UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Estudio de la relación entre popularidad online y cuota de mercado. Aplicación al sector agroalimentario

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Lopez Llorens, Ivan

Tutor/a: Doménech i de Soria, Josep

CURSO ACADÉMICO: 2023/2024

Resumen

El paradigma de la sociedad actual en cuanto a los productos que consume y como éstos se dan a conocer está variando mucho desde la llegada de las redes sociales. Este motivo junto con el auge de las tecnologías digitales o la publicidad online están cambiando la visión de las empresas en cómo enfocar las estrategias de venta y conseguir el mayor impacto en el consumidor para hacerse con la mayor cuota de mercado. Por ello, es muy importante conocer cómo las redes sociales y los medios digitales influyen en las decisiones que toman los consumidores a la hora de decantarse por uno u otro producto.

En este Trabajo de Fin de Grado se afrontará este cambio de paradigma desde diversos puntos considerados dentro del "mundo digital": distintas redes sociales o búsquedas en navegadores web, entre otras; en concreto dentro del sector agroalimentario. Entre los diversos objetivos el principal es buscar la relación entre métricas asociadas a la popularidad online y la cuota de mercado que empresas del sector obtienen. Para ello se utilizarán datos de diversas fuentes de información como Google Trends, Kantar WorldPanel o SocialBlade extraídos mediante técnicas de web scrapping o utilizando scripts de Python para obtener los datos desde las propias APIs de las plataformas. Tras la obtención de datos y una exhaustiva limpieza del DataFrame resultante, se abordará el problema desde diversos puntos de vista: visualizaciones de gráficos de dispersión, correlaciones entre variables asociadas a la popularidad online o la creación de un modelo de regresión utilizando varios algoritmos serán herramientas que se utilizarán para explicar las relaciones.

Palabras clave: sector agroalimentario, Kantar WorldPanel, Google Trends, API, web scrapping, correlación, modelos de regresión, redes sociales.

Abstract

The paradigm of today's society in terms of the products it consumes and how these are made known is changing a lot since the arrival of social networks. This, together with the rise of digital technologies and online advertising, is changing the vision of companies on how to focus sales strategies and achieve the greatest impact on the consumer in order to gain the greatest market share. For this reason, it is very important to understand how social networks and digital media influence the decisions that consumers make when deciding on one product or another.

This Project will address this paradigm shift from different points considered within the 'digital world': different social networks or searches in web browsers, among others; specifically within the agri-food sector. Among the various objectives, the main one is to search for the relationship between metrics associated with online popularity and the market share that companies in the sector obtain. To do this, data from various sources of information such as Google Trends, Kantar WorldPanel or SocialBlade will be used, extracted using web scraping techniques or using Python scripts to obtain the data from the platforms' own APIs. After obtaining the data and an exhaustive cleaning of the resulting DataFrame, the problem will be approached from different points of view: visualisations of scatter plots, correlations between variables associated with online popularity or the creation of a regression model using various algorithms will be used to explain the relationships.

Keywords: agri-food sector, Kantar WorldPanel, Google Trends, API, web scraping, correlation, regression models, social networks.

Tabla de contenido

Índice de ilustraciones

Índice de tablas

1. INTRODUCCIÓN

1.1 Resumen

Las Redes Sociales y otros factores derivados de la sociedad moderna están cambiando el paradigma de las empresas alrededor de todo el mundo. Tener seguidores, impacto en ellos y popularidad online son ahora tres características que toda empresa (y particular) busca con gran determinación. (Hill, Troshani, & Chandrasekar, 2017)

Desde hace varios años atrás, desde el nacimiento de Internet a finales de la década de los 70, y más marcado aún con el continuo crecimiento de las redes sociales, las personas se han vuelto más dependientes de lo que ven en las plataformas online. Hace cuarenta años, para que una empresa se diera a conocer, tenía que crear buena publicidad entre sus potenciales clientes cercanos y que, con el boca a boca, acabasen siendo populares en su zona. De esta forma conseguirían sus clientes potenciales y su empresa podría empezar a triunfar. Este *modus operandi* podría durar meses o incluso años. Otra opción era, sin duda alguna, los medios de masas, a los que era complicado acceder para empresas no punteras en el sector.

En la actualidad todo ha cambiado. Y es que con que una empresa se haga conocida a través de una red social o la publiciten ciertas personas encargadas de hacerlo en redes sociales (comúnmente conocidos como “influencers”) ya pueden dar un salto de popularidad abismal, y ponerse al nivel de otras que llevan en el sector mucho más tiempo que ellas. El boca a boca funciona de la misma forma, la diferencia es que ahora todo el mundo dispone de un smartphone y esa publicidad te llega a través de un click, mucho más rápido, y, a veces, engañoso. (Grover, Kar, & Dwivedi, 2022) (Gong, Zhang, Zhao, & Jiang, 2017)

En este Trabajo de Final de Grado se analizará cuál es el potencial de las redes sociales en la cuota de mercado de las empresas, en este caso y más concretamente, en las empresas del sector agroalimentario en España. Se obtendrán métricas de redes sociales para analizar si tienen alguna relación con la cuota de mercado que la empresa tiene o si no tienen relación. Además, se utilizarán otras métricas derivadas de la popularidad online utilizando herramientas como Google Trends para observar si verdaderamente las empresas más populares en Internet son aquellas que tienen mayor cuota de mercado. Se utilizarán herramientas como matrices de correlación o distintos tipos de regresiones para intentar explicar las relaciones entre las variables de popularidad online y la cuota de mercado.

1.2 Motivación

Actualmente, la transformación digital ha cambiado enormemente los modelos de negocio en todo el mundo. El papel de las redes sociales y otros factores de la sociedad moderna ha adquirido una relevancia sin precedentes. Para cualquier empresa o individuo que aspire a destacar en el entorno competitivo actual, encontrar seguidores, influir en ellos y hacerse visible en línea se han convertido en aspectos cruciales. Desde la llegada de Internet a finales de los años 1970, y especialmente con el auge de las redes sociales, la dependencia de las personas de la información obtenida a través de estas plataformas ha crecido exponencialmente. (Cheung & Thadani, 2012)

El objetivo de este trabajo fin de grado (TFG) es analizar el impacto potencial de las redes sociales en la cuota de mercado de las empresas de la industria agroalimentaria española. Para ello, se obtendrán y analizarán métricas de varias plataformas de redes sociales para determinar si existe una correlación significativa entre estas métricas y la participación de mercado de la empresa en cuestión. Además, se utilizarán herramientas como Google Trends para evaluar si las empresas más populares en el entorno digital coinciden con aquellas con mayores cuotas de

mercado. Se utilizarán matrices de correlación y varios modelos de regresión para explorar y explicar la relación entre las variables de popularidad en línea y la participación de mercado.

Este análisis no sólo permite comprender el impacto de la presencia digital en el éxito comercial de las empresas agroalimentarias, sino que también proporciona una perspectiva cuantitativa sobre la importancia de las estrategias de marketing digital en un importante sector de la economía española, como es el sector agroalimentario. Muchas empresas podrán hacer útil esta información para poder trabajar a partir de las métricas de sus redes sociales y crecer con respecto a su competencia.

1.3 Objetivos

El objetivo principal de este TFG es analizar la relación entre la popularidad online y la cuota de mercado de las empresas del sector agroalimentario en España. A partir de ahí, encontrar cualquier tipo de patrón o de relación entre variables y cuota de mercado ocupada será útil para el estudio. Más en concreto, se plantean los siguientes objetivos específicos:

- Estudiar, interpretar y visualizar la relación entre una variable que cuantifica la popularidad online (Google Trends) y la cuota de mercado de las empresas del sector agroalimentario en España.
- Estudiar, interpretar y visualizar la relación entre variables que miden la popularidad en redes sociales: Twitter, Facebook o Instagram; y la cuota de mercado de las empresas del sector agroalimentario en España.
- Analizar para el contexto de una empresa concreta, si las subidas y bajadas de las métricas influyen en su cuota de mercado a lo largo del tiempo.
- Crear un modelo de regresión que prediga la cuota de mercado que ocupa una empresa del sector a partir de las métricas que arrojan las redes sociales.
- Utilizar el modelo de regresión para crear una especie de escenario ficticio en el que un usuario indica las métricas de su empresa y el modelo devuelve el valor de cuota de mercado que ocupará.

Para llegar a realizar todos esos objetivos se han tenido que realizar múltiples tareas secundarias, a las que se podría nombrar objetivos secundarios, los cuales han sido:

- Recopilar y extraer datos de actividad online de las empresas del sector agroalimentario en España (redes sociales, tráfico web y medios de comunicación).
- Modificar los datos con el objetivo de poder trabajar con ellos de la manera más cómoda y eficiente posible.
- Aplicar técnicas de análisis, desde métodos de regresión simples hasta modelos de regresión complejos, con el fin de que arrojen los mejores resultados posibles y alcanzar un nivel de confianza óptimo.

1.4 Estructura del TFG

Con la finalidad de tener la mayor claridad y el mejor formato posible este Trabajo de Fin de Grado se ha estructurado de la siguiente manera:

Se comenzará con una breve introducción, para conocer de qué va a tratar el Trabajo de Fin de Grado más a fondo. Tras la introducción, se hablará del marco teórico que involucra a todo el estudio. En el marco teórico se definirán los conceptos que son interesantes conocer a la hora de abordar el estudio y que los usuarios deben conocer para sacar la mayor información posible del trabajo. Entre estos conceptos se definirá el sector agroalimentario, las redes sociales

y sus principales componentes y las plataformas que se utilizarán a la hora de la extracción de datos como son Kantar WorldPanel y Google Trends. Para cada una de ellas se analizará su uso y su relación con el sector agroalimentario.

Una vez conocidos y entendidos los conceptos teóricos se pasará a la parte de metodología en donde se explicará cuáles son los datos y cuál ha sido la extracción y el tratamiento de los datos extraídos y desde qué lugar. Finalmente se mostrará cuál va a ser el aspecto de los datasets finales y cómo se ha llegado hasta ellos. Dentro de la metodología también se hablará de los métodos estadísticos empleados y que serán de utilidad en el grueso del estudio. En este punto se comprenderá, de una forma más teórica, cuáles son los métodos estadísticos utilizados en el estudio y el porqué de su uso.

El próximo punto que recalcar es el punto de los resultados y visualización. En este momento ya se conoce de forma teórica todos los conceptos y la forma en la que vamos a abordar los objetivos propuestos y solo queda una cosa: plasmarlo en visualizaciones y modelos. En este punto se observará cuáles modelos son óptimos para los objetivos propuestos y se tratará de responder a todos los objetivos principales del estudio, ya sea mediante visualizaciones gráficas, modelos estadísticos o mediante cualquier otra utilidad, pero siempre con el fin de responder a los objetivos principales. Por último, la parte final de este reporte consta de una conclusión en la que se sintetizan todos los resultados obtenidos y se resume el contenido de los resultados del trabajo.

2. MARCO TEÓRICO

2.1 Sector agroalimentario.

2.1.1. Definición y actividades incluidas en este ámbito. Relación con la economía.

El sector agroalimentario es una de las partes fundamentales de la economía de cualquier país o territorio y abarca todas las actividades relacionadas con la producción, transformación, distribución y comercialización de alimentos, desde el sector primario (ganadería, agricultura), hasta el transporte, almacenamiento, comercialización e investigación y desarrollo en este ámbito. Es un sector crucial para el desarrollo económico de una nación por todos estos motivos, entre los que se incluye, además, un papel importante en la generación de empleo, que además abarca todas las generaciones desde gente joven sin experiencia laboral hasta personas de avanzada edad, en cualquier lugar del país, de forma homogénea, aunque quizá distribuida de la forma en que las actividades primarias generan más empleo en zonas rurales y las actividades de distribución, almacenamiento o I+D en zonas más desarrolladas. Dentro de la importante función de este sector en la economía de los países, a esto hay que añadir la capacidad de exportación en la que muchos de los territorios basan gran parte de su economía. (Montoriol Garriga, 2019)

En concreto, se va a analizar el comportamiento del sector agroalimentario en territorio español, y como varían las cuotas de mercado dados diversos factores. Por ello y en adelante, se va a hablar centrándose en el marco español.

2.1.2. Datos relevantes observados. Análisis de las cifras relacionadas a la cantidad de empresas, volumen de ventas y otros datos de importancia.

(Ministerio de Agricultura, Pesca y Alimentación, Gobierno de España, 2023) Siguiendo el Informe Anual de la Industria Alimentaria Española durante el PERIODO 2022 – 2023 publicado por el Ministerio de Agricultura, Pesca y Alimentación del Gobierno de España se pueden obtener datos claves y relevantes acerca del sector agroalimentario.

En la UE, la industria alimentaria es la principal actividad manufacturera, con una impresionante facturación de más de 1.121 millones de euros, lo que representa el 14,3% del sector manufacturero. Está formado por 294.000 empresas y ofrece oportunidades de empleo a 4,62 millones de personas. En esta amplia red, las PYME desempeñan un papel fundamental: representan el 95,7% del total de empresas y aportan el 39,4% de la facturación de la industria alimentaria.

Por otro lado, la industria alimentaria española ocupa una posición destacada, ocupando el cuarto lugar en la UE en términos de facturación, después de Francia, Alemania e Italia. La facturación fue de 142.073 millones de euros, lo que representa el 24,2% de la industria manufacturera española. La industria no es sólo un pilar de la economía sino también un importante empleador, con el 22,6% de los trabajadores manufactureros empleados en la industria. Además, su contribución al PIB del país también es significativa, alcanzando el 2,4%.

España cuenta con 28.335 empresas del sector alimentario, de las cuales el 96,1% son pequeñas y medianas empresas de menos de 50 empleados y el 77,7% de menos de 10 empleados. En términos de empleo, la industria de alimentos, bebidas y tabaco emplea a 549.300 personas, lo que representa el 21,5% de la industria manufacturera y el 2,6% de la economía total. Un hecho digno de mención es que la tasa de empleo femenino en esta industria es del 37,2%, significativamente más alta que el 28,3% en otras industrias manufactureras. Esta industria dinámica no sólo impulsa la economía, sino que también proporciona importantes oportunidades de empleo y desarrollo para el país.

De aquí en adelante se pueden visualizar distintas tablas en los que se desagregan distintos aspectos del sector, a fecha máxima 2022.

Subsectores	Menos de 10		De 10 a 49		De 50 a 199		De 200 a 249		250 y más		Total	
	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%
Industria cárnica	2.085	9,5	916	17,6	166	21,0	16	27,1	61	24,5	3.244	11,4
Industria del pescado	250	1,1	223	4,3	54	6,8	6	10,2	17	6,8	550	1,9
Prep. y conservación frutas y hortalizas	831	3,8	349	6,7	103	13,0	5	8,5	30	12,0	1.318	4,7
Aceites y grasas	1.290	5,9	286	5,5	25	3,2	s	1,7	7	2,8	1.609	5,7
Productos lácteos	1.336	6,1	218	4,2	41	5,2	s	3,4	23	9,2	1.620	5,7
Molinería y almidones	267	1,2	80	1,5	8	1,0	s	3,4	6	2,4	363	1,3
Panadería y pastas alimenticias	9.215	41,8	1.560	30,0	120	15,2	s	6,8	25	10,0	10.924	38,6
Fabricación otros productos alimenticios	2.305	10,5	665	12,8	117	14,8	11	18,6	49	19,7	3.147	11,1
Productos de alimentación animal	440	2,0	214	4,1	52	6,6	s	6,8	6	2,4	716	2,5
Fabricación de bebidas	4.010	18,2	695	13,3	106	13,4	8	13,6	25	10,0	4.844	17,1
Total Industria Alimentaria	**22.029**	**100**	**5.206**	**100**	**792**	**100**	**59**	**100**	**249**	**100**	**28.335**	**100**

***Nota**: (S*) Al ser menor de 5 es secreto estadístico INE*

Ilustración 1. Directorio Central de Empresas DIRCE 2023. INE (14/12/2023).

