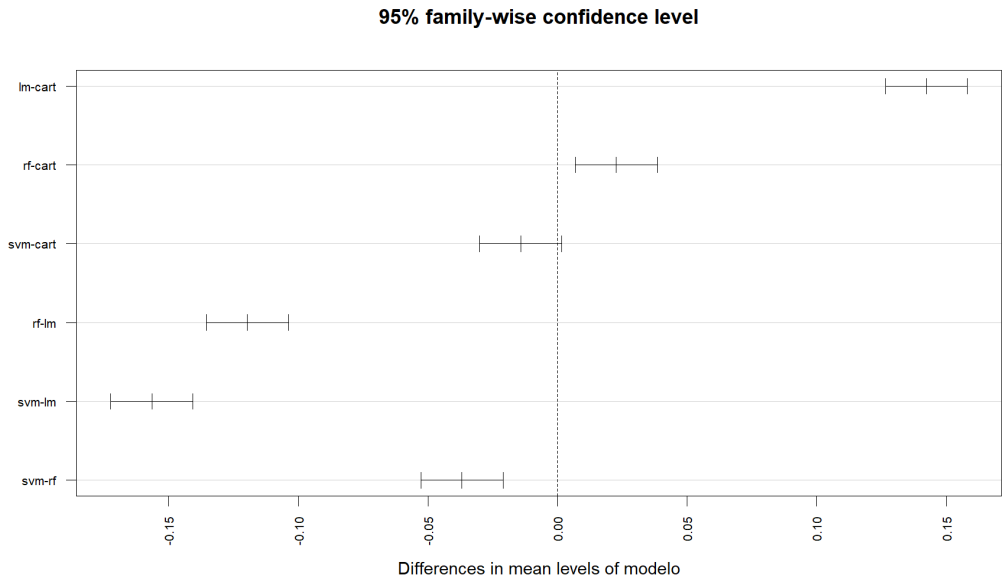


Figura 51: Prueba de Tukey para evaluar las diferencias de medias entre modelos

Además, estos resultados indican que, en términos de RMSE, el modelo svm es el mejor, seguido por rf, cart y finalmente lm, aunque son muy similares en cuanto a precisión, con valores muy similares en sus medidas de bondad de ajuste.

5. Conclusiones

Tras realizar el análisis para responder a los objetivos propuestos, a continuación, se presentan las conclusiones obtenidas.

El objetivo general de este estudio fue investigar la posible relación entre el rendimiento financiero (ROA) de las bodegas e indicadores de huella digital de bodegas de vino de España. Por ello la investigación se centró en un principio en la exploración y búsqueda de indicios de interés que permitan relacionar qué variables o factores presentes en las páginas web de dichas bodegas tienen relación o pueden ser indicadores del nivel de competitividad medida a través del ROA de la empresa. Posteriormente, los modelos multivariantes y de minería de datos confirman estas relaciones y nos permiten extraer información relativa a los indicadores de huella digital.

Al explorar la muestra de empresas utilizando los indicadores de huella digital (variables o factores presentes en las páginas web de dichas bodegas), los resultados del PCA obtenidos ponen de manifiesto la importancia de varios indicadores de huella digital mediante la presencia de palabras clave como "_valor", "_marca", "_vino", "_bodega", "_calidad", "_propiedad", "_derechos", así como del papel relevante de las redes sociales mediante la presencia de los indicadores "hrefwords". Además, se agrupan las variables "keywords" "_productividad", "competitividad", "_eficiencia", "_innovación", "_novedad", "_marketing", "_posicionamiento", "_investigacion_y_desarrollo", "_sistemas_de_información" o "_flexibilidad", entre otras, siendo indicadores que guardan relación con el desarrollo, la competitividad y la productividad de la empresa.

Los modelos de regresión implementados tanto en Python como en R presentan resultados similares. Además, ha sido posible replicar todas las técnicas con ambos lenguajes. En ambos casos, los modelos de regresión implementados no presentan resultados demasiado satisfactorios para predecir la variable respuesta "Roa Mean". Aunque si ha sido posible la comparación de los modelos, observando que el peor de los ellos corresponde al de regresión lineal. Mientras que, los modelos de svm, random forest y árbol de regresión, presentan valores muy similares en sus medidas de bondad de ajuste y serían ligeramente mejores que el modelo de regresión lineal.

Por otro lado, mediante los modelos multivariantes y de minería de datos implementados, se ha podido seleccionar las variables con mayor importancia en el modelo, entre las que destacan algunas "keywords" como, por ejemplo, marketing, tecnología, rendimiento,

recursos, exportación, eficiente, marca, moderna, redes y las “hrefwords”, en relación con su existencia en las diferentes redes sociales.

Por lo que, a partir de los resultados del análisis PCA y la importancia de las variables en los diferentes modelos, se ha confirmado la relación entre indicadores digitales como la innovación, el marketing y la investigación, y la presencia en las redes con competitividad, productividad y eficiencia empresarial. Por otro lado, también ha sido posible intuir la relación de la competitividad de la bodega con su marca y la calidad del vino pues han sido palabras clave importantes en los modelos.

Finalmente, hay que subrayar la dificultad de encontrar modelos que expliquen adecuadamente la rentabilidad de las bodegas, concepto que depende de muchos factores que no están siempre disponibles. Sin embargo, los indicadores de huella digital a pesar de su limitado valor explicativo de la rentabilidad tienen una ventaja fundamental pues son datos fácilmente accesibles y baratos de conseguir. Este avance ayuda a las bodegas en aspectos estratégicos de toma de decisiones y abre posibilidades de aplicación en diversos sectores industriales, promoviendo una visión más amplia, comprensión y adopción de la analítica digital en el análisis competitivo. Así pues, abren una línea prometedora de futuros estudios que mejoren los modelos y que se apliquen a otros tipos de empresas y en otros espacios geográficos.

Bibliografía

Abuín, J. R. (2007). *Regresión lineal múltiple. IdEyGdM-Ld Estadística, Editor*, 32.

Alonso, M. H. (2023). La Transformación Digital como Oportunidad para el Sector Vitivinícola Riojano en el Enoturismo (Tesis doctoral). Universidad Internacional de La Rioja.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational statistics & data analysis*, 52(4), 2249-2260.

Boqué, R., & Maroto, A. (2004). El análisis de la varianza (ANOVA) 1. Comparación de múltiples poblaciones. *Técnicas de Laboratorio*, 294, 680-683.

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

Blázquez Soriano, M. D. (2020). Design and Evaluation of Web-Based Economic Indicators: *A Big Data Analysis Approach* (Tesis doctoral). Universitat Politècnica de València.

Burge, R. C. (2023). *Forecasting Urgent Care Patient Volume Using Traditional Techniques and Machine Learning to Improve Resource Allocation* (Tesis doctoral, The George Washington University). Accedido el 18 de Diciembre de 2023.

Bezdek, J. C., Coray, C., Gunderson, R., & Watson, J. (1981). *Detection and characterization of cluster substructure: Linear structure: Fuzzy c-lines.* SIAM Journal on Applied Mathematics, 40(2), 339-357

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984*), Classification and Regression Trees*, CHAPMAN & HALL/CRC, Boca Raton.

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

Castro Barrantes, L. E. (2023). *Estudio de la relación entre indicadores de competitividad y huella digital en bodegas valencianas mediante técnicas multivariantes y de minería de datos*. https://riunet.upv.es/handle/10251/192896. Accedido el 18 de Octubre de 2023.

De La Fuente, S. (2011). *Análisis conglomerados*. Facultad de Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid (UAM). Madrid, España.

De La Hoz Suárez, B., Ferrer, M. A., & De La Hoz Suárez, A. (2008). Indicadores de rentabilidad: herramientas para la toma decisiones financieras en hoteles de categoría

media ubicados en Maracaibo. *Revista de Ciencias Sociales*, 14(1), 88-109. https://www.redalyc.org/pdf/280/28011673008.pdf

Esbensen, K. H., & Geladi, P. (2020). 2.02-Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice. *In Comprehensive chemometrics: Chemical and biochemical data analysis* (Vol. 2).

Fernández, D. S. (2016). Big Data y la estadística oficial: retos. Indice*: Revista de Estadística y Sociedad,* (68), 14-17.

Ferrer-Riquelme, A. (2009). Statistical Control of Measures and Processes. *Comprehensive Chemometrics; Elsevier: Amsterdam, The Netherlands*,pages 97–126.

FEV. (2022). *El sector en cifras*. https://www.fev.es/sector-cifras/. Accedido el 11 de Octubre de 2023.

Franco-Árcega, A., Sobrevilla-Sólis, V. I., de Jesús Gutiérrez-Sánchez, M., García-Islas, L. H., Suárez-Navarrete, A., & Rueda-Soriano, E. (2021). Sistema de enseñanza para la técnica de agrupamiento k-means. *Pädi Boletín Científico de Ciencias Básicas e Ingenierías Del ICBI*, 9(Especial), 53-58.

Frutos Serrano, Sergio (2021-2022). *Comparación entre XGBoost y Regresión Lineal Múltiple para la predicción de la evolución del precio de las acciones*. https://docta.ucm.es/rest/api/core/bitstreams/40882ad0-7a8b-425b-a455-ab723770c351/content. Accedido el 20 de Febrero de 2024.

Gomez-Conde, J., Lopez-Valeiras, E., Gonzalez-Sanchez, M. B., & Alguacil, M. (2013). El ajuste contingente entre los sistemas contables de gestión y la estrategia. Un análisis empírico en el sector enoturístico. *Revista Europea de Dirección y Economía de la Empresa*, 22(2), 89-96.

Gómez, M., & Molina, A. (2013). Estrategias de gestión del valor de marca en los destinos enoturísticos. *Revista Europea de Dirección y Economía de la Empresa*, 22(2), 69-79.

González, L. D., Vega, V. D. R. S., Melgar, D. C., Velázquez, M. D. C. C., Nav, I. S. H., & Castill, M. L. S. (2016). Utilización del BMA en el modelo de regresión logística y su comparación con otros criterios de selección de modelos. *Investigación Operacional*, 37(1).

Granados, R. M. (2016). *Modelos de regresión lineal múltiple*. Granada, España: Departamento de Economía Aplicada, Universidad de Granada.

Gupta, S., & Sedamkar, R. R. (2020). Machine learning for healthcare: Introduction. In *Machine learning with health care perspective: Machine learning and healthcare*, 1-25.