ESPAÑA - TOTAL PAÍSES	EXPORTACIONES (M €)		IMPORTACIONES (M €)		SALDO (M€)		TASA (%)	
	2021	2022	2021	2022	2021	2022	2021	2022
Comercio alimentario total	57.135	64.451	38.381	50.382	18.755	14.069	149	128
Comercio alimentario transformado	**37.985**	**44.493**	**25.228**	**32.453**	**12.757**	**12.040**	**151**	**137**
Comercio alimentario no transformado	19.150	19.958	13.153	17.929	5.998	2.030	146	111

Ilustración 2. Exportaciones e Importaciones en territorio español (Dpto. de Aduanas e Impuestos Especiales, 2022)

El valor de las exportaciones, en el total de la industria de alimentación y bebidas, considerados por grupos de productos según TARIC, los más significativos son los correspondientes a Carne de porcino con 5.835 M€, seguido de Aceite de oliva 4.336 M€, Vino total 2.980 M€, Resto de aceites No oliva 1.863M€, Carne de bovino 1.212 M€, Preparaciones para alimentación animal 1.111 M€, Preparaciones alimenticias diversas 1.046 M€ y Otros productos cárnicos comestibles 913 M€. (Informe anual de Comercio exterior 2022. MAPA.)

ESPAÑA - TOTAL PAÍSES	EXPORTACIONES		
	(Millones €)	(Miles Tn)	
02.002 Carne de porcino fresca, refrigerada o congelada	5.835,0	2.084,1	1
15.012 Aceite de oliva	4.336,8	1.166,5	2
22.006 Vino total	2.980,2	2.116,7	3
15.013 Resto de Aceites	1.863,8	949,5	4
02.001 Carne de bovino fresca, refrigerada o congelada	1.212,3	222,8	5
23.010 Preparaciones para alimentación animal	1.111,2	871,0	6
21.012 Preparaciones alimenticias no expresadas ni comprendidas en otras partidas	1.046,5	220,9	7
02.010 Despojos comestibles	913,4	556,8	8
20.013 Aceitunas preparadas o conservadas (excepto en vinagre o en ácido acético)	800,4	482,1	9
16.2.005 Preparaciones y conservas de túnidos	784,5	131,6	10
17.013 Artículos de confitería	777,9	257,0	11
04.012 Quesos	709,0	117,2	12
03.015 Túnidos frescos, refrigerados o congelados	641,0	216,2	13
20.004 Tomates preparados o conservados (excepto en vinagre o en ácido acético)	558,0	541,1	14
16.1.003 Embutidos	548,2	79,6	15

Ilustración 3. 15 principales exportaciones en España. (Dpto. de Aduanas e Impuestos Especiales, según TARIC, 2022)

En cuanto al valor de las importaciones el valor de los productos más significativos según TARIC corresponden a Aceite no oliva o resto de aceites con 3.428 M€, seguido de Quesos 1.640 M€, Tortas de soja 1.437 M€, Café y sucedáneos 1.374 M€, Calamares y potas congelados 1.147 M€, Camarones, langostinos y quisquillas congelados 1.143 M€, Preparaciones alimenticias

diversas 1.022 M€ y Preparaciones para alimentación animal 923 M€ (Departamento de Aduanas e Impuestos Especiales. Según TARIC. Años 2021 y 2022)

ESPAÑA - TOTAL PAÍSES	IMPORTACIONES		
	(Millones €)	(Miles Tn)	
15.013 Resto de Aceites	3.428,2	2.324,2	1
04.012 Quesos	1.640,5	358,5	2
23.005 Tortas y demás resíduos sólidos de la extracción del aceite de soja	1.437,3	2.824,6	3
09.001 Café y sucedáneos	1.374,1	376,3	4
03.083 Calamares y potas congelados, secos, salados o en salmuera	1.147,9	256,6	5
03.070 Camarones, langostinos y quisquillas congelados y sin congelar	1.143,2	164,7	6
21.012 Preparaciones alimenticias no expresadas ni comprendidas en otras partidas	1.022,6	226,7	7
23.010 Preparaciones para alimentación animal	923,2	595,3	8
16.2.005 Preparaciones y conservas de túnidos	876,9	177,5	9
02.001 Carne de bovino fresca, refrigerada o congelada	847,3	113,0	10
15.012 Aceite de oliva	814,2	279,2	11
03.010 Salmones frescos, refrigerados o congelados	747,9	90,7	12
03.085 Pulpos congelados, secos salados o en salmuera	713,6	60,7	13
19.013 Otros productos de panadería, pastelería o galletería	512,7	167,7	14
03.015 Túnidos frescos, refrigerados o congelados	506,6	145,0	15

Ilustración 4. 15 principales importaciones en España. (Dpto. de Aduanas e Impuestos Especiales, según TARIC, 2022)

En la próxima figura se puede visualizar también desagregando por Comunidad Autónoma el número de empresas destinadas al sector agroalimentario dependiendo del número de empleados del que disponen.

Comunidad Autónoma	Menos de 10		De 10 a 49		De 50 a 199		De 200 a 249		250 y más		Total	
	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%
Andalucía	4.188	19,0	994	19,1	89	11,2	3	5,1	28	11,2	5.302	18,7
Aragón	744	3,4	174	3,3	36	4,5	1	1,7	8	3,2	963	3,4
Principado de Asturias	511	2,3	119	2,3	11	1,4	0	0,0	3	1,2	644	2,3
Illes Balears	477	2,2	108	2,1	7	0,9	0	0,0	0	0,0	592	2,1
Canarias	745	3,4	155	3,0	25	3,2	2	3,4	3	1,2	930	3,3
Cantabria	285	1,3	82	1,6	10	1,3	0	0,0	2	0,8	379	1,3
Castilla y León	2.298	10,4	483	9,3	67	8,5	3	5,1	16	6,4	2.867	10,1
Castilla-La Mancha	1.722	7,8	332	6,4	43	5,4	6	10,2	14	5,6	2.117	7,5
Cataluña	2.697	12,2	777	14,9	168	21,2	11	18,6	64	25,7	3.717	13,1
Comunitat Valenciana	1.786	8,1	462	8,9	72	9,1	6	10,2	34	13,7	2.360	8,3
Extremadura	1.061	4,8	156	3,0	27	3,4	0	0,0	2	0,8	1.246	4,4
Galicia	1.801	8,2	359	6,9	60	7,6	6	10,2	19	7,6	2.245	7,9
Comunidad de Madrid	1.093	5,0	330	6,3	57	7,2	1	1,7	17	6,8	1.498	5,3
Región de Murcia	711	3,2	227	4,4	44	5,6	6	10,2	15	6,0	1.003	3,5
Comunidad Foral Navarra	392	1,8	123	2,4	31	3,9	3	5,1	14	5,6	563	2,0
País Vasco	1.041	4,7	215	4,1	24	3,0	7	11,9	7	2,8	1.294	4,6
La Rioja	447	2,0	106	2,0	21	2,7	4	6,8	3	1,2	581	2,1
Ceuta y Melilla	30	0,1	4	0,1	0	0,0	0	0,0	0	0,0	34	0,1
Total Industria Alimentaria	**22.029**	**100**	**5.206**	**100**	**792**	**100**	**59**	**100**	**249**	**100**	**28.335**	**100**

Ilustración 5. Número de empresas por tamaño y CCAA. (DIRCE, 2023)

2.1.3. Evolución del sector

2.1.1.1 Impacto de la era digital

El sector agroalimentario ha experimentado una notable evolución a lo largo del tiempo, y en la era digital, este cambio ha sido aún más pronunciado debido al impacto de la tecnología y la digitalización en toda la sociedad mundial. Antiguamente, en siglos anteriores, el sector agroalimentario ha sido caracterizado por métodos de producción tradicionales y una cadena de suministro lineal. Sin embargo, con la llegada de la era digital, ha habido una gran transformación en la forma en que se cultivan, procesan, distribuyen y comercializan los alimentos.

Uno de los principales impactos de la era digital en el sector agroalimentario ha sido el surgimiento de tecnologías avanzadas que permiten a los agricultores monitorear y gestionar sus cultivos de manera más eficiente, optimizando recursos y mejorando la calidad de los cultivos.

Estos avances tecnológicos han llegado, en gran parte, debido a la digitalización de la sociedad, que hacía mucho más fácil esta posibilidad. También ha facilitado la integración de la cadena de suministro agroalimentaria, permitiendo una mayor coordinación y colaboración entre los diferentes actores involucrados, desde agricultores y procesadores hasta minoristas y consumidores. Esto ha dado lugar a una mayor transparencia y trazabilidad en la cadena de suministro. (Guo, y otros, 2023) (Xia, Baghaie, & Sajadi, 2024)

La era digital también ha abierto nuevas oportunidades en términos de comercialización y distribución de productos agroalimentarios gracias a plataformas de comercio electrónico, redes sociales y aplicaciones móviles han permitido a los agricultores y empresas agroalimentarias llegar directamente a los consumidores, eliminando a los intermediarios. (Kosior, 2018)

2.2 Redes sociales

2.2.1 Definición y evolución en el tiempo.

La evolución de las redes sociales ha sido un enorme fenómeno que ha transformado la manera en que se comunica la sociedad, interactuando y compartiendo información en línea con el resto de la sociedad. En un inicio simplemente la definición de "red social" se resumía en una conexión social con el resto, pero actualmente desempeñan un papel clave en la sociedad digital moderna. Las redes sociales no tienen por qué ser online, de hecho, parte del concepto de red social es que se define como un conjunto de personas y conexiones entre ellas. Fue con la popularización de internet cuando se empezó a extender los sitios de redes sociales.

Las primeras redes sociales como Six Degrees, permitían a los usuarios crear perfiles y conectarse con otros individuos. Sin embargo, a principios de los 2000, la llegada de plataformas como Friendster y MySpace hizo que las redes sociales comenzaran a capturar la imaginación del público en general, siendo las primeras prácticamente de nicho. Las nuevas plataformas ofrecían nuevas formas de expresión personal, permitiendo a los usuarios personalizar sus perfiles, compartir fotos, música, mensajes, conectarse con amigos…

Sin embargo, estos detalles son insignificantes en comparación con lo que estaba por llegar. En 2004 nació Facebook, que fue la que realmente marcó un punto de inflexión en la historia de las redes sociales. Facebook abrió sus puertas a cualquier persona con una dirección de correo electrónico. Con su diseño intuitivo, características innovadoras y enfoque en la privacidad y la seguridad, Facebook se convirtió rápidamente en la red social dominante, atrayendo a millones de usuarios en todo el mundo y estableciendo un nuevo estándar para la interacción online.

Desde entonces, las redes sociales han experimentado una enorme evolución de forma constante, impulsada por avances tecnológicos del momento, cambios en el comportamiento del usuario y la aparición de nuevas plataformas y aplicaciones. Tras Facebook, la siguiente fue la llegada de Twitter en 2006, donde se introdujo el concepto de microblogging, permitiendo a los usuarios compartir pensamientos y actualizaciones en tiempo real en mensajes cortos. Tras ella aparecieron muchas más, siendo importante mencionar a Instagram, lanzado en 2010, que se desmarcó de las anteriores y se centró en la fotografía y el contenido visual. Otras redes sociales que destacar podrían ser YouTube, principal plataforma de vídeos online; o LinkedIn, red social favorita centrada en el ámbito profesional. (Meire, Ballings, & Van den Poel, 2017)

En resumen, las redes sociales han evolucionado desde sus comienzos, donde se podían definir como simples plataformas de conexión social entre amigos, hasta convertirse en una importante fuerza en la sociedad digital moderna. Han transformado la forma en que nos comunicamos, compartimos información, consumimos contenido y nos conectamos con el mundo que nos rodea, y continúan desempeñando un papel fundamental en nuestra vida cotidiana y en la manera en que interactuamos con el mundo. (Ayedee & Kumar, 2020) (Pourkhani, Abdipour, & Baher, 2019)

2.2.2 Principales plataformas y su relación con el sector agroalimentario.

Dentro de las principales plataformas de redes sociales en la actualidad, podemos fijar una gran base con cuatro pilares. Facebook, como la red social favorita internacionalmente, donde se ofrecen múltiples capacidades desde conexión con tus amigos hasta disponibilidad para jugar a videojuegos o comprar ropa. Otra de las más importantes es Twitter (también llamada X), donde millones y millones de usuarios interactúan cada día entre ellos. YouTube, como se ha descrito anteriormente, es la plataforma líder en visualización de vídeos en línea. Por último, Instagram, red social favorita para compartir contenido visual. Analizando su relación con el sector agroalimentario, se va a centrar en las dos primeras, Facebook y Twitter, puesto

que Instagram o YouTube no tienen suficiente importancia a la hora de su relación con el sector agroalimentario, pese a que al igual que el resto de las redes sociales, tiene la ventaja de poder contactar directamente con el consumidor.

La red social Facebook se posiciona como una plataforma clave para las empresas del sector agroalimentario. Su versatilidad permite a estas empresas desde promocionar sus productos hasta interactuar directamente con sus potenciales clientes. A través de la publicación de contenido relevante las empresas pueden mantenerse en contacto con su audiencia y fomentar la lealtad de marca. Además, y como parte clave de esta red social, Facebook ofrece herramientas de segmentación de audiencia que permiten a las empresas dirigirse específicamente a consumidores interesados en alimentos orgánicos, productos locales o prácticas agrícolas sostenibles, entre otras, buscando así maximizar el impacto de sus campañas de marketing.

Twitter también se ha convertido en un canal importante para la difusión de noticias e información relevante en el sector agroalimentario (al igual que con en el resto de los sectores). Las empresas pueden utilizar Twitter para compartir actualizaciones breves y en tiempo real sobre eventos, tendencias y desarrollos en la industria. Además, Twitter es un espacio propicio para participar en conversaciones sobre temas relacionados con el sector, como la agricultura sostenible o la seguridad alimentaria, entre otras. Otra herramienta valiosa que nos proporciona Twitter es la capacidad para generar conversaciones en tiempo real, permitiendo a las empresas responder rápidamente a las preguntas y preocupaciones de los consumidores. Por último, gracias a las impresiones de los usuarios sobre un tema concreto, las empresas pueden permitirse reaccionar a tiempo real a cualquier inconveniente que surja, desde polémicas internas hasta productos defectuosos, etc. pudiendo así conocer las sensaciones sobre sus productos o sobre los de la competencia, del usuario general.

2.2.3 Influencia de las redes sociales en el comportamiento del consumidor.

La influencia de las redes sociales en el comportamiento del consumidor en el sector agroalimentario va más allá de la simple interacción; se extiende a la forma en que los consumidores perciben, evalúan y eligen productos y marcas en un mercado cada vez más competitivo y cambiante. A continuación se puede ver cómo las redes sociales impactan a la hora de tomar decisiones en los consumidores.

Para comenzar, las redes sociales se han convertido en una fuente de información y educación para los consumidores de cualquier ámbito, y en este caso para los consumidores en el sector agroalimentario. Hay multitud de información que se puede obtener, desde aprender sobre los beneficios nutricionales de ciertos alimentos hasta descubrir nuevas técnicas de cultivo o recetas innovadoras, los consumidores recurren a plataformas como Facebook, Twitter o YouTube para obtener información relevante y actualizada. Por esta razón, esta información puede influir directamente en sus decisiones de compra, ya que los consumidores buscan productos que se alineen con sus valores y preferencias. (Boyd & Ellison, 2008) (Kaplan & Haenlein, 2010)

Por otro lado, las redes sociales permiten una conexión directa entre los consumidores y las empresas, agricultores y otros usuarios del sector agroalimentario. Los consumidores pueden expresar sus opiniones, hacer preguntas y compartir experiencias, lo que fomenta la transparencia y la confianza entre las marcas y sus clientes. Esta interacción no solo fortalece la relación entre las marcas y los consumidores, sino que también brinda a las empresas la oportunidad de recibir retroalimentación en tiempo real y adaptarse a las necesidades y preferencias del mercado de manera más efectiva. (Smith, Fischer, & Yongjian, 2012)

Otro punto es que la presencia activa en redes sociales puede tener un impacto significativo en la percepción de una marca en el sector agroalimentario. Las empresas que mantienen una presencia sólida y positiva en las redes sociales pueden generar mayor interés, confianza y preferencia entre los consumidores. Además, el contenido compartido en redes sociales, como fotos de productos frescos, testimonios de clientes satisfechos y publicaciones sobre prácticas agrícolas sostenibles, puede influir en la percepción de calidad, autenticidad y compromiso con valores éticos y medioambientales de una marca.