Hernández Córdoba, E. (2019). *Selección del mejor conjunto de Regresión*.

Jaramillo, H. A. L., Pinos, C. A. E., Sarango, A. F. H., & Román, H. D. O. (2023). Histograma y distribución normal: Shapiro-Wilk y Kolmogorov Smirnov aplicado en SPSS: Histogram and normal distribution: Shapiro-Wilk and Kolmogorov Smirnov applied in SPSS. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, 4(4), 596-607.

Judge, T., Robbins, S. P., & SOBRAL, F. (2009). *Comportamiento organizacional. Estados Unidos: Pearson.*

Kourti, T., & MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of quality technology*, 28(4), 409-428.

Lara Parra, A. C., & Torres Parra, Y. A. (2015). Análisis de competitividad bajo criterios e indicadores financieros del sector de las confecciones de Bogotá para el periodo 2009–2013.

Lawson, R., y Peter, J. (1990). New Index for Clustering Tendency and Its Application to Chemical Problems. *Journal of Chemical Information and Modeling*, 30(1), 36-41.

Larico Soncco, Y. (2021). *Caracterización de clientes de la marca TGI Fridays, mediante los algoritmos K-Means y K-Medoids.* (Tesis de maestría, Universidad Nacional Mayor de San Marcos). https://cybertesis.unmsm.edu.pe/backend/api/core/bitstreams/95e7ab0b-36c7-4f79-890c-a44ea7773903/. Accedido el 12 de Marzo de 2024.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.

López Sánchez, V. (2019). *Aplicación y comparativa de cuatro modelos de clustering para datos GTEx. (Tesis de maestría, Universitat Oberta de Catalunya*). http://hdl.handle.net/10609/90626. Accedido el 21 de Enero de 2024.

López-Guzmán Guzmán, T. J., Vázquez de la Torre, G. M., & Caridad y Ocerín, J. M. (2008). Análisis econométrico del enoturismo en España: un estudio de caso. *Estudios y perspectivas en turismo*, 17(2), 34-54.

Martínez Castellar, O. (2022). *Activos intangibles y huella digital en el sector vitivinícola* (Doctoral dissertation, Universitat Politècnica de València). https//www.riunet.upv.es/bitstream/handle/10251/184731/Martinez%20-

%20Activos%20Intangibles%20y%20huella%20digital%20en%20el%20sector%20vitivinicola.pdf?sequence=1

Medina, F. X. (2017, July). Reflexiones sobre el patrimonio y la alimentación desde las perspectivas cultural y turística. *In Anales de antropología* (Vol. 51, No. 2, pp. 106-113). No longer published by Elsevier.

Navarro, O. (2009, June). Selección de variables en regresión componentes principales. *In 7th Latin American and Caribbean Conference for Engineering and Technology* (pp. 1-8).

Niño, F. A. P. (2020). *Introducción al análisis clúster: una aplicación en la clasificación de campos petroleros.* Universidad Industrial de Santander.

Pal, N. R., Bezdek, J. C., & Hathaway, R. J. (1996). Aprendizaje competitivo secuencial y algoritmos difusos de agrupamiento de c-medias. *Redes Neuronales*, 9(5), 787-796.

Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J. P., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista de teledetección*, (44), 55-65.

Puebla, J. G. (2018). Big Data y nuevas geografías: la huella digital de las actividades humanas. *Documents d'anàlisi geogràfica*, 64(2), 195-217.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1), 21-33.

Rodríguez Zamora, E. E. (2022). *Influencia de la actividad digital en la competitividad de las bodegas valencianas.* (Tesis de máster, Universidad Politécnica de Valencia). https://riunet.upv.es/handle/10251/186563. Accedido el 5 de Diciembre de 2023.

Soto Andreu, S. (2023). Análisis de la relación entre la competitividad de las empresas y su presencia online (Tesis doctoral, Universidad Politécnica de Valencia).

Therneau, T. M., & Atkinson, E. J. (1997). *An introduction to recursive partitioning using the RPART routines* (Vol. 61, p. 452). Mayo Foundation: Technical report.

Tobón Perilla, N., Urquía Grande, E., & Cano Montero, E. I. (2022). ¿Qué factores de gestión interna favorecen la competitividad de las pymes? Evidencia en Colombia. *Revista Universidad y Empresa*, 24(42). https://doi.org/10.12804/revistas.urosario.edu.co/empresa/a.11102

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

Van Dijk, B., & Informa, S. A. (2017). SABI. *Sistemas de Análisis de Balances Ibéricos*. Accedido el 19 de Diciembre de 2023.