En resumen, las redes sociales son una herramienta poderosa que moldea el comportamiento del consumidor en el sector agroalimentario al proporcionar información y educación, facilitar la interacción y participación entre consumidores y empresas, e influir en la percepción de marca. Las empresas que comprenden y aprovechan esta influencia pueden posicionarse de manera más efectiva en el mercado, establecer relaciones más sólidas con los consumidores y diferenciarse de la competencia. (Qualman, 2019)

2.3 Google Trends

2.3.1 Definición y funcionamiento.

(Cebrián & Domenech, 2023) Google Trends es una herramienta lanzada por Google que proporciona series temporales mostrando el volumen de búsqueda de ciertas palabras clave en el buscador de Google. La frecuencia de las propias series temporales puede variar de forma diaria o mensual, a gusto del consumidor, además de poderse delimitar el área geográfica para la que se deseen obtener los datos. Gracias al retroalimentarse con el buscador más popular online nos proporciona también datos en tiempo real, si es de nuestro interés.

Fue lanzada a mediados de 2006, y desde ese primer lanzamiento ha sido modificada en múltiples ocasiones buscando su mayor eficiencia, hasta una última modificación en septiembre de 2012.

2.3.2 Interfaz

La interfaz de Google Trends es muy intuitiva y puede ser utilizada fácilmente por cualquier individuo acostumbrado al uso de las nuevas tecnologías. La forma más popular de utilizar la herramienta es desde la interfaz de usuario que se encuentra en Google y que se muestra en la siguiente figura:

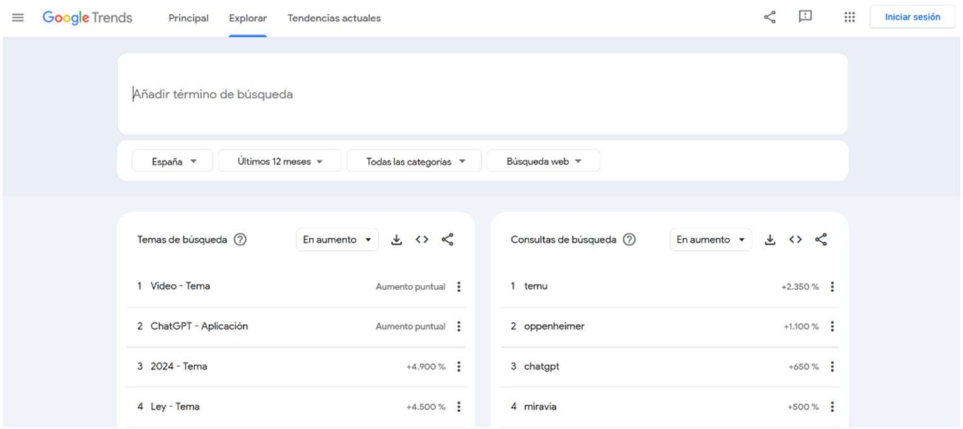


Ilustración 6. Interfaz Google Trends (Google Trends)

En la interfaz general se puede observar varios apartados como los temas de búsqueda o las consultas más populares del momento (en la parte inferior de la ilustración 6). Además, puede observarse en la parte inferior de la búsqueda, una serie de desplegables con filtros que puedes utilizar si es importante para las distintas aplicaciones. Los posibles filtros son, de izquierda a derecha:

- **Lugar**: puedes escoger tanto el país, como la región concreta en el que nos interesa ahondar.
- **Tiempo**: puedes escoger el rango de tiempo que interese para la búsqueda.
- **Categoría**: puedes escoger la categoría que más interese.
- **Plataforma**: permite diferenciar resultados de distintas plataformas.

Para conocer mejor el funcionamiento de la herramienta, se muestra la búsqueda de la palabra “agricultura” con los filtros de lugar: España; tiempo: últimos 5 años; todas las categorías y búsqueda web.

Google Trends muestra la evolución temporal de la búsqueda deseada como se observa en la siguiente figura:

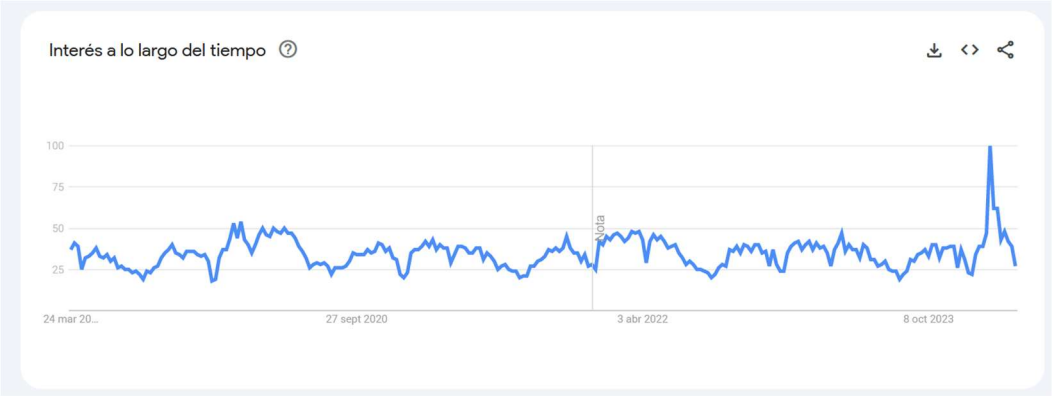


Ilustración 7: Evolución Temporal de la Búsqueda "agricultura" en Google Trends. (Google Trends)

El siguiente apartado es un desglose por subregión del lugar de la búsqueda.

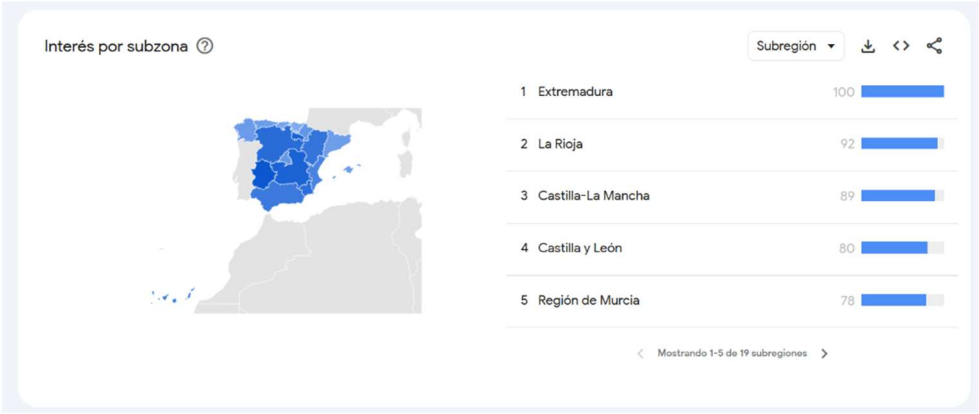


Ilustración 8. Interés por Subzona Geográfica de la Búsqueda "agricultura" en Google Trends. (Google Trends)

Y el último de los apartados muestra posibles temas o consultas relacionadas con el tema que se ha buscado.

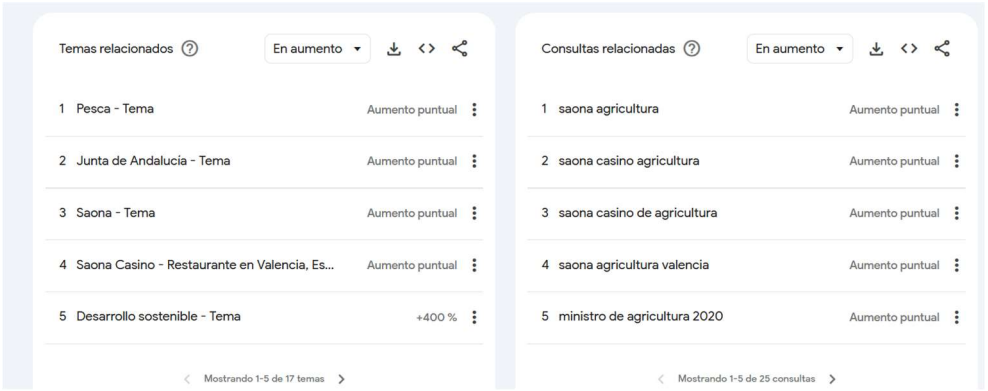


Ilustración 9: Temas y Consultas relacionadas tras la Búsqueda "agricultura" en Google Trends. (Google Trends)

Gracias a la búsqueda por popularidad que proporciona Google Trends, se pueden observar cambios a lo largo de tiempo o zona geográfica, que puede ser de utilidad para multitud de proyectos de distintos ámbitos.

2.3.3 Utilización de esta plataforma en el análisis de mercados

Google Trends brinda información valiosa sobre las tendencias de búsqueda en uno de los motores de búsqueda más importantes del mundo como Google. Esta herramienta es fundamental para posicionar a una empresa en Internet, ya que brinda datos confiables y objetivos que te ayudarán a tomar ciertas decisiones.

Esta herramienta ofrece una visión de demandas y ofertas para facilitar el posible negocio, además de proporcionar información sobre términos de búsqueda de potenciales clientes, consiguiendo así adaptar posibles estrategias de marketing. Estas estrategias pueden venir dadas por contenido relevante o palabras clave que se obtienen utilizando Google Trends. Además, y gracias a su posibilidad de filtrar por zona geográfica o intervalo temporal, la herramienta permite conocer posibles tendencias a lo largo de un período, y si estas tendencias siguen una estacionalidad. Gracias a todo esto, la utilización de Google Trends puede brindar a quien esté interesado una ventaja competitiva significativa, ya que podrás optimizar tu presencia en las redes y llegar a un público más amplio, adaptarte a las posibles tendencias y demandas del mercado en el que estés interesado, y tomar las mejores decisiones basadas en datos confiables, entre otras muchas utilidades. (Cebrián & Domenech, 2023)

2.3.4 Aplicaciones en el sector agroalimentario

Google Trends es una plataforma multiusos y es de gran utilidad en muchos de los sectores de la sociedad, pero para este caso se van a enumerar las posibles aplicaciones más interesantes que nos ofrece esta plataforma.

Tal y como se indica en su mismo nombre, Google Trends, nos permite identificar tendencias emergentes y actuales: Google Trends permite identificar términos de búsqueda relevantes en el ámbito agroalimentario y analizar su popularidad a lo largo del tiempo.

Además de nuevas tendencias, también ayuda al análisis de patrones estacionales. La estacionalidad desempeña un papel clave en el sector agroalimentario, debido a que muchos de las empresas pertenecientes al sector tienen productos con esta característica (depende mucho de la temporada, ya sea para su fabricación/cultivo como para su venta y distribución). Google Trends puede ayudar a identificar patrones estacionales en las búsquedas relacionadas con dichos productos específicos, ya sea además por zona geográfica o por período de tiempo concreto.

También para la evaluación de eventos externos debido a que Google Trends puede ayudar a evaluarlos en el sector agroalimentario, como en muchos otros. Como eventos externos se puede incluir seguridad alimentaria, demanda de productos no perecederos, o algo de actualidad, como los efectos de la pandemia. (Ghose, 2019) (Mavragani, Ochoa, & Tsagarakis, 2018)

2.4 Kantar World Panel

Kantar WorldPanel es una herramienta de investigación de mercado líder en el sector, que ofrece datos y análisis sobre el comportamiento del consumidor en el ámbito agroalimentario y otros sectores. Fundada en la década de 1960, Kantar WorldPanel recopila información detallada sobre las compras de los consumidores, sus preferencias y tendencias, proporcionando una visión integral del mercado. Esta visión es especialmente útil para el análisis del consumidor, beneficiando a cualquier tipo de empresa o asociación interesada en optimizar la eficiencia y rentabilidad de sus negocios.

La solidez de los datos y análisis proporcionados por Kantar WorldPanel convierte esta herramienta en un recurso muy útil para el estudio del comportamiento del consumidor, siendo relevante para una amplia variedad de sectores, desde grandes corporaciones multinacionales hasta pequeñas asociaciones locales. La capacidad de esta plataforma para desglosar los matices del mercado y prever las necesidades y deseos de los consumidores la posiciona como un activo estratégico crucial para cualquier entidad que busque maximizar la eficiencia y rentabilidad en sus operaciones comerciales.

En un entorno empresarial como el actual, altamente competitivo y en constante cambio, disponer de una visión integral del mercado y una comprensión profunda del comportamiento del consumidor es fundamental para el éxito y la sostenibilidad a largo plazo. Por lo tanto, una herramienta como Kantar WorldPanel se considera clave.

2.4.1 Definición y funcionamiento.

Kantar WorldPanel se destaca por su capacidad para recopilar datos detallados sobre las compras de una muestra representativa de hogares, abarcando tanto transacciones en tiendas físicas como en entornos online. A través de paneles de consumidores que son seleccionados, esta plataforma captura información clave sobre las marcas adquiridas, la cantidad de productos comprados, los precios pagados y una variedad de otros datos relevantes relacionados con el comportamiento de compra de los consumidores, en sea cual sea la necesidad que se plantee al interesado. Estos paneles de consumidores actúan como observadores activos del mercado, registrando cada interacción de compra para proporcionar una visión completa y actualizada del panorama comercial. Entre los distintos apartados que nos proporciona esta plataforma se tiene la elección de marcas, las preferencias de productos o las fluctuaciones en los precios, entre otras.

Como ya se ha comentado, los datos recopilados por Kantar son estratégicamente significativo y pueden ayudar a las empresas a conseguir una ventaja competitiva. Al analizar estos datos y generar informes detallados, la plataforma ofrece a las empresas una comprensión profunda del comportamiento del consumidor, lo que les permite identificar patrones, tendencias y oportunidades de mercado emergentes. Esta inteligencia de mercado permite a las empresas anticipar las necesidades y deseos de los consumidores, adaptar sus estrategias comerciales en consecuencia y tomar decisiones informadas y estratégicas que impulsen el crecimiento y la rentabilidad a largo plazo.

2.4.2 Aplicaciones en el sector agroalimentario

Tras conocer el funcionamiento y la ayuda que puede proporcionar Kantar WorldPanel a aquel que esté interesado, es el momento de ver cómo puede esta plataforma ser de utilidad en

el sector agroalimentario, que es el que compete en este caso. Para el caso a analizar en este TFG es de gran importancia analizar cómo se comporta el consumidor en aquello relacionado con este sector, para analizar cuáles son las estrategias que la herramienta puede proporcionar. Entre las diversas aplicaciones que se pueden encontrar, las cuatro posiblemente más importantes sean las siguientes que se van a enumerar y que se explicarán a continuación: Segmentación del mercado, evaluación de productos, análisis de precios y seguimiento de tendencias.

En el primer caso es la segmentación del mercado: Kantar WorldPanel permite identificar segmentos de consumidores basados en sus hábitos, preferencias o lugar de origen. Esto ayuda a las empresas agroalimentarias o del sector a adaptar sus productos y estrategias de marketing para satisfacer las necesidades específicas de cada segmento.

Evaluación de productos: Se puede también obtener datos detallados sobre el desempeño de los productos en el mercado, ya sea las ventas o su frecuencia de compra, entre otras opciones. Esto permite a las empresas del sector evaluar el rendimiento de sus productos y realizar ajustes según sea necesario.

De forma similar a la evolución de productos, Kantar WorldPanel también puede ayudar al análisis de precios: Kantar WorldPanel ofrece información sobre los precios de los productos y el impacto de las promociones en las ventas. Esto ayuda a las empresas a fijar precios competitivos y a diseñar promociones efectivas para aumentar la demanda de sus productos.

Por último, gracias a los datos proporcionados por la herramienta, que pueden estar segmentados, como ya se ha comentado, de forma demográfica y, por supuesto, temporal, las empresas agroalimentarias pueden utilizar esta ventaja para analizar tendencias, ya que permite a las empresas monitorear las tendencias del mercado agroalimentario. Esto ayuda a las empresas a anticipar cambios en la demanda del consumidor y a adaptar sus estrategias en consecuencia. (Green, 2015) (Ghose, 2019) (WorldPanel, s.f.)

3. METODOLOGÍA

3.1 Datos.

Para este estudio el objetivo principal es buscar la relación entre distintas variables que cuantifiquen la popularidad online con la cuota de mercado que poseen las distintas empresas del sector alimentario. El estudio se centrará en empresas del territorio español, por lo que todos los datos extraídos serán en consonancia. Para conseguir los datos que se emplearán próximamente se ha extraído información de varias fuentes distintas, en concreto 3 distintas, empleando distintas técnicas de extracción de datos como web scrapping o extracción de datos de una API Web. En los siguientes apartados se explica más a fondo cada una de las extracciones.

3.1.1 Cuotas de mercado. Kantar WorldPanel.

Kantar WorldPanel, cómo se ha explicado en apartados anteriores, es una herramienta que nos permite medir y entender las cuotas de mercado de diversos sectores. En este caso, y lo que es de interés de estudio son las cuotas de mercado de los supermercados españoles. Para ello es necesario entrar en la página web de Kantar WorldPanel y acudir a la página con gráficos de interés del estudio, para comprender los datos que se deben extraer para comenzar con el estudio. Como se puede visualizar en la siguiente figura, así es la interfaz sobre la que se quiere obtener los datos.

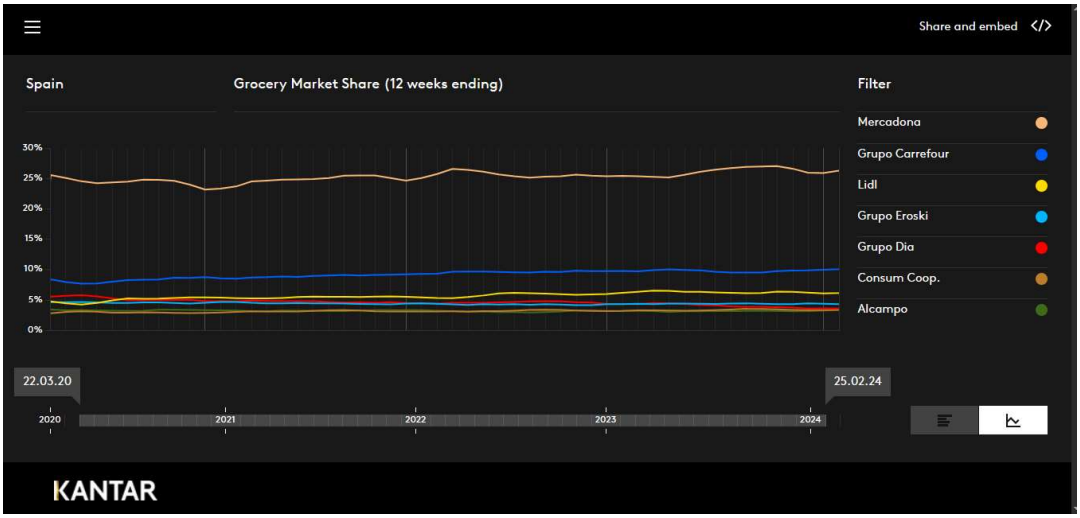


Ilustración 10. Interfaz de Kantar WorldPanel (Kantar WorldPanel)

En la ilustración 10, se puede observar el evolutivo de cuota de mercado de los diferentes supermercados españoles (Grocery Market Share), desde el año 2020 hasta la actualidad, como se puede observar en el eje horizontal. En la parte derecha de la pantalla se puede también observar los principales actores de este gráfico, entre los que se encuentran Mercadona, Grupo Carrefour… entre otros. Los datos de Kantar WorldPanel no están disponibles para descargar de forma abierta, pero inspeccionando la página y echando una mirada a su código HTML, se puede observar que todos los datos están alojados en un archivo JSON al que se puede acceder, en este caso, mediante Python.

3.1.1.1 Extracción y preparación de datos. Técnicas empleadas.

Tal y como se ha introducido en el apartado inmediatamente anterior, para lograr descargar todos los datos, tendrá que realizarse un script de Python en el que se consiga llamar

al JSON y descargar los datos que nos sean convenientes. En la siguiente tabla se enumeran las librerías y los paquetes utilizados en este script de Python.

Tabla 1. Librerías utilizadas extracción datos Kantar WorldPanel, fuente: Propia

Paquete	Librería	Utilidad
requests	requests	Permite realizar solicitudes HTTP en Python.
json	json	Trabajar con datos en formato JSON.
pandas	pandas	Herramientas para el manejo de los datos.
matplotlib	matplotlib	Visualizaciones gráficas.

Para realizar la descarga para el acceso a los datos del JSON se ha realizado un script de Python. Los pasos para seguir han sido los siguientes:

1. Se establece la URL de la que se desea descargar el contenido JSON de Kantar WorldPanel.
2. Se realiza una solicitud GET de la URL utilizando la función requests.get() de la librería requests para enviar una solicitud HTTP GET a la URL especificada y obtener el contenido de la respuesta, para verificar si la solicitud fue exitosa o no.
3. Si esta solicitud efectivamente es exitosa, se comienza con la extracción del contenido JSON mediante el método .json() para extraer el contenido JSON de la respuesta y convertirlo en un objeto Python, con el que sea más fácil trabajar posteriormente.
4. Se procesa el contenido JSON para extraer la información relevante. En este caso, el objeto de interés son las cuotas de mercado solamente de España. Por ello, se busca el país "Spain", obviando los resultados que ofrece Kantar WorldPanel para el resto de países y se accede a los datos de sus distintos periodos (desde 2020 hasta la actualidad). En cada uno de estos periodos se recopilan los datos de cada empresa en una lista.
5. Mediante la librería pandas, se procede a la creación de un DataFrame a partir de los datos recopilados. Cada fila del DataFrame representa una empresa en un periodo determinado, con columnas que incluyen el número de periodo, el año, el nombre de la empresa y su valor (la cuota de mercado correspondiente a ese periodo de ese año concreto)..
6. Por último se asigna ese DataFrame a una variable, para que posteriormente durante el estudio se pueda acceder de forma rápida y concisa a dicha información.

Todo ello se encuentra bajo un marco de manejo de errores, mediante el cual en caso de que algún error se produzca, desde una URL incorrecta, una solicitud GET no exitosa, o cualquier otro problema, el script de Python nos mostrará por pantalla que ha encontrado un error y nos mostrará el código de dicho error.

Una vez obtenidos los datos y almacenados en el DataFrame deseado ya se puede comenzar a modificar los datos a gusto del estudio. Con los datos originales se pueden realizar ciertos gráficos visuales (mediante la librería matplotlib), pero ahora van a comentarse ciertos cambios que se realizan sobre los datos originales con el fin de realizar el estudio final y llegar a los objetivos propuestos.

Primero de todo, se crea una columna “Código” que incluya un valor sin faltantes ni valores duplicados, que sea el valor que una esta tabla con el resto de tablas posteriores. Este código unitario tendrá el formato “Nombre Empresa”+”Mes”+”Año”, resultando un valor único que será de utilidad para unir (mediante la función merge de la librería pandas) distintos DataFrames a este ya creado.

El DataFrame final, con los datos ya correctamente descargados y las variables necesarias creadas tendría un formato similar al siguiente:

Periodo	Año	Nombre Empresa	Valor Empresa	Codigo
1 mm	yyyy	string	float	string
2 mm	yyyy	string	float	string

Siendo, “Periodo” el mes, “Año” el año de estudio, “Nombre Empresa” el nombre del supermercado correspondiente, “Valor Empresa” será la cuota de mercado de esa empresa durante ese periodo de ese año, y por último, “Codigo” será el código unitario y no repetido que se encontrará en los DataFrame que se busque juntar.

3.1.2 Google Trends.

El caso de Google Trends es similar a la descarga de datos de Kantar WorldPanel, en el sentido de que desde la interfaz de Google Trends, como se ha mostrado en apartados anteriores, es muy intuitivo visualizar todo aquello que sea necesario a la hora de ver o analizar el comportamiento de una consulta (un supermercado en este caso); pero tampoco se pueden descargar los datos de manera inmediata y desde la propia interfaz, se debe utilizar herramientas externas para conseguir esos valores que se necesiten en cada momento. En este caso, se ha utilizado una librería de Python, Pytrends, que se explicará al detalle en las próximas líneas.

3.1.2.1 Extracción y preparación de datos. Técnicas empleadas.

- **Pytrends**:

Esta librería Python es una API no oficial de Google Trends. Nos permite extraer datos de Google Trends de todo tipo relacionados con el interés del usuario por un tema o una consulta específica. Lo más útil que tiene es que, al igual que en la propia herramienta de Google desde su interfaz, mediante Pytrends también podemos extraer información a nivel mundial o geolocalizada, en tiempos concretos e incluso consultas relacionadas, además de otras funcionalidades que no se utilizarán en el contexto actual.

En la siguiente tabla se muestran las librerías utilizadas para la extracción y preparación de los datos de Google Trends:

Tabla 2. Librerías utilizadas extracción datos Google Trends, fuente: Propia

Paquete	Librería	Utilidad
TrendReq	pytrends.request	Proporciona una interfaz para interactuar con la API de Google Trends. Permite acceder a datos de tendencias de búsqueda de Google
pandas	pandas	Herramientas para el manejo de los datos.
matplotlib	matplotlib	Visualizaciones gráficas.

Una vez conocidas las librerías que se van a utilizar para extraer y descargar los datos de Google Trends, es el momento de hablar de la estructura de este sencillo script de Python, desde el que se descargan estos datos. En esta parte se verá la estructura concreta que se ha utilizado para la extracción.

Tras la importación de las librerías que se van a utilizar, el script comienza conectando nuestro script a la API de Google Trends. De esta forma, se crea una instancia de la clase

TrendReq. A esta instancia hay que establecer el lenguaje de las consultas: en nuestro caso, al parámetro llamado “hl” (host language - o en español idioma anfitrión) se le establece el valor "es", para indicar que el idioma de las consultas será español.

Una vez creada la instancia TrendReq, se define la lista de palabras clave que se quiere utilizar para consultar sus tendencias en Google Trends. La lista máxima es de 5 términos, por lo que se ha decidido que los supermercados con mayor cuota de mercado durante los 4 años del estudio sean aquellos que se busquen y se descarguen los datos de Google Trends. Esta lista de términos serán los siguientes: "mercadona", "carrefour", "lidl", "eroski", "dia supermercado". Una vez se consulte y se extraigan los datos, el valor que aparecerá asignado cada fecha será el de dicho supermercado en relación con el resto de supermercados, es decir, saldrá la cantidad (sobre 100) que se ha buscado el término “mercadona” en comparación al resto de términos durante ese periodo concreto. A partir de varias métricas (búsquedas del término en concreto, tiempo de permanencia durante la búsqueda o click en páginas web relacionadas con el término de búsqueda, entre otras) se define ese valor que a lo largo del trabajo se asumirá como la popularidad asociada a Google Trends.

El siguiente paso es construir la solicitud. Para ello se emplea el método build_payload, también de la librería pytrends, con tal fin. A este método se le pasa como argumento los siguientes parámetros: la lista de palabras que queremos buscar y encontrar sus tendencias, definida anteriormente; se especifica cat=0 para no tener en cuenta ninguna categoría específica. Otro de los argumentos es timeframe, en donde se indica el periodo de tiempo para el que se solicitan los datos. En este caso concreto el periodo que se solicita es desde marzo de 2020 hasta abril de 2024, el intervalo de fecha que se encuentra en la plataforma Kantar WorldPanel. Por último, se establece como ubicación geográfica “ES” haciendo referencia a que se quiere obtener los datos de las búsquedas en territorio español.

El siguiente y último paso de la extracción de datos es ejecutar la consulta de tendencias. Para ello se utiliza en este caso el método interest_over_time, también de la librería pytrends, para ejecutarla y obtener los datos de tendencia a lo largo del tiempo estipulado en la construcción de la solicitud.

A partir de aquí solo queda aproximar el formato de los datos al formato de DataFrame que buscamos. Para ello necesitamos que los datos estén calculados por mes y año, como en el DataFrame extraído de Kantar WorldPanel. Para ello se hace uso de la función resample, junto al cálculo de su media. Utilizando estas dos funciones se procesan los datos para agruparlos por mes y calcular la media de dicho mes concreto. También eliminamos la columna “isPartial” que no será de utilidad para nuestro estudio. (Fahrudin, Asniar, & Faizul Ula, 2022)

Lo siguiente que se realiza es una modificación del DataFrame original que se acaba de extraer para asemejarse al DataFrame extraído de Kantar WorldPanel. Para ello, queremos que nuestro DataFrame tenga 3 columnas - fecha, nombre de la empresa y la puntuación asignada por Google Trends a esa fecha. Utilizando la librería pandas construimos el nuevo DataFrame al que vamos introduciendo los datos del DataFrame sin modificar, con el objetivo de dejarlo al mismo formato. Una vez obtenido el DataFrame como se desea, se debe crear la columna código, que, al igual que anteriormente, será la columna con la que se junten los DataFrames para el estudio final. Una vez creada la columna, y asegurando que se tienen los mismos valores en la columna código de este DataFrame que los valores en la columna código en el resto de DataFrames se puede dar por concluida la extracción y la preparación de datos de Google Trends.

3.1.3 Redes sociales.

3.1.3.1 Extracción y preparación de datos. Técnicas empleadas.

Para la extracción de datos acerca de las métricas que pudiesen ayudar a contabilizar la popularidad online en redes sociales de las distintas empresas del sector agroalimentario, se ha utilizado la técnica de web scrapping, desde una página web que recoge las métricas de las distintas cuentas en redes sociales (tanto de Twitter, Youtube, Facebook…). La web en concreto es www.socialblade.com.

Como se ha comentado en el marco teórico, las redes sociales son una fuente de información y de negocio muy importante en la actualidad, y para cualquier empresa que quiera alcanzar una popularidad alta debe ser consciente de que tiene que tener un equipo de redes sociales o similar muy preparado. En el caso que compete este estudio se han analizado las métricas de 3 redes sociales (Twitter, Facebook e Instagram) de las cinco principales empresas del sector en España: Mercadona, Carrefour, Lidl, Eroski y Día. En concreto se han extraído para cada una de las empresas las siguientes métricas:

En **Twitter**:

- *TWEETS*: Número de tweets posteados por la empresa en el período concreto.
- *VARIACION SEGUID. TWITTER:* Variación de seguidores en la cuenta en un período concreto.

En **Facebook**:

- *LIKES FB:* Número de "Me gustas" recibidos por la página de la empresa en un período concreto.
- *TALK FB:* Cantidad de interacciones o de veces que se ha hablado de la empresa en un período concreto.

En **Instagram**:

- *NUEVOS SEG IG*: Variación de seguidores en la cuenta en un período concreto.
- *CONTENIDO IG*: Cantidad de posts que la cuenta ha publicado en un período concreto.

Para obtener todas las métricas que se han mencionado anteriormente, se ha realizado un script de Python, en el que, mediante la biblioteca BeautifulSoup de Python, se ha llevado a cabo una tarea de web scrapping sobre los datos que se buscaban. Estos datos obtenidos desde la script de Python se exportan a un archivo CSV (Comma Separated Values), en donde se almacenan todos los datos. Este CSV será finalmente importado desde la script de Python donde se realiza la extracción de todos los datos que se necesitan y a partir de ahí ya tendremos el DataFrame que se necesita.

Ilustración 11. Esquema Extracción Datos Redes Sociales, (Fuente: Propia)

Ahora que ya se entiende el proceso de extracción de los datos, se está en la situación de modificarlos al gusto que se desee para poder luego operar en los distintos objetivos que se buscan. Ahora mismo se tiene 6 columnas, una por cada métrica que se han mencionado anteriormente, y cada una de las filas o instancias serán las métricas para un mes concreto una empresa concreta.

Como en los casos anteriores, se tiene el período entre marzo de 2020 y marzo de 2024, por lo que en total se tendrán 245 instancias (filas) por 6 columnas de variables. Ahora se necesita, igual que en los casos anteriores, crear una columna "Código" en la que estén los valores unitarios con el que se puedan juntar estas variables al resto de variables que nos interesen. La columna "Código" igual que anteriormente tendrá el formato de *NombreEmpresaMesAño* (por ejemplo: Mercadona032020). Una vez se tiene los datos extraídos y la columna "Código" correctamente creada, se puede dar por concluida la extracción de métricas de redes sociales.