Anexos

Anexo I. Listado de palabras clave (huella digital)

Las palabras corresponden al siguiente listado: 'activos', 'adaptabilidad', 'agilidad organizacional', 'alianzas estrategicas', 'analisis de datos', 'bodega', 'cadena de suministro', 'calidad', 'cambio', 'capacidad', 'capacidad de analisis', 'capacidad financiera', 'competencia', 'competitividad', 'compromiso con la calidad', 'comunicacion efectiva', 'creatividad', 'crecimento', 'crecimiento sostenido', 'cultura organizacional', 'derechos', 'diferenciar', 'diversificacion', 'economia de escala', 'eficiencia', 'eficiente', 'empatia con el cliente', 'emprendimiento interno', 'estrategia competitiva', 'expansion internacional', 'experiencia del cliente', 'exportacion', 'fidelizacion de clientes', 'flexibilidad', 'formacion continua', 'gestion de crisis', 'gestion del cambio', 'gestion del talento', 'gestion de riesgos', 'ideas', 'imagen corporativa', 'inclusion y diversidad', 'iniciativa', 'inmaterial', 'innovacion', 'inteligencia competitiva', 'investigacion y desarrollo', 'liderazgo efectivo', 'localizacion', 'logistica integrada', 'marca', 'marketing', 'mentalidad global', 'mercado', 'moderna', 'negociacion', 'novedad', 'nuevo', 'optimizacion de procesos', 'orientacion al resultado', 'original', 'pais', 'participacion de mercado', 'patente', 'penin', 'perfeccionar', 'planificacion estrategica', 'popular', 'posicionamiento', 'posicionamiento de marca', 'potencia', 'potente', 'practicas eticas', 'precios competitivos', 'proactividad', 'produccion', 'productividad', 'propiedad', 'reconocimiento de marca', 'recursos', 'redes', 'redes sociales y digitales', 'relaciones publicas', 'rendimiento', 'rendimiento financiero', 'rentabilidad', 'resiliencia', 'responsabilidad corporativa', 'responsabilidad social', 'robert parker', 'segmentacion de mercado', 'servicio al cliente', 'sistemas de informacion', 'sostenibilidad', 'tecnologia', 'trabajo en equipo', 'transparencia', 'valor', 'valor agregado', 'ventaja', 'vino', 'vision de futuro’.

Anexo II. Código Python, Código R y Base de Datos

El código de ambos lenguajes y la base de datos utilizada para este trabajo se incluye en el siguiente enlace de GitHub: https://github.com/JorgeVizu/TFM-UPV.git

Anexo III. Resultados modelos de regresión con R

1. Regresión múltiple – R

En la Tabla 26 se presentan los resultados o medidas de bondad de ajuste del modelo de regresión lineal múltiple:

RMSE	MAE	MSE	NMSE	NMAE	MAPE
3.082	2.375	9.500	1.014	0.975	206.525

Tabla 26: Medidas de bondad de ajuste modelo de regresión lineal múltiple - R

A continuación, se presentan los resultados de los coeficientes obtenidos en el modelo:

Coef	Estimate Std.	Error	t value	P value
keywords_activos	0.717	0.375	1.910	0.056
keywords_calidad	-0.549	0.327	-1.680	0.093
keywords_competitividad	-0.9178	0.3628	-2.530	0.0116 *
keywords_derechos	0.7492	0.3027	2.476	0.0135 *
keywords_eficiencia	0.8272	0.3239	2.554	0.0108
keywords_eficiente	-0.5185	0.3018	-1.718	0.0861
keywords_gestion_del_talento	4.3245	2.2701	1.905	0.0571
keywords_innovacion	0.3724	0.2498	1.491	0.1364
keywords_localizacion	-0.3967	0.2495	-1.590	0.1122
keywords_marca	0.4879	0.2500	1.952	0.0512
keywords_moderna	-0.4966	0.2557	-1.942	0.0524
keywords_optimizacion_de_procesos	2.8377	1.6507	1.719	0.0859
keywords_potente	0.3804	0.2322	1.638	0.1017
keywords_servicio_al_cliente	-1.0371	0.5376	-1.929	0.0540
keywords_sistemas_de_informacion	1.2492	0.5377	2.323	0.0204 *

keywords_trabajo_en_equipo	0.8471	0.4851	1.746	0.0811
keywords_valor	-0.6068	0.2578	-2.354	0.0188 *
hrefwords_twitter	-0.3599	0.2234	-1.611	0.1075

Tabla 27: Coeficientes Modelo de Regresión lineal Múltiple – R

El mejor modelo de regresión lineal obtenido presenta 18 variables. Observamos también la influencia de las variables sobre la respuesta, marcadas con un *asterisco, todas aquellas variables con p valor bajo, por debajo del nivel de significancia 0,05% (keywords_competitividad, keywords_derechos, keywords_eficiencia, keywords_sistemas_de_informacion, keywords_valor), son estadísticamente significativas en el modelo, lo que indica que tienen un efecto influyente sobre la variable respuesta estudiada.

Además, cabe destacar que la mayoría de las variables seleccionadas para el modelo, coinciden con las que destacan en la realización del PCA para las dos primeras componentes principales, siendo las variables que más explican en la variabilidad de los datos, destacando las marcadas en el modelo como variables con un mayor nivel de significancia,"hrefwords_twitter" y algunas "keywords" relevantes como "_valor, "_marca", "_derechos", "_competitividad", "_eficiencia" y "_sistemas_de_información", además del resto incluidas en el modelo como por ejemplo, "_calidad", "_propiedad", "_innovación", "_marketing", "_posicionamiento", "_investigacion_y_desarrollo", o "_flexibilidad", entre otras. De este modo, verificamos que estas variables incluidas en el modelo y agrupadas en el PCA como variables con una gran relación en las 2 primeras componentes principales son las variables con una mayor incidencia sobre la variable respuesta roa_mean.