3.2 Unificación de datos y DataFrame definitivo.

Tras la extracción de datos desde las 3 distintas fuentes que se han extraído (Kantar WorldPanel, Google Trends y SocialBlade) se debe ahora crear DataFrames conjuntos para poder estudiar la relación entre las cuotas de mercado de las empresas obtenidas en Kantar WorldPanel con el resto de variables de popularidad online.

Para ello y como se ha indicado en cada uno de los apartados del 3.2. se ha creado una columna "Código" en cada uno de los DataFrames resultantes de la extracción de datos de cada fuente. Esta columna es la que se utilizará para unificar las variables de cada DataFrame y poder realizar los estudios posteriores.

Utilizando la librería pandas de Python, y gracias a la función "merge" es fácil unir los DataFrames necesarios. Para ello se le pasa a la función como parámetros los DataFrames que se quiere unir, junto a otros dos parámetros, el parámetro "on" en el que se referirá a la columna que hará de nexo de unión entre ambos DataFrames (en este caso será la columna "Código"); y el parámetro "how" en el que se debe especificar cómo se quiere que sea la unión (Left, Right, Outer, Inner, Cross). Left utiliza solo las claves del primer DataFrame, Right al contrario, las del segundo; outer utiliza la unión de las claves de ambos DataFrames, inner utiliza la intersección de las claves de ambos DataFrames y, por último, cross que crea el producto cartesiano de ambos DataFrames. En el caso de estudio, y sabiendo cómo están formadas las distintas tablas se realizará un merge con inner. A continuación se muestra el código del merge de los DataFrames de Kantar y de los datos de redes sociales.

Código: df_total = pd.merge(df_final_kantar, df_sm, on="Codigo", how="inner")

	Periodo	Año	Nombre Empresa	Valor Empresa	Codigo	TWEETS	VARIACION SEGUID. TWITTER	LIKES FB	TALK FB	NUEVOS SEG IG	CONTENIDO IG
0	03	2020	Mercadona	25.55	Mercadona032020	2516	11307	677490.0	17263.0	40465	12
1	03	2020	Grupo Carrefour	8.38	Grupo Carrefour032020	206	3022	NaN	NaN	15362	78
2	03	2020	Lidl	4.76	Lidl032020	112	2453	2241378.0	27635.0	30646	115
3	03	2020	Grupo Eroski	4.64	Grupo Eroski032020	364	685	NaN	NaN	1465	31
4	03	2020	Grupo Dia	5.55	Grupo Dia032020	453	1158	1166379.0	9503.0	3257	23
...	...	...	...	...	...	...	...	...	...	...	...
240	03	2024	Mercadona	26.80	Mercadona032024	724	19	807208.0	15934.0	4825	5
241	03	2024	Grupo Carrefour	10.41	Grupo Carrefour032024	141	-571	NaN	NaN	6128	45
242	03	2024	Lidl	6.36	Lidl032024	214	-68	2472617.0	32714.0	3957	12
243	03	2024	Grupo Eroski	4.13	Grupo Eroski032024	186	-56	290359.0	4496.0	1574	20
244	03	2024	Grupo Dia	3.68	Grupo Dia032024	236	85	1380359.0	12376.0	2309	29

245 rows × 11 columns

Ilustración 12. DF resultante de la unión de datos de Kantar WorldPanel y de Redes Sociales. En verde: datos de métricas de redes sociales; en amarillo: datos de KWP. (Fuente: Propia)

3.3 Métodos estadísticos empleados.

Para resolver los problemas de regresión que se plantean en el trabajo se va a realizar diversos métodos para lograr resolverlos de la mejor manera posible.

Antes de ello se empleará el algoritmo de perturbación aleatoria para obtener instancias sintéticas a partir de las instancias iniciales, para lograr obtener un mayor número de filas con las que poder realizar la tarea de regresión posterior.

Tras esta generación de datos sintéticos, se emplearán distintas técnicas de regresión basándose en distintos algoritmos desde el algoritmo Random Forest, hasta técnicas más complejas de regresión como Support Vector Regression (Regresión basada en vectores soporte) o Partial Least Squares Regression (Regresión de Mínimos Cuadrados Parciales). En los próximos apartados se explicará detalladamente el funcionamiento y el uso de cada una de ellas y se dirimirá cuál es la técnica utilizada finalmente.

3.3.1 Generación de nuevas instancias.

Previo a comenzar con los distintos análisis en profundidad hay que tener en cuenta un par de consideraciones previas:

- Se han eliminado las observaciones con algún dato nulo, para que no influyeran en las predicciones. Los valores nulos suponían el desconocimiento del dato concreto de la métrica de cualquier tipo de red social con respecto a esa empresa concreta. En lugar de introducir las instancias al estudio con valores aleatorios, se ha optado por eliminar dichas instancias. La eliminación de estas instancias hará que tengamos un número menor de observaciones para conseguir la regresión, pero, como se observará posteriormente, se aplicará un método para conseguir ampliar el número de observaciones para hacer el estudio más fiable.
- Para la realización del estudio en concreto, se han eliminado las columnas con algún identificador no relevante como el nombre de la empresa, el período o el año, quedándose así con los valores que se utilizarán para la regresión.

Tras la eliminación de las filas con algún valor nulo, queda un DataFrame con 185 instancias únicamente. Estas son pocas instancias para poder realizar un correcto tratamiento de los datos para la posterior regresión. Por esta razón es útil realizar una ampliación de las instancias de forma sintética, generando nuevos datos de forma aleatoria, pero siguiendo la estructura de los datos reales. Gracias a ello, se dispondrá de una mayor cantidad de observaciones y en la posterior regresión se ajustará mejor a lo que se busca. Para conseguir estos datos sintéticos, se utilizará perturbación aleatoria, un método que, como su propio nombre indica, genera datos aleatorios a partir de una pequeña perturbación o cambio de los datos reales. A continuación se define más a fondo cómo utilizamos esta herramienta:

Las perturbaciones aleatorias se pueden definir como cambios en un sistema dados factores incontrolables e impredecibles. Estas perturbaciones se caracterizan por seguir distribuciones de probabilidad concretas, tales como la distribución normal, entre otras. En matemáticas y otras ciencias, estas perturbaciones se representan como variables aleatorias agregadas a un modelo. Una ecuación de un modelo f(x), se le añade perturbación aleatoria de la siguiente manera $f(x)+\epsilon$, siendo ϵ la variable de ruido añadida a la observación real. (Goodfellow, y otros, 2014) (Shorten & Khoshgoftaar, 2019)

3.3.2 Random Forest

El Bosque Aleatorio, también conocido como Random Forest en inglés, es un algoritmo de aprendizaje automático ampliamente utilizado y reconocido por su eficacia y estabilidad en una variedad de aplicaciones. Fue introducido por Leo Breiman y Adele Cutler, y se ha convertido en una herramienta esencial para los científicos de datos debido a sus numerosas ventajas. Pese a que los árboles de decisión (o decision trees en inglés) pueden ser propensos a problemas como el sobreajuste, cuando se combinan en un Bosque Aleatorio, producen resultados más precisos y confiables.

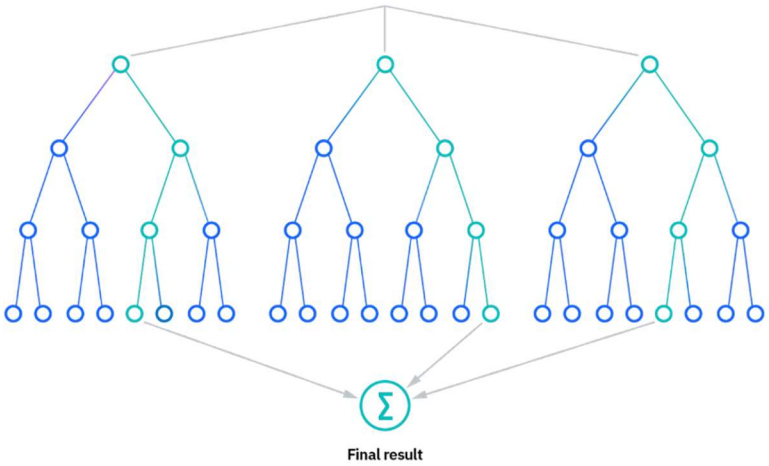


Ilustración 13. Esquema Funcionamiento Random Forest. (IBM)

El Bosque Aleatorio se basa en árboles de decisión, un algoritmo común en el aprendizaje supervisado. Estos árboles comienzan con una pregunta inicial y realizan una serie de preguntas adicionales hasta llegar a una conclusión. Cada pregunta se utiliza para dividir los datos en subconjuntos más pequeños, y las decisiones finales se toman en las hojas del árbol. Dicho de otra forma, cada pregunta hace que el árbol se ramifique hasta que llegue a una situación en la que deje de hacerlo, lo que llamamos hojas, en donde se tomarán las decisiones pertinentes.

Random Forest es un ejemplo de un método de conjunto, que como ya se ha comentado, combina los resultados de varios clasificadores para obtener una predicción final lo más ajustada posible. Este enfoque ayuda a reducir la varianza y mejorar la precisión del modelo.

Ya se ha explicado qué es el algoritmo de Random Forest, en los próximos apartados, se analizará más a fondo el funcionamiento de este algoritmo, y se verá la forma de utilizarse ya sea para tareas de clasificación o de regresión, como es este caso.

Random Forest utiliza la técnica de bagging para crear múltiples árboles de decisión no correlacionados entre sí. Esto se logra seleccionando aleatoriamente un subconjunto de características para cada árbol, asegurando la diversidad y solidez del modelo. En el anexo 4, se explica más a fondo cómo funciona la técnica de bagging.

Los bosques aleatorios pueden estar compuestos por cientos de árboles de decisión y la cantidad de árboles generalmente se ajusta durante el entrenamiento. Cada árbol se entrena con un subconjunto aleatorio de datos y genera una predicción, que se combina con las predicciones de otros árboles mediante votación o promedio. (Zou & Schonlau, 2020)

Este algoritmo tiene varias ventajas ventajas clave:

- Reduce el riesgo de sobreajuste (overfitting) del modelo: El algoritmo de Bosque Aleatorio reduce este riesgo al promediar múltiples árboles los cuales no están correlacionados.

- Flexibilidad: Puede manejar una gran cantidad y variedad de tareas de regresión y clasificación con buena precisión.
- Determinación de la importancia de las características: El algoritmo permite la evaluación de la importancia de las características en el modelo.

Como todo algoritmo también tiene inconvenientes, entre los que se encuentran la lentitud del proceso; dado que el entrenamiento de múltiples árboles puede ser un proceso muy lento, especialmente con grandes conjuntos de datos. De aquí también se traduce en la necesidad de disponer de más recursos de computación para poder procesar grandes conjuntos de datos. Además, puede ser más difícil de interpretar que un solo árbol de decisión.

Como ya se ha comentado, Random Forest se utiliza en muchos campos, desde finanzas hasta comercio electrónico, pasando por otros campos tan distintos como la atención sanitaria. Es un algoritmo muy utilizado ya que suele traducirse en modelos muy robustos y confiables. (Yoon, 2021)

3.3.3 Métodos de regresión.

Los modelos de regresión que se han utilizado para valorar cuál es la mejor opción son los siguientes:

1. Regresión lineal

La regresión lineal es una técnica fundamental muy utilizada en el análisis estadístico y el aprendizaje automático que se usa para modelar la relación entre una variable dependiente y una o varias variables independientes. Su simplicidad y versatilidad lo convierten en una herramienta muy utilizada en varias disciplinas para predecir valores futuros, comprender las relaciones entre variables y tomar decisiones basadas en datos. La ecuación que define cómo funciona la regresión lineal es algo como esto:

$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\varepsilon$, siendo Y la variable a predecir, X1, X2… las variables dependientes o predictoras, B0, B1… coeficientes asociados a esas variables predictoras, y ε la diferencia entre el valor real y el valor predicho.

Como se ha comentado, este método es un método muy sencillo y simple a la hora de ser implementado por lo que se utiliza en varios sectores del conocimiento. Desde la economía modelando relaciones entre variables económicas hasta la Ciencia de la Salud, en donde se pueden prever de forma mucho más eficiente y menos invasiva para el paciente utilizando modelos que utilizan este método entre muchos otros. Otras de las aplicaciones de esta regresión lineal podrían ser la ingeniería, las ciencias sociales, la informática… (Montgomery, Peck, & Vining, 2012)

2. Support Vector Regression (SVR)

Otro tipo de regresión utilizada es la regresión mediante vectores soporte o Support Vector Regression (SVR). SVR es una técnica de aprendizaje supervisado que se utiliza para realizar distintas tareas de regresión. Al igual que su homólogo Support Vector Machine (SVM) para la clasificación, SVR utiliza vectores de soporte en un espacio de alta dimensión para aproximar la función de regresión. SVR busca encontrar una función de regresión que se ajuste a los datos de entrenamiento lo mejor posible, y que, además, al mismo tiempo tenga un margen de holgura o tolerancia en relación con un margen de error predefinido.

Una de las utilidades más extendidas de SVR es el análisis de series temporales y predicción de valores a futuro.

En cuanto a similitudes y diferencias con la regresión lineal se puede observar que la principal similitud es que ambas regresiones tienen el mismo fin: relacionar las variables regresoras y las variables predictoras, buscando minimizar el error al máximo posible. Pero esta similitud es algo común a la hora de analizar los distintos tipos de regresión.

Diferencias se pueden enumerar varias pero las dos más importantes son: la regresión lineal asume que las variables tienen una relación lineal, mientras que en la regresión con vectores soporte se pueden modelar también relaciones no lineales, haciendo uso de funciones de mapeo no lineales. La otra principal diferencia es que SVR es mucho menos sensible a valores atípicos en comparación con la regresión lineal, lo cual puede ser de gran utilidad en diversas ocasiones. (Dash, Nguyen, & Cengiz, 2023) (Zhang & O'Donnell, 2020)

3. Partial Least Squares Regression (PLSR)

La regresión de mínimos cuadrados parciales (PLSR) es una técnica estadística que combina características del análisis de componentes principales (PCA) y la regresión múltiple. Es particularmente útil cuando las variables independientes son altamente colineales o cuando el número de variables independientes excede el número de observaciones. La regresión de mínimos cuadrados parciales (PLSR) tiene como objetivo proyectar la variable independiente X y la variable dependiente Y en un nuevo espacio de componentes latentes. Estos componentes latentes se eligen de tal manera que se busca maximizar la covarianza entre X e Y.

Al igual que las anteriores tipos de regresión, la PLSR tiene multitud de aplicaciones que pueden ir desde la economía o la ciencia de la salud hasta el estudio de químicos o la informática.

Las principales diferencias de este tipo de regresión frente a las mencionadas anteriormente podrían ser:

- En cuanto a la linealidad, PLSR proyecta los datos en un espacio de componentes latentes y, pese a que sigue siendo una técnica lineal, puede llegar a capturar complejas relaciones entre datos. Por otra parte, SVR sí puede hacer relaciones no lineales.
- PLSR está diseñado para trabajar con la colinealidad, mientras que la regresión lineal tiene problemas con ella y la SVR es menos afectada.
- Por otro lado, PLSR es relativamente sencillo de interpretar una vez se seleccionan el número de componentes latentes. La regresión lineal es muy sencilla mientras que la SVR es una técnica relativamente compleja.
- Por último, en relación a los valores atípicos, SVR es la técnica que mejor los mitiga, pese a que PLSR los trabaja de una forma distinta a la regresión lineal y afectan menos que en ese método.

(Kamboj, Paramita, & Mishra, 2022)

3.3.4 Métricas de Evaluación.

Todo modelo necesita ser evaluado para finalmente poder ser lanzado. En este caso, para el modelo de regresión, se van a tener en cuenta dos métricas a la hora de evaluar el modelo, el

MSE (Mean Squared Error) y el R^2 (coeficiente de determinación). Estas métricas serán las que se tendrán en cuenta a la hora de la elección de un modelo por encima de otro.

- **MSE**:

El Mean Squared Error (MSE) es una métrica que evalúa la precisión de los distintos modelos de regresión. Su objetivo es medir la magnitud promedio de los errores cuadráticos entre los valores reales y los predichos. Es una métrica muy utilizada dada su simplicidad y su fácil interpretación. Por tanto, el MSE proporciona una medida de la precisión de un modelo de regresión. Su interpretación es la siguiente: un MSE bajo indica un modelo más preciso, mientras que un MSE más alto sugiere un modelo con mayor error en general. El MSE penaliza los errores grandes de mayor forma dada la operación de cuadrado. Este hecho hace que esta métrica sea sensible a valores atípicos.

También cabe indicar que el MSE se mide en las mismas unidades que la variable a predecir, por lo que hay que tener en cuenta la unidad en la que se tiene la variable a predecir para interpretar el modelo de forma correcta.

- **R^2:**

Por otra parte, el coeficiente de determinación es una medida que indica la proporción de la variabilidad en la variable dependiente (la variable a predecir) que es explicada por el modelo de regresión en concreto. Está métrica es muy importante para evaluar la bondad de ajuste del modelo.

Su interpretación también es igual de sencilla que la del MSE. El coeficiente de determinación varía entre 0 y 1, siendo 1 indicador de que el modelo explica perfectamente la variabilidad de los datos, mientras que un coeficiente de 0 indica que el modelo no explica ninguna variabilidad. En resumen, valores cercanos a 1 indican un buen ajuste del modelo, mientras que valores cercanos a 0 sugieren que el modelo no es adecuado para explicar los datos.

Explicado de otra forma, supongamos que el coeficiente de determinación es de 0,9. Este coeficiente indica que aproximadamente el 90% de la variabilidad en la variable dependiente (en el caso de estudio la cuota de mercado) puede ser explicada por el modelo, lo que se traduce por un buen ajuste del modelo a los datos.

Cabe tener presente a la hora de la evaluar y seleccionar los parámetros para el modelo de regresión que también hay que tener en cuenta el tiempo de ejecución de cada uno de los casos, ya que modelos que expliquen una gran variabilidad pero que tardan en ejecutarse mucho más tiempo que otros modelos que quizá no expliquen ese porcentaje de variabilidad pero se quedan cerca, serán menos óptimos a la hora de afrontar el estudio. El objetivo es encontrar el máximo balance entre las métricas explicadas anteriormente y el tiempo de ejecución de cada uno de los modelos. (James, Witten, & Tibshirani, 2013)

3.4 Análisis de riesgos.

Dentro del contexto de análisis de riesgos para este proyecto, se debe tener en cuenta múltiples factores.

Al tratar con una importante cantidad de datos de diversas fuentes de datos distintas, existe el riesgo de que los datos recopilados de fuentes en línea no sean precisos o sean

incompletos. Por hablar de un ejemplo, si se quisiera extender el estudio a nivel de provincia o de Comunidad Autónoma, esto sería imposible por la incompletitud de los datos: en la fuente de datos no se está sesgando por Comunidad, sino a nivel de país. Para minimizar este riesgo, es importante validar la fuente de los datos y considerar la posibilidad de usar múltiples fuentes para corroborar la información. Además, conocer bien los objetivos a alcanzar ayuda a que luego no sea un problema el que los datos no estén completos.

Dentro del apartado tecnológico, el uso de técnicas como web scraping para la descarga de datos o el procesamiento de lenguaje natural puede ser técnica y compleja. Más aún en aquellas páginas donde se quiere obtener información y no existe una manera sencilla para obtenerlos. Existe el riesgo de que surjan problemas técnicos durante la implementación de estas técnicas, como errores en el código o cambios en la estructura de los sitios web. Para mitigar este riesgo, es crucial realizar pruebas exhaustivas de las herramientas y scripts utilizados y estar preparado para realizar ajustes en caso de problemas inesperados.

El análisis de grandes volúmenes de datos puede ser desafiante y propenso a distintos errores, desde el empleo de técnicas estadísticas complejas como la regresión hasta la interpretación y el análisis de los mensajes en el análisis de sentimiento. Existen varios riesgos: desde el punto de vista interpretativo hasta el punto de vista de fallos externos a la interpretación. Dentro del análisis de sentimiento es posible filtrar por localización, pero hay riesgo de que haya multitud de usuarios utilizando una VPN y conectándose a la red social estudiada de una forma que molestaría a nuestro estudio. El otro riesgo es la interpretación incorrecta de los resultados o la utilización de modelos inadecuados para el análisis. Este último riesgo asociado al análisis de datos sí es individual y para mitigar al máximo este riesgo, es importante conocer las técnicas estadísticas utilizadas y validar los resultado stanto como sea posible.

A la hora de interpretar unos datos en estudios como este, es probable realizar una interpretación sesgada, especialmente si hay una predisposición hacia ciertas conclusiones. Para reducir al máximo este riesgo, es importante mantener una postura imparcial durante el análisis y considerar todas las perspectivas posibles antes de sacar conclusiones definitivas.

Otro riesgo asociado a este estudio más concreto es el riesgo a utilizar datos para el estudio cuando realmente estos datos vienen asociados a otro tipo de sector. Centrando la vista al reporte, dentro del sector agroalimentario hay multitud de empresas que se dedican a ello. Más en concreto dentro del sector de los supermercados y la venta al individuo se encuentran múltiples empresas. El problema es que estas múltiples empresas muchas de ellas además de tener la parte principal dedicada al sector agroalimentario, tienen multitud de otras secciones dentro de otros departamentos. Esto hace que a la hora de obtener datos de importancia social o influencia online, estos datos no reflejan solo lo que produce la empresa dentro del sector agroalimentario, sino que refleja todo lo que la empresa repercute hacia el usuario, sea este usuario parte de los clientes del sector agroalimentario o sean parte de un departamento de ventas externo a ello. Por ejemplo: la empresa “Carrefour” no sólo se dedican a la venta de alimentos, sino que tienen gran cantidad de otros productos como ropa o electrodomésticos, lo que hace que a la hora de tener repercusión social, tengan una mayor repercusión que otras empresas que solamente se dedican a la venta de alimentos. Pueden tener un mayor potencial de seguidores en redes sociales pero cabe la posibilidad de que dichos seguidores sean clientes del sector de electrodomésticos o de ropa y no del sector agroalimentario. Quizá estos valores se tomen en el estudio como valores generales en alguna ocasión y se estén comparando con los valores de impacto social de otras empresas que no tienen esos sectores.

3.5 Marco legal y ético.

Cuando se realiza un estudio en donde se necesitan obtener datos de multitud de fuentes distintas, se corre un riesgo muy alto dentro del marco legal y ético.

Es fundamental cumplir con las leyes de privacidad de datos, como el Reglamento General de Protección de Datos (GDPR) en la Unión Europea o leyes similares en otros territorios. Para el caso que nos compete, en España, se tiene la Agencia Española de Protección de Datos que es la autoridad encargada de supervisar y garantizar el cumplimiento de la normativa de protección de datos en territorio español. La ley que establece las reglas en cuanto a la protección de los datos en España es la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD). Se deben tomar las suficientes medidas para garantizar que los datos recopilados se obtengan de manera ética y legal, respetando la privacidad de los individuos y las políticas de privacidad de las plataformas en línea.

Además de las leyes de protección de datos, es importante cumplir con cualquier otra regulación aplicable relacionada con la investigación, la publicación y el uso de datos en línea. Se deben considerar, para nuestro caso, las leyes y regulaciones específicas del sector agroalimentario.

Cuando sea necesario recopilar datos de usuarios o clientes, se debe obtener un consentimiento informado de acuerdo con las leyes y regulaciones pertinentes. Es importante para lograr tener el consentimiento, explicar cómo se utilizarán los datos y ofrecer a los participantes la opción de optar por no participar en la recopilación de datos si así lo desean. En el caso de este Trabajo Final de Grado no se utiliza datos personales, por lo que se evita este riesgo en cierta medida.

Por otra parte, al utilizar técnicas como el web scraping para recopilar datos en línea, se deben respetar los derechos de propiedad intelectual de los propietarios de los sitios web objetivo, evitando infringir derechos de autor u otras formas de propiedad intelectual.

Se debe mantener la integridad y la objetividad en todas las etapas del estudio, evitando sesgos y conflictos de intereses que puedan comprometer la validez de los resultados. Para ello, se debe asegurar que la investigación se lleve a cabo de manera ética y responsable.

4. RESULTADOS Y VISUALIZACIÓN (Evaluación, validación y despliegue)

4.1 Caso concreto: Top-5 supermercados España

Antes de comenzar con el estudio propiamente práctico, se va a hacer una pequeña introducción en base a que se conozca cuál es la situación de los 5 supermercados con mayor cuota de mercado en España. Tras obtener los datos desde Kantar WorldPanel, se puede visualizar gráficas mes a mes durante los 3 años del estudio con datos completos de todos los meses para poder visualizar cuáles son las empresas punteras.

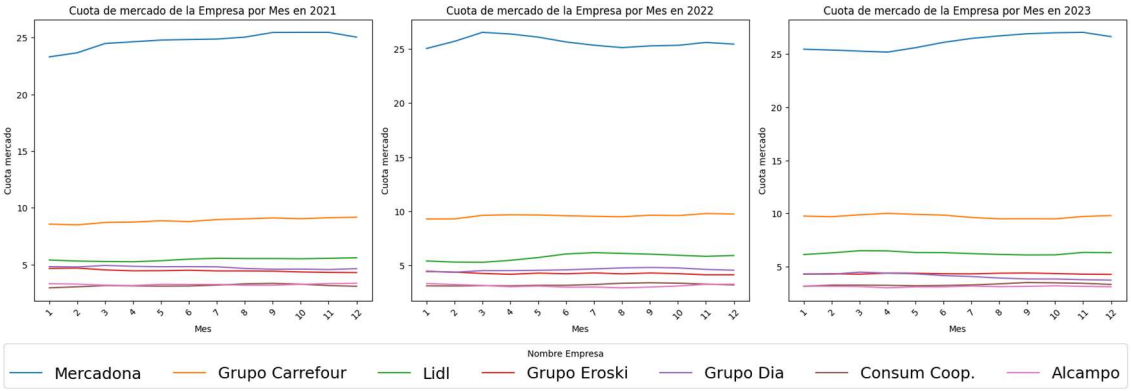


Ilustración 14. Evolución de la cuota de mercado por empresa del sector (Fuente: Propia)

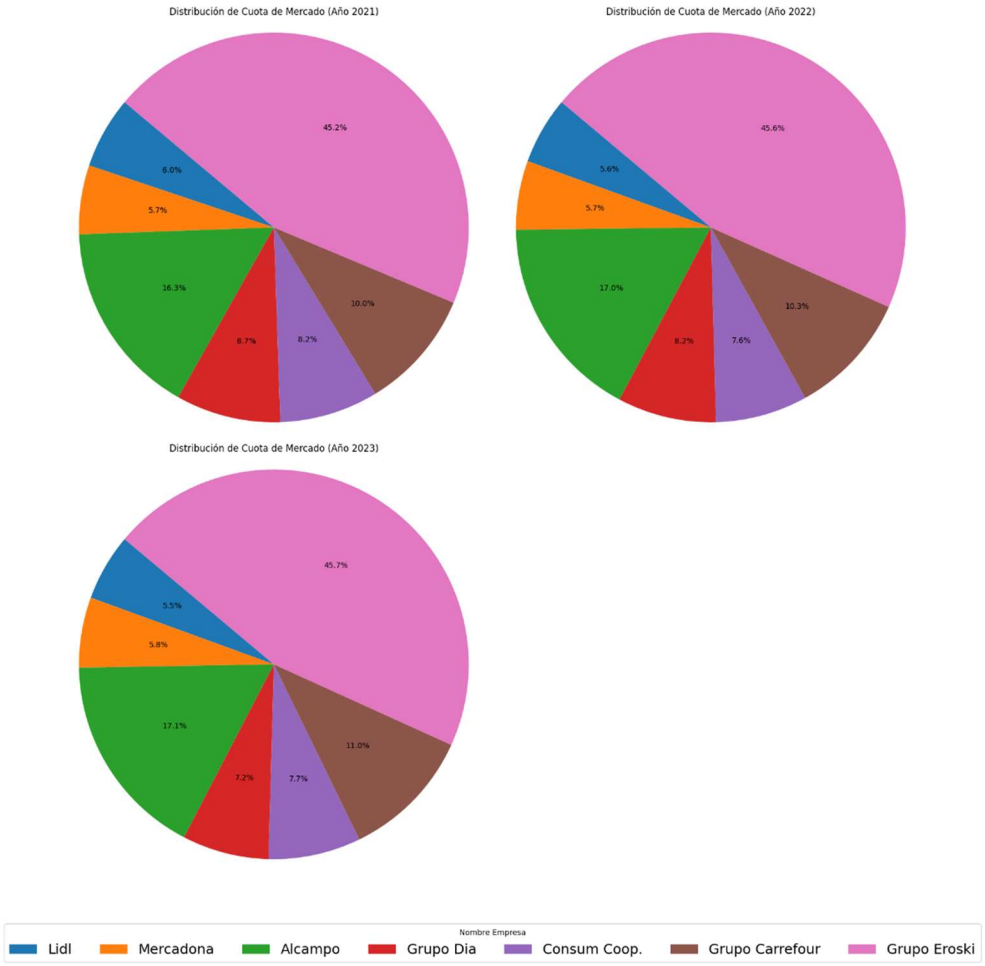


Ilustración 15. Distribución de cuota de mercado por empresa (Fuente: Propia)

Se puede visualizar para todos esos años que se mantiene como una constante que el supermercado con una cuota de mercado mayor es, y a mucha diferencia del siguiente, la empresa Mercadona; siendo la segunda con también bastante clarividencia el grupo Carrefour, y siguiéndole el resto de empresas del sector, con porcentajes similares entre ellas.

Sin duda una de las conclusiones que se desea obtener mediante este reporte es saber si esta cuota de mercado está relacionada con la popularidad online o qué cantidad de importancia tiene esta popularidad online a la hora de establecer las cuotas de mercado, para conocer cuán importante es para una empresa del sector manejar correctamente toda la infraestructura online y que repercuta positivamente en las ventas. En los próximos apartados se verá con más detenimiento la influencia que tienen, tanto en redes sociales como en búsquedas online desde Google Trends y se responderá a dicha pregunta.

4.1.1 Subidas en popularidad online de una de ellas y la influencia en la otra. ¿Hay relación con Google Trends?

Una forma clara y concisa de medir la popularidad online en cualquier sector es la búsqueda de las distintas empresas en Google, para lo que se hace uso la herramienta Google Trends. Como bien se ha explicado en el apartado de extracción y preparación de datos, Google Trends asigna una puntuación a la búsqueda de un término, teniendo en cuenta varias variables como la cantidad de usuarios que han realizado la búsqueda, el tiempo que han estado dentro de la búsqueda, entre otras muchas variables. En el siguiente gráfico se puede observar la evolución de la puntuación asignada a cada uno de los 5 supermercados con mayor cuota de mercado en España durante los últimos 5 años.

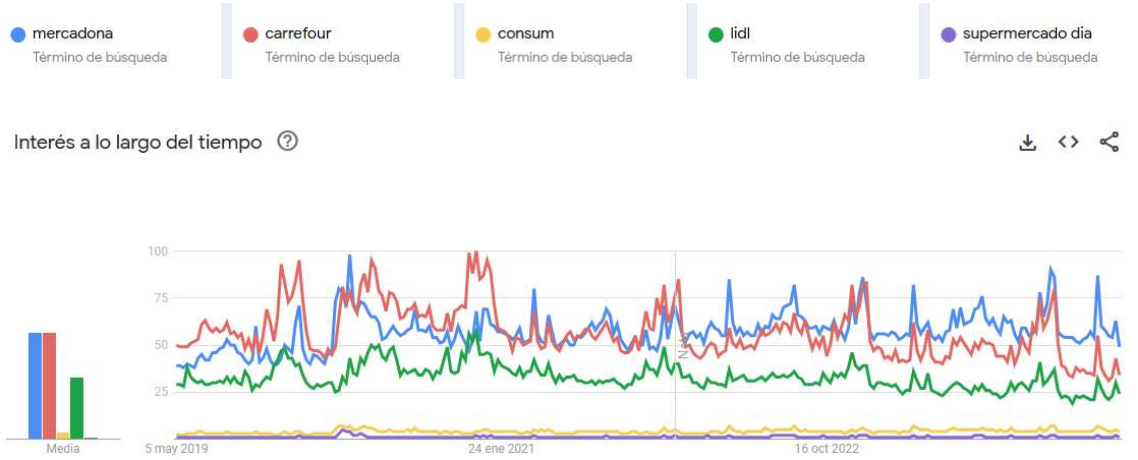


Ilustración 16. Evolución Interés Principales Supermercados desde 2019 (Google Trends)

Es por ello que sería interesante para responder a la pregunta de si hay o no relación entre la cuota de mercado y la popularidad online, hacer uso de las puntuaciones que asigna Google Trends. Si la relación existiese, las empresas como Mercadona o Carrefour con mayores cuotas de mercado, también deberían ser las que tengan mayor puntuación de Google Trends asignada.