Resultados de las pruebas para la evaluación de hipótesis asociadas al modelo:

Del mismo modo, comprobamos la normalidad y vemos que los residuos no siguen una distribución normal según las pruebas Shapiro-Wilk y Kolmogorov-Smirnov, como sigue (Tabla 28 y 29):

1- Normalidad:

Resultado de la prueba de Shapiro-Wilk	
Estadístico de prueba	0.986
P-valor	9.566e-08

Tabla 28: Resultado de la prueba de Shapiro-Wilk

Resultado de la prueba de Kolmogorov-Smirnov	
Estadístico de prueba	0.209
P-valor	0

Tabla 29: Resultado de la prueba de Kolmogorov-Smirnov

2- Homocedasticidad:

Mientras que, para la evaluación de la homocedasticidad se observó el gráfico de residuos frente a predichos en la Figura 52:

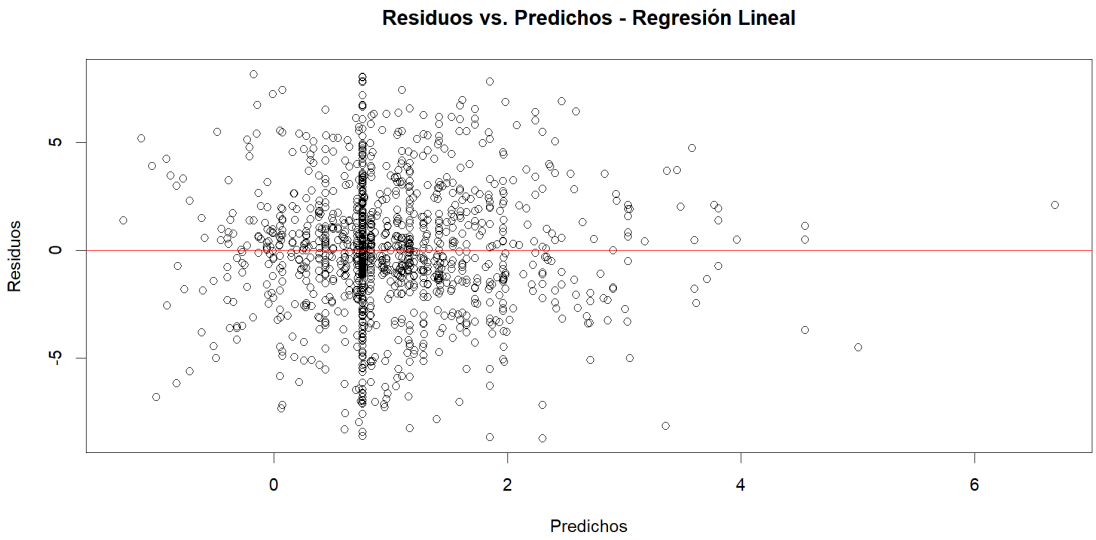


Figura 52: Gráfico Residuos vs Valores predichos – Regresión multiple con R

Donde se observó la misma distribución aleatoria de los datos, sin un patrón en función de los residuos, por lo que se confirma la homocedasticidad.

3 – Autocorrelación:

Posteriormente se analizó la autocorrelación mediante el test Durbin-Watson (Tabla 30):

Resultados del Test de Durbin-Watson	
Estadístico de Durbin-Watson	2.014

Tabla 30: Test de Durbin-Watson

Confirmando así, que los residuos no tienen autocorrelación significativa.

4 – Multicolinealidad:

Por último, de igual modo, verificamos la multicolinealidad entre las variables, ya que los valores de la VIF son bastante bajos (Tabla 31).

Variable	VIF
keywords_activos	1.461596
keywords_calidad	2.196779
keywords_competitividad	1.207635
keywords_derechos	2.211783
keywords_eficiencia	1.371363
keywords_eficiente	1.292546
keywords_gestion_del_talento	1.154424
keywords_innovacion	1.432308
keywords_localizacion	1.181351
keywords_marca	1.722484
keywords_moderna	1.265380
keywords_optimizacion_de_procesos	1.218302
keywords_potente	1.409162
keywords_servicio_al_cliente	1.094786
keywords_sistemas_de_informacion	1.185376
keywords_trabajo_en_equipo	1.110268
keywords_valor	1.815743
hrefwords_twitter	1.373252

Tabla 31: Resultados VIF – Evaluación multicolinealidad

Finalmente, en la Figura 53 observamos mediante el gráfico de valores observados frente a valores predichos.