Para eso se va a mostrar ahora un gráfico de dispersión donde en el eje X se mostrará el valor de la cuota de mercado obtenida de Kantar WorldPanel, y en el eje Y la puntuación asignada por Google Trends. Cada uno de los valores que se muestran son para un mes de un año concreto, y para mayor facilidad de interpretación se pintarán las instancias de cada una de las empresas de un color.

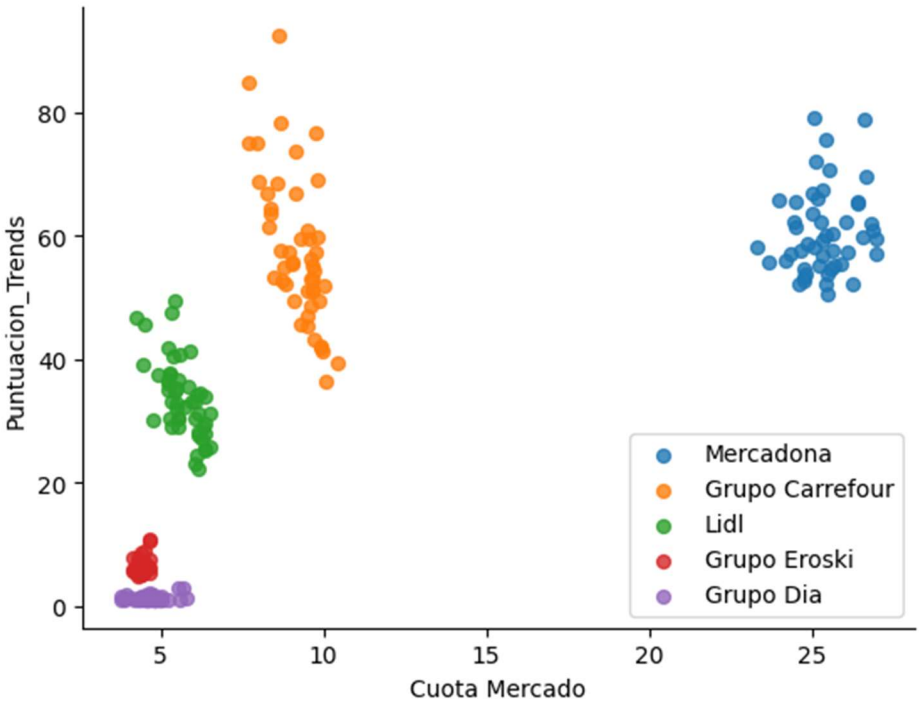


Ilustración 17. Scatter Plot Cuota de Mercado vs Puntuación Asignada por Google Trends por supermercado (Fuente: Propia)

Tras la visualización del gráfico, se obtienen un par de conclusiones claras. Todos los valores tanto de cuota de mercado como de popularidad de Google Trends siguen un cierto patrón; es decir, ninguna empresa tiene observaciones que disten mucho del grueso de las observaciones de esa misma empresa. Dicho de otra forma, están bien segmentadas, se pueden distinguir perfectamente los grupos. Y en respuesta a la pregunta, ¿hay relación entre la cuota de mercado y la popularidad online? A grandes rasgos: sí. Desgranando un poco más, no hay una relación directa completamente, pues si así fuera lo que se vería sería una especie de línea diagonal con las instancias de cada empresa sobre esa línea o formando una "nube" de puntos sobre esa línea de tendencia, y como se observa no es el caso. Hay varias instancias de la empresa "Carrefour" que, pese a tener mucha menos cuota de mercado que la empresa "Mercadona", tiene una puntuación de Trends superior, aunque no es la tónica habitual. Se observa que las que menor cuota de mercado tienen también tienen una muy baja puntuación de Trends.

En el siguiente gráfico de caja y bigotes se puede observar las diferencias de puntuación para cada una de las empresas. Para poder comprender el uso de este tipo de gráficos, en el anexo 4 se explica minuciosamente cómo utilizarlos.

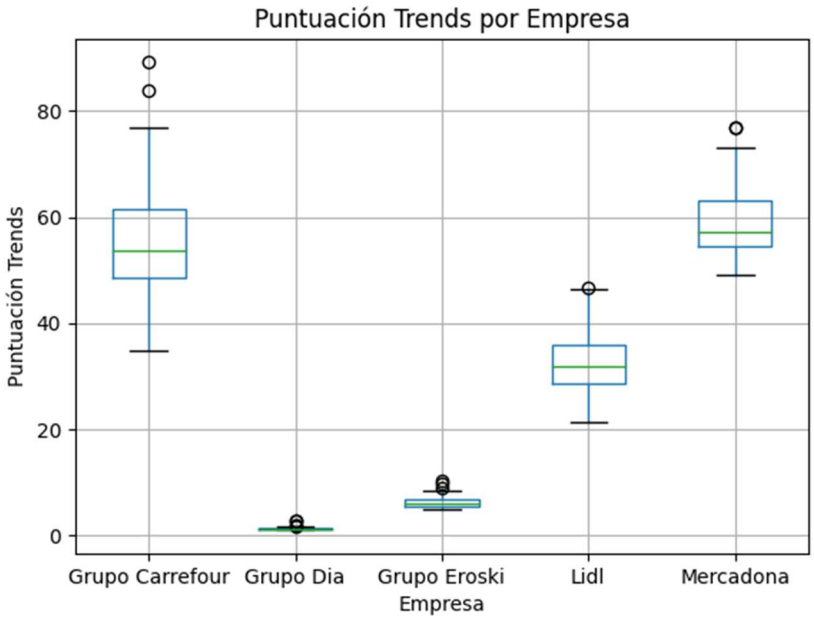


Ilustración 18. Diagrama Box-and-Whisker puntuación asignada por Google Trends por empresa (Fuente: Propia)

Se puede observar lo mismo que en el gráfico de dispersión anterior, pero con la peculiaridad que aquí se observa que la puntuación de "Mercadona" siempre está en un rango mucho más estrecho que el caso de "Carrefour", pese a que Carrefour también llega muchas veces a la misma puntuación de Mercadona. Tanto en Mercadona como en Carrefour se observan datos atípicos, un par de meses en Carrefour y un mes en Mercadona, se ha obtenido más puntuación en Trends de la que marca el resto. Con esto se quiere mostrar que, pese a que hay algunos valores de Carrefour superiores a Mercadona, en general, los valores de Mercadona son mayores que los de Carrefour, al igual que su cuota de mercado, pese a que la diferencia en cuota de mercado es muchísimo superior a la diferencia en popularidad online medida por Trends. (Vignesh, Pavithra, Dinakaran, & Thirumalai, 2017)

En conclusión, a grandes rasgos sí hay relación entre la puntuación de Trends y la cuota de mercado pese a que no hay una relación directa (hay mucha más diferencia en la cuota de mercado que en la popularidad online). El resto de empresas sí que se nota que son muy inferiores a estas dos primeras. Como ya se ha comentado en el análisis de riesgo, un posible peligro de este análisis era que empresas como Carrefour además de pertenecer al sector agroalimentario, pertenecen a múltiples sectores y a la hora de medir la popularidad online estos otros sectores también los tenemos en cuenta, por lo que es una posible razón de la cercanía que tiene con otra empresa como Mercadona, la cual sí se dedica íntegramente al sector agroalimentario.

4.1.2 Relación con métricas de Redes Sociales.

Las redes sociales suponen una gran herramienta a la hora de impulsar una cuenta tanto personal como profesional a un nuevo nivel de popularidad. Redes sociales como las ya mencionadas anteriormente como Twitter o Instagram suponen una fuente de ingresos y popularidad para muchas personas, de donde ha surgido el nuevo término de "influencers", los cuales gracias a publicidad y a sus propios contenidos por redes sociales, llegan a mucha gente alrededor del planeta y se ha convertido en su principal trabajo.

¿Sucede también con las empresas del sector agroalimentario? Esta es la cuestión que se va a intentar resolver en este apartado, donde utilizando gráficos de dispersión se buscará

dilucidar si alguna de las métricas de redes sociales que se han extraído suponen una gran diferencia a la hora de establecerse con la mayoría de la cuota de mercado del sector. En los siguientes gráficos se observará la comparación de la cuota de mercado frente a cada una de las métricas de redes sociales: tweets publicados, variación de seguidores en twitter, “me gustas” en Facebook, impresiones en Facebook, nuevos seguidores en Instagram y publicaciones en Instagram, en ese orden. Cabe destacar que para el caso de la empresa “Carrefour” no se ha podido incluir en el análisis debido a la falta de información sobre ella.

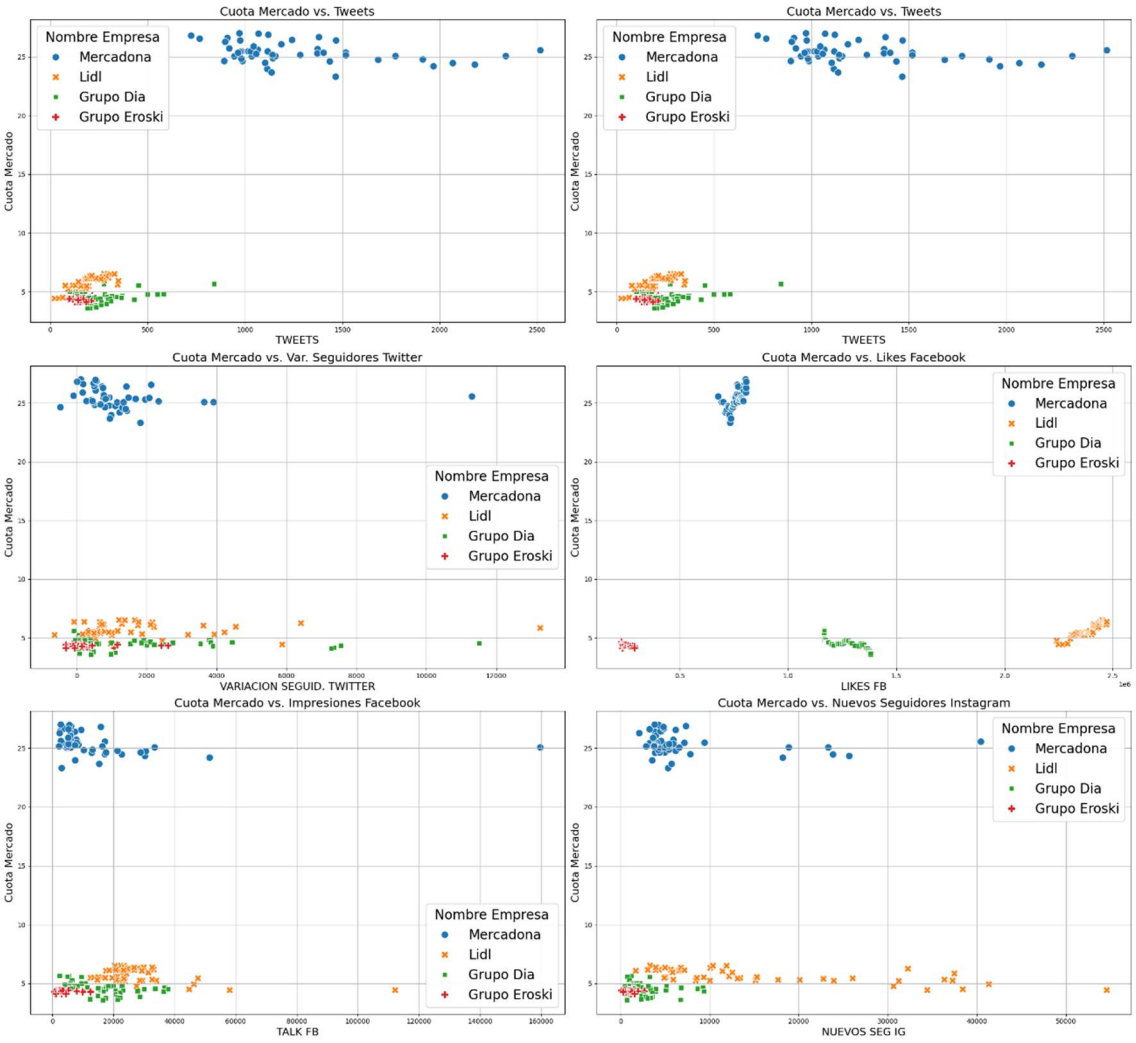


Ilustración 19. Scatter Plot métricas de Redes Sociales vs Cuota de Mercado por empresa (Fuente: Propia)

Para mayor comodidad y comprensión, se ha decidido marcar en una leyenda las observaciones dependiendo del nombre de la empresa a quien hacen referencia.

Para extraer conclusiones se irá visualizando gráfico a gráfico si se tiene o no algún tipo de relación directa. Empezando con el primero de los gráficos, el que enfrenta número de tweets posteados frente a la cuota de mercado correspondiente sí se puede observar cierta relación y es que todas las observaciones con gran cuota de mercado (aproximadamente el 25%) que se corresponden con Mercadona, suelen tener una gran cantidad de tweets, superior a cualquiera de las otras empresas del sector. Por otra parte, las 3 empresas restantes que comparten números similares en cuanto a cuota de mercado, también tienen un número de tweets similares, siempre inferiores a los de Mercadona. En conclusión, hay una relación positiva en cuanto a los tweets publicados por la cuenta del supermercado en un mes concreto y la cuota de mercado que se obtiene. Además, poniendo el foco en el análisis de cada empresa, se puede observar

cierta correlación entre el número de tweets publicados en ese período y la cuota de mercado obtenida. Esta correlación se puede observar claramente en las empresas Lidl y Grupo Dia, mientras que en Mercadona se observa que no se tiene esa correlación. Esta se puede observar de la forma en que a mayor número de tweets publicados mayor cuota de mercado se obtiene, dentro de la misma empresa.

El segundo de los gráficos, que muestra la variabilidad de seguidores frente a la cuota de mercado se observa que no hay relación. Las cuatro empresas tienen un valor de variación de seguidores similar (viendo todas las observaciones en promedio), mientras que la cuota de mercado de una de ellas es muy superior. Esto nos hace indicar que efectivamente no hay especial relación entre esta métrica y la cuota de mercado de cada empresa. Algo similar ocurre en el cuarto gráfico que muestra el número de impresiones en Facebook frente a la cuota de mercado. La mayoría de valores tienen impresiones similares mientras que en la cuota de mercado sí hay mucha diferencia.

El caso del número de “me gustas” en Facebook frente a la cuota de mercado es un caso especial y es que observamos que las distintas empresas agrupan sus observaciones lo que significa que siempre tienen un valor de esta métrica similar a lo largo de los meses. Todos los valores de Lidl tienen un valor de esta métrica muy superior a los de Supermercados Día, que a su vez tienen valores superiores a los de Mercadona, que son superiores a los de Grupo Eroski. Pero esta superioridad en me gustas no se traduce en una mayor cuota de mercado.

En el caso de las dos métricas provenientes de Instagram sucede algo similar a lo del segundo y cuarto gráfico, y es que todas las empresas tienen valores similares en la métrica pero sí se distingue una gran diferencia entre la cuota de mercado de Mercadona y el resto, siendo que los valores de las métricas de redes sociales son similares.

En conclusión, el análisis de las métricas de redes sociales en relación con la cuota de mercado de las distintas empresas del sector revela diversos patrones. Por una parte, existe una correlación positiva entre el número de tuits y la cuota de mercado, lo que se traduce que a mayor actividad en Twitter, mayor cuota de mercado. Sin embargo, otros indicadores, como los cambios de seguidores en redes sociales o de impresiones en Facebook, carecen de correlación con la cuota de mercado. Además, si bien algunas empresas pueden tener una gran cantidad de Me gusta en Facebook o métricas similares en Instagram, esto no se traduce necesariamente en una mayor participación de mercado. En resumen, si bien algunas métricas de las redes sociales pueden estar relacionadas con el éxito en el mercado, otras métricas no reflejan claramente este vínculo, lo que resalta la complejidad de la relación entre la presencia en las redes sociales y el desempeño comercial de las empresas de supermercados.