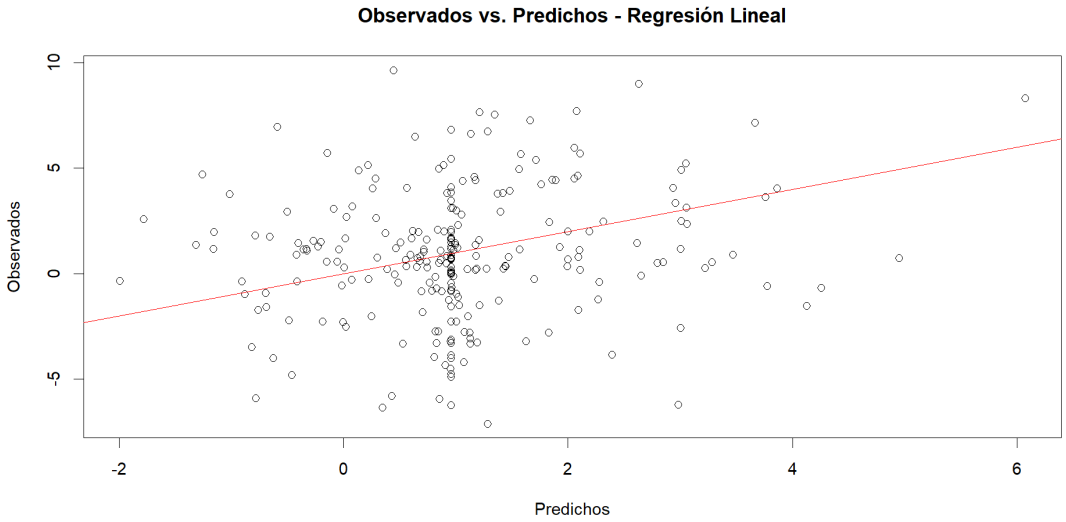


Figura 53: Valores observados vs Valores predichos

Del mismo modo que en el modelo realizado con Python, y tal y como indican exactamente los mismos resultados en sus medidas de bondad de ajuste, el modelo no sería un buen modelo para predecir la variable respuesta “Roa_mean”, con resultados muy elevados en sus métricas. Por lo que se observan los puntos distribuidos de forma aleatoria.

2. Arbol de regresión – R

Por otro lado, se presentan en la Tabla 32 los resultados obtenidos para el modelo árbol de regresión en R.

RMSE	MAE	MSE	NMSE	NMAE	MAPE
3.387	2.714	11.472	1.224	1.114	310.745

Tabla 32: Medidas de bondad de ajuste modelo árbol de regresión – R

Observamos que los resultados son ligeramente diferentes a los obtenidos en Python, esto puede ser justificado ya que los algoritmos utilizados pueden diferir ligeramente entre las bibliotecas de Python (scikit-learn) las de R (rpart), siendo que pueden utilizar diferentes criterios de división.

Los resultados son similares, obteniendo valores elevados que indican que el modelo no es preciso al realizar sus predicciones. En la Figura 54 vemos el dibujo del árbol obtenido.

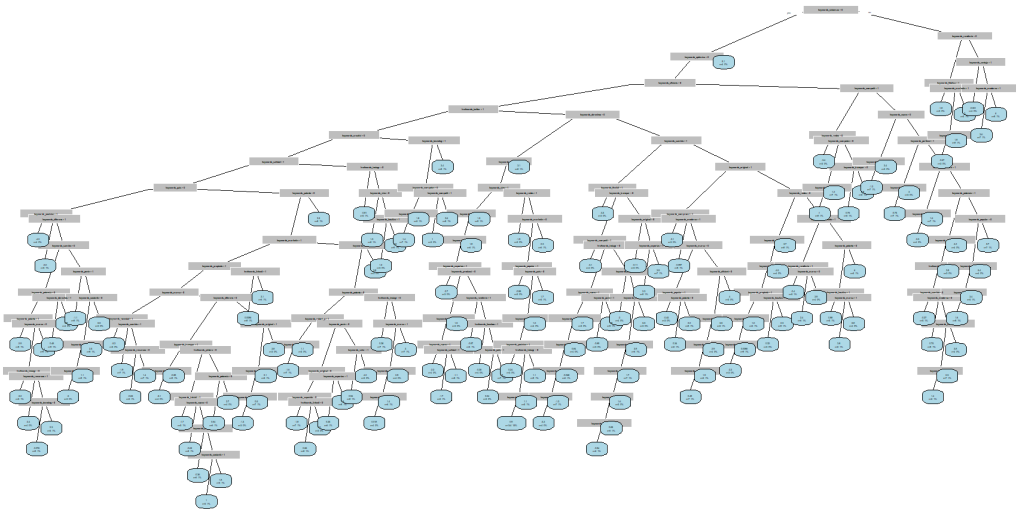


Figura 54: Arbol de regresión - R

Además, en la Tabla 33 vemos las 10 variables más importantes y en la Figura 55 el gráfico de barras de su representación.

Variable	Overall
keywords_localizacion	3.614916
keywords_potente	3.496224
keywords_tecnologia	2.542863
keywords_rendimiento	2.501208
keywords_recursos	2.305230
keywords_moderna	2.295072
keywords_penin	2.217652
keywords_exportacion	2.173460
hrefwords_linkedin	2.168602
keywords_popular	2.077563