En el próximo apartado se tratará más a fondo el tema de las correlaciones entre las variables y la cuota de mercado, utilizando gráficos de correlación y métodos algorítmicos como Random Forest

4.1.3 Análisis de las variables de popularidad online en la cuota de mercado.

Uno de los factores más importantes a tener en cuenta en cuanto a cómo afectan las redes sociales y la popularidad online con respecto a la cuota de mercado es analizar cuáles variables tienen más influencia a la hora de tener mayor o menor valor a razón de ella. Para obtener valores numéricos que indiquen cuán importante es la variable concreta a la hora de establecer una cuota de mercado se va a emplear un análisis de correlación de variables. En el caso de tener muchas más variables que pudiesen alterar el valor de la variable objetivo se podría aplicar otros métodos como el Análisis de Componentes Principales o el Análisis Factorial de

Correspondencias (si las variables fuesen numéricas se usaría el primero y en caso de ser explicativas se utilizará el segundo). Estos métodos reducen la dimensionalidad de las variables predictoras y muestran cuáles de las variables originales tienen mayor influencia en las principales dimensiones, y por tanto, cuáles explican mayor variabilidad de la variable objetivo.

En este caso, se utilizará el DataFrame completo: las cuotas de mercado extraídas desde Kantar WorldPanel, los datos extraídos desde Google Trends y las métricas extraídas de redes sociales de SocialBlade. Uniendo todos estos datos se obtiene un DataFrame sobre el que se realizarán las correlaciones necesarias, buscando ver cuáles de las variables hacen que aumente o disminuya el valor de la cuota de mercado.

Para tenerlo en cuenta y pese a que es muy conocida, la correlación es una medida estadística que describe la relación entre dos variables y nos indica la fuerza y la dirección de esa asociación. Una correlación fuerte significa que los valores de una variable tienden a cambiar de manera consistente con los valores de la otra variable. Una correlación débil indica que hay poca relación entre las variables. La correlación puede ser positiva o negativa: una correlación positiva significa que a medida que aumenta el valor de una variable, también tiende a aumentar el valor de la otra variable, y viceversa. Los valores varían entre -1 y 1, siendo -1 correlación negativa perfecta y 1 correlación positiva perfecta. Valores como 0 o cercanos a este valor indican que no hay o hay muy poca correlación entre ambas variables.

Las variables con las que se buscará relación a la variable de Cuota de Mercado son: 'TWEETS', 'VARIACION SEGUID. TWITTER', 'LIKES FB', 'TALK FB', 'NUEVOS SEG IG', 'CONTENIDO IG','Puntuacion Trends'. Ya se conoce cada una de las variables que explican, debido a que están explicadas en el apartado 3, así que se pasa a continuación a la correlación.

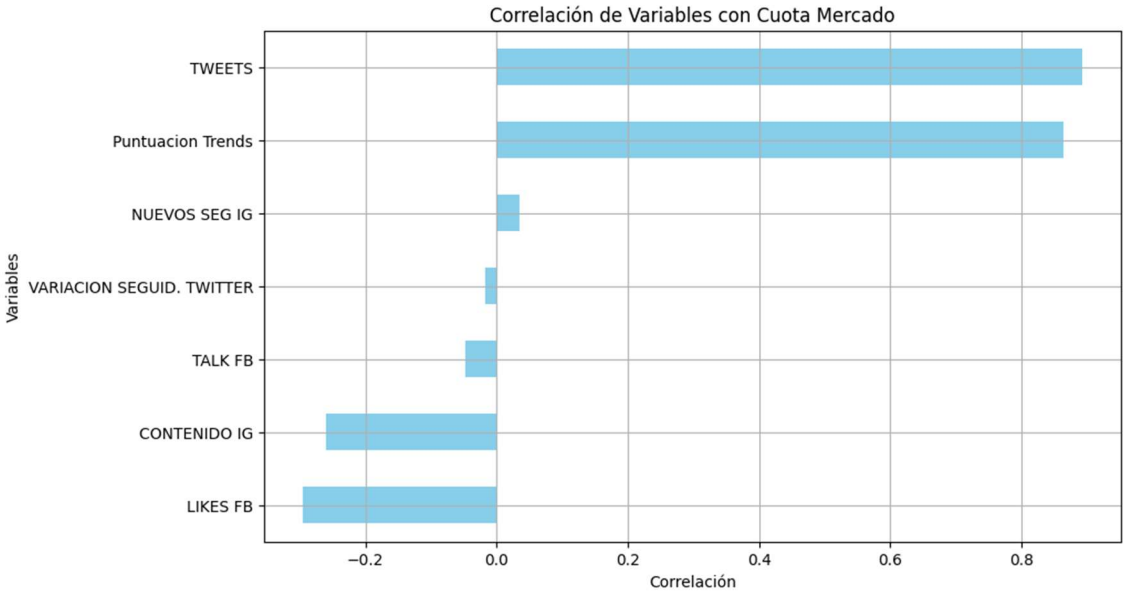


Ilustración 20. Correlaciones de variables con la Cuota de Mercado (Fuente: Propia)

Tal y como se observa, se tiene un gráfico de correlación entre el Valor de la Cuota de Mercado de la Empresa y las siete variables de popularidad (en el eje Y). En el eje X se tiene el valor de correlación. Tal y como se explica en la parte superior, valores cercanos a 1 ó -1 se traducen en mucha correlación y los cercanos a 0 en poca.

Teniendo en cuenta el gráfico se puede corroborar lo establecido en los dos apartados anteriores, donde se podía observar de forma más visual: y es que los valores de tweets posteados y la puntuación que indica Google Trends (explicada anteriormente) tienen una

correlación positiva superior al 80%, lo que explica que altos valores de estas variables están asociados con altos valores de cuota de mercado. Por otra parte, el contenido publicado en Instagram o los likes recibidos en Facebook tienen una pequeña correlación negativa, lo cual hará que se produzca lo contrario a los tweets. Las otras tres variables no muestran apenas correlación.

Otra opción para visualizar las correlaciones es no partir de la base de la relación lineal (como sí hace el gráfico de correlación anterior) y ver la importancia de las variables utilizando un algoritmo, en este caso se ha utilizado Random Forest.

Creando un regresor de tipo randomForest y haciendo uso de la función feature_importances, se obtiene la importancia relativa de cada variable con respecto a la variable objetivo: Cuota de Mercado en este caso. Ambos gráficos tienen similitudes y explican lo mismo y a mayor valor de correlación o importancia mayor será la relación, pero también se tiene alguna diferencia que cabe destacar: Los valores de correlación pueden ser positivos o negativos, mientras que los valores de importancia solo pueden ser positivos, sea una importancia "positiva" o "negativa", simplemente mide cómo esa variable hace que disminuya el error del modelo, sin importar el sentido. Además, no toma una supuesta relación lineal entre las variables como sí hace las correlaciones anteriores. En el siguiente gráfico se observa los valores de importancia de cada variable al modelo de Random Forest.

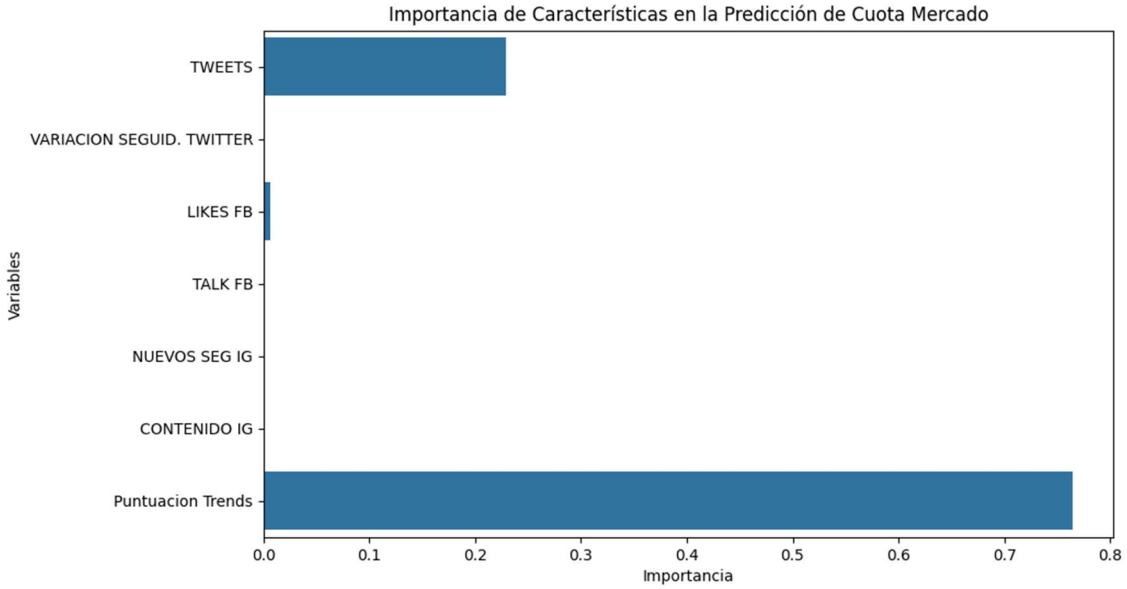


Ilustración 21. Importancia de Variables frente a Cuota de Mercado (Fuente: Propia)

Como se ve el gráfico es similar al anterior excepto que el eje X ahora es el valor de importancia y sólo contiene valores positivos. Las conclusiones son similares y es que los valores con altas correlaciones tienen altos valores de importancia también aquí, y los que apenas tenían correlación no tienen ninguna importancia. En este caso en concreto la puntuación de Trends y los tweets publicados tienen bastante importancia, los Likes de Facebook tienen algo de importancia, pero el resto de variables ninguna tiene.

También podemos visualizar estos valores en un gráfico de tarta, que es más intuitivo visualmente (el valor de importancia de cada variable está en la leyenda):

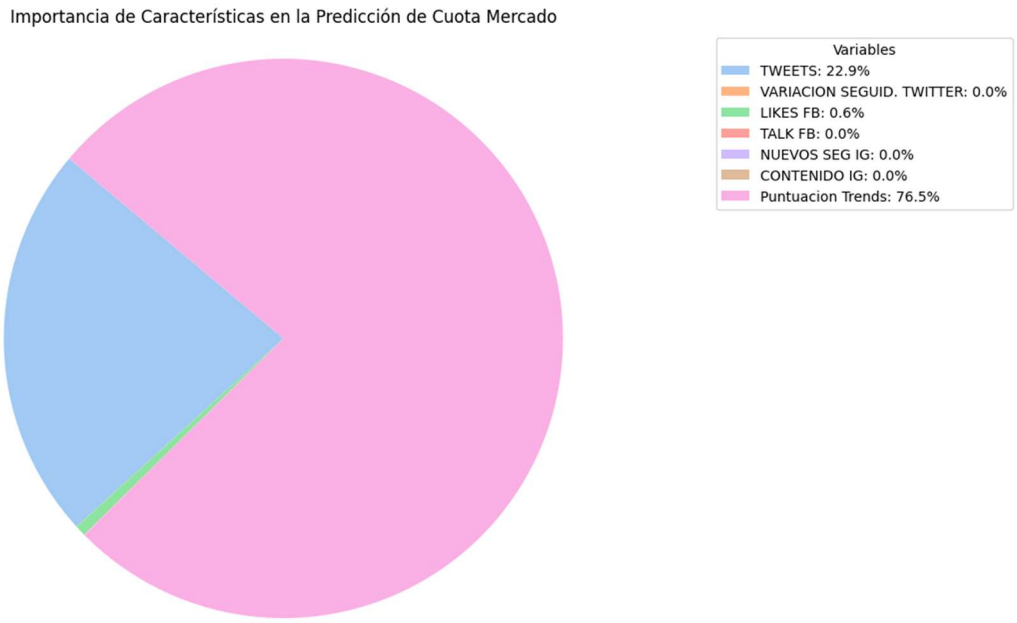


Ilustración 22. Importancia frente a Cuota de Mercado. Gráfico de Tarta (Fuente: Propia)

4.1.4 Subidas y bajadas de la cuota de mercado en la propia empresa, ¿influyen el resto de variables?

En esta sección se va a profundizar más a fondo en una empresa concreta y en cómo se puede observar si para esa empresa concreta del sector le afectan las variables de popularidad online. Para ello se va a seleccionar los valores correspondientes a la empresa "**Mercadona**", en los que se tienen desde 2020 hasta marzo de 2024 datos mes a mes tanto de la cuota de mercado que ocupan como del resto de variables provenientes de métricas de redes sociales o de impacto en Google Trends. Con el DataFrame formado por todas las variables a estudiar junto con la cuota de mercado de dicho mes, se puede comenzar el estudio de observar cuáles variables afectan más en la cuota de mercado de Mercadona.

Una forma sencilla de visualizar patrones de comportamiento de las variables con respecto a la variable a predecir es graficar los valores de la variable que corresponde a la cuota de mercado frente al resto, pudiendo observar así si cuando la variable nos interesa sube si hay alguna de las otras variables que siga el mismo patrón.

Para poder observar de forma precisa, se ha utilizado de la librería sklearn el método MinMaxScaler, gracias al cual se escalan los valores de las distintas columnas numéricas del DataFrame. En este caso, se ha elegido escalar de 0 a 100 todas las variables. De esta forma a la hora de visualizarlo todas las variables estarán escaladas a los mismos valores. En el gráfico inferior se observa la tendencia de la variable que interesa conocer (la cuota de mercado, en color azul y ligeramente más gruesa que el resto), frente al resto de variables.

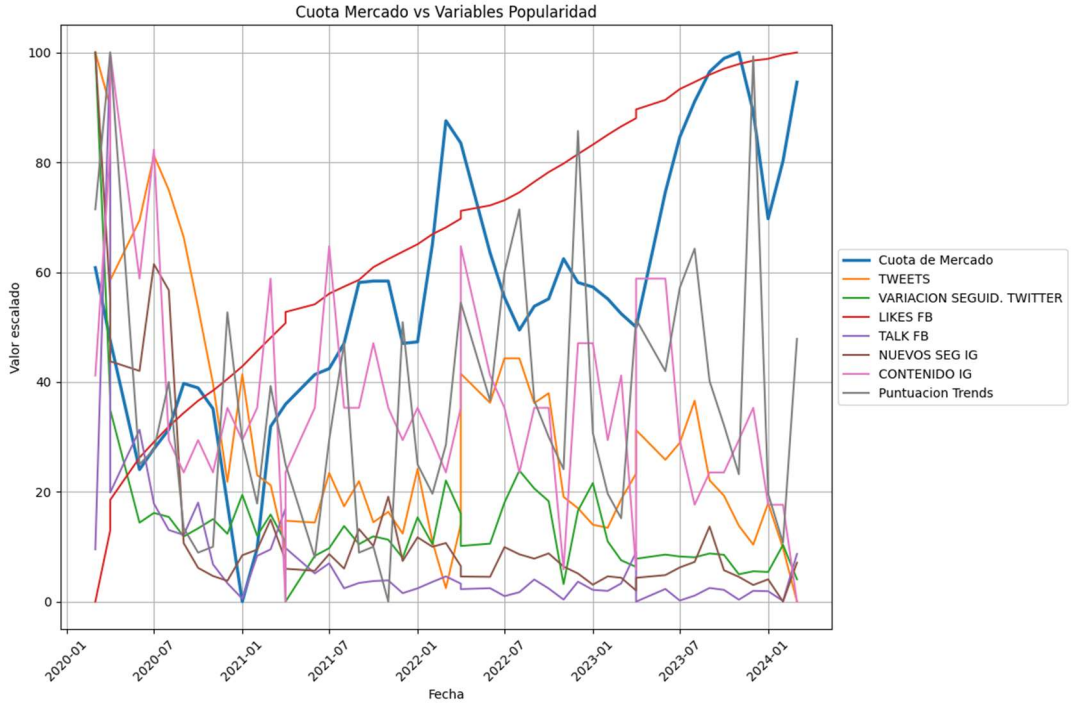