Tabla 33: Importancia de las variables - R

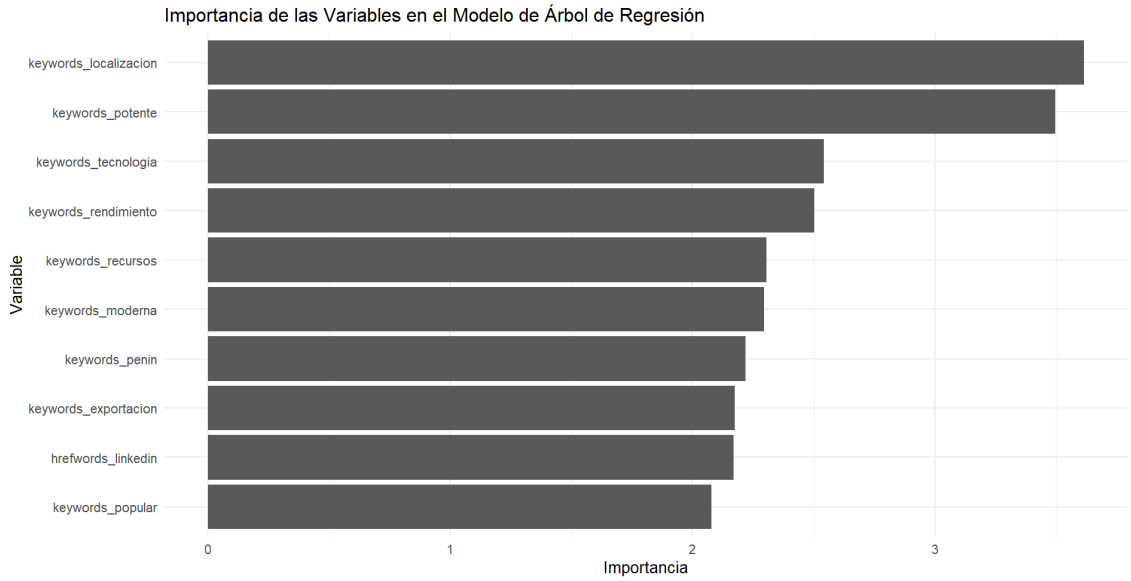


Figura 55: Gráfico de barras importancia de las variables - R

En cuanto a la importancia de las variables, observamos las variables con mayor importancia en el modelo, entre las que destacan algunas como, tecnología, rendimiento, recursos o exportación, entre otras.

Además, vemos que alguna de ellas coincide con las de mayor importancia en el modelo de Python, como por ejemplo “keyword_rendimiento”, “keyword_moderna” o “keyword_popular”.

3. Bosques aleatorios (Ramdom forest) – R

Mientras que en la Tabla 34, se presentan los resultados del modelo de bosques aleatorios implementado en R.

RMSE	MAE	MSE	NMSE	NMAE	MAPE
2.948	2.250	8.693	0.928	0.924	190.160

Tabla 34: Medidas de bondad de ajuste en el modelo Ramdom forest - R

En la Figura 56 se presentan los resultados obtenidos para ambas medidas de importancia de las variables.

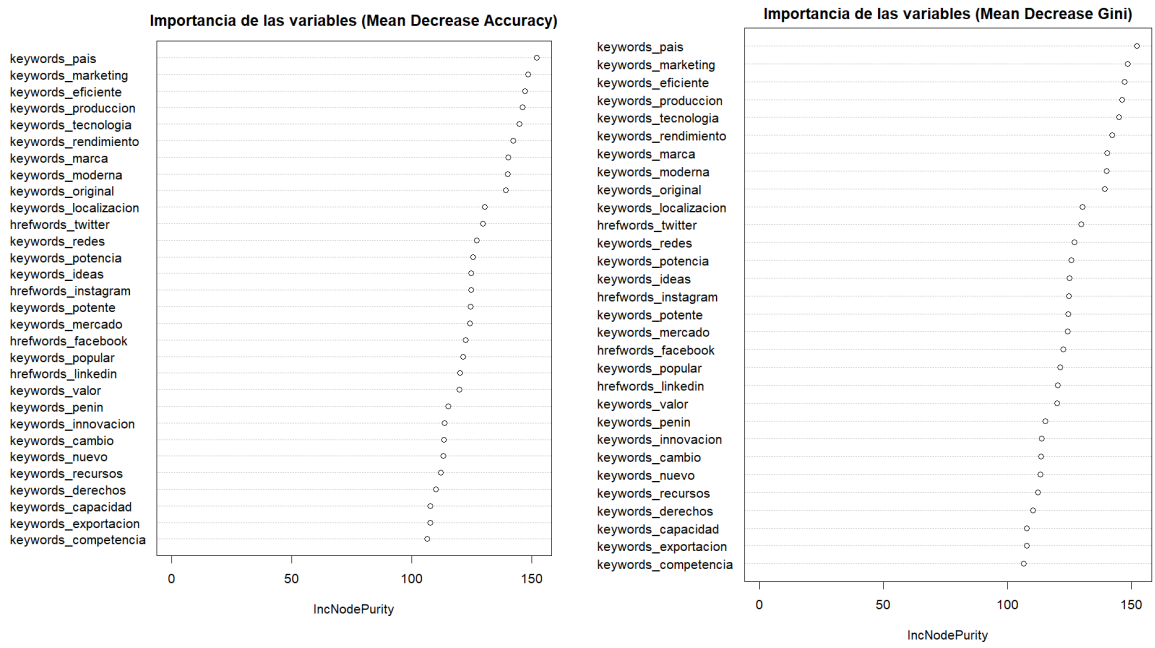


Figura 56: Medidas de importancia de variables: Mean Decrease Accuracy y Mean Decrease

Del mismo modo que en los árboles de regresión, observamos que los modelos de Python y R difieren muy ligeramente, debido a la implementación de los diferentes algoritmos y sus respectivas librerías, siendo muy similares y mostrando unos resultados no satisfactorios como predictores de la variable respuesta. Aunque también podemos extraer información similar, visualizando las variables de mayor importancia, como las keywords (marketing, rendimiento, eficiente, tecnología, marca, moderna, redes) y las hrefwords con relación a su existencia en las diferentes redes sociales.

4. Máquinas de soporte vectorial (SVM) – R

Mientras que en la Tabla 35, se presentan los resultados del modelo SVM implementado en R.

RMSE	MAE	MSE	NMSE	NMAE	MAPE
3.003	2.291	9.022	0.963	0.940	175.212

Tabla 35: Medidas de bondad de ajuste en el modelo SVM - R

Por otro lado, en la Figura 57 observamos mediante el gráfico de valores observados frente a valores predichos:

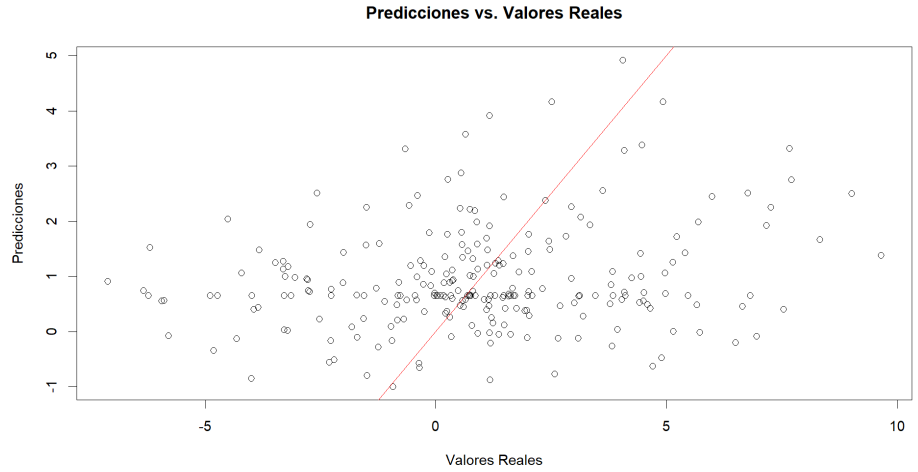


Figura 57: Gráfico de valores predichos vs valores reales con R

Finalmente, en la Figura 58 observamos el gráfico de residuos frente a valores predichos:

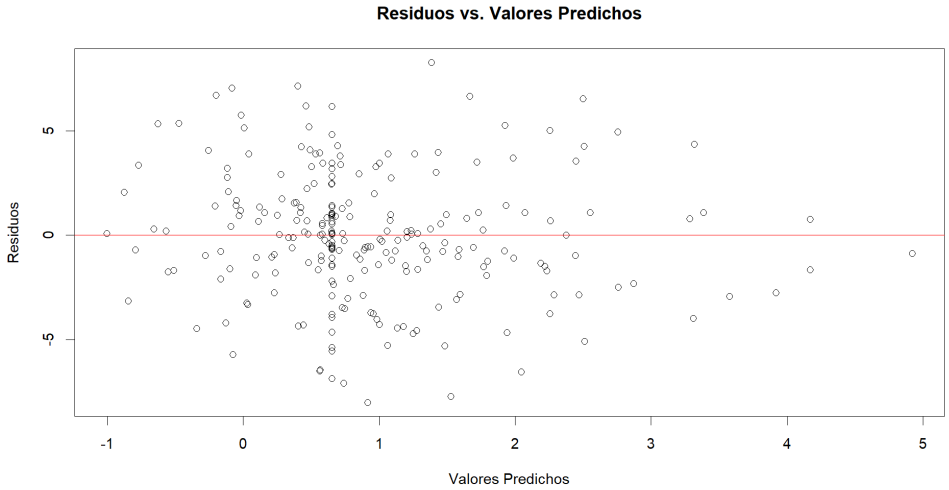


Figura 58: Gráfico de residuos vs valores predichos con R

Igual que en Python observamos un MSE muy elevado (9.0222) y los puntos en los gráficos distribuidos aleatoriamente, mostrando una gran distancia entre valores reales y predichos, así que tampoco sería un buen modelo para predecir la variable respuesta.

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ANEXO IV. RELACIÓN DEL TRABAJO CON LOS OBJETIVOS DE DESARROLLO SOSTENIBLE DE LA AGENDA 2030

Anexo al Trabajo de Fin de Grado y Trabajo de Fin de Máster: Relación del trabajo con los Objetivos de Desarrollo Sostenible de la agenda 2030.

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

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

Descripción de la alineación del TFG/TFM con los ODS con un grado de relación más alto.

***Utilice tantas páginas como sea necesario.

Este Trabajo Final de Máster tiene un alto grado de relación con el Objetivo de Desarrollo Sostenible número 9 de la Agenda 2030: Industria, innovación e infraestructuras. Se relaciona con este objetivo dado el interés en la innovación de la industria (industria vitivinícola) utilizando herramientas, técnicas multivariantes y de minería de datos, para analizar la relación de la huella digital y la competitividad (medida como el rendimiento financiero sobre los activos, ROA) en bodegas de vino de España, de manera que se pueda facilitar la toma de decisiones para favorecer a la industria e incrementar su desarrollo económico.

Editor: Secretaria General / UPV · D.L.: V-5092-2006 · ISSN: 1887-2298 · www.upv.es/secgen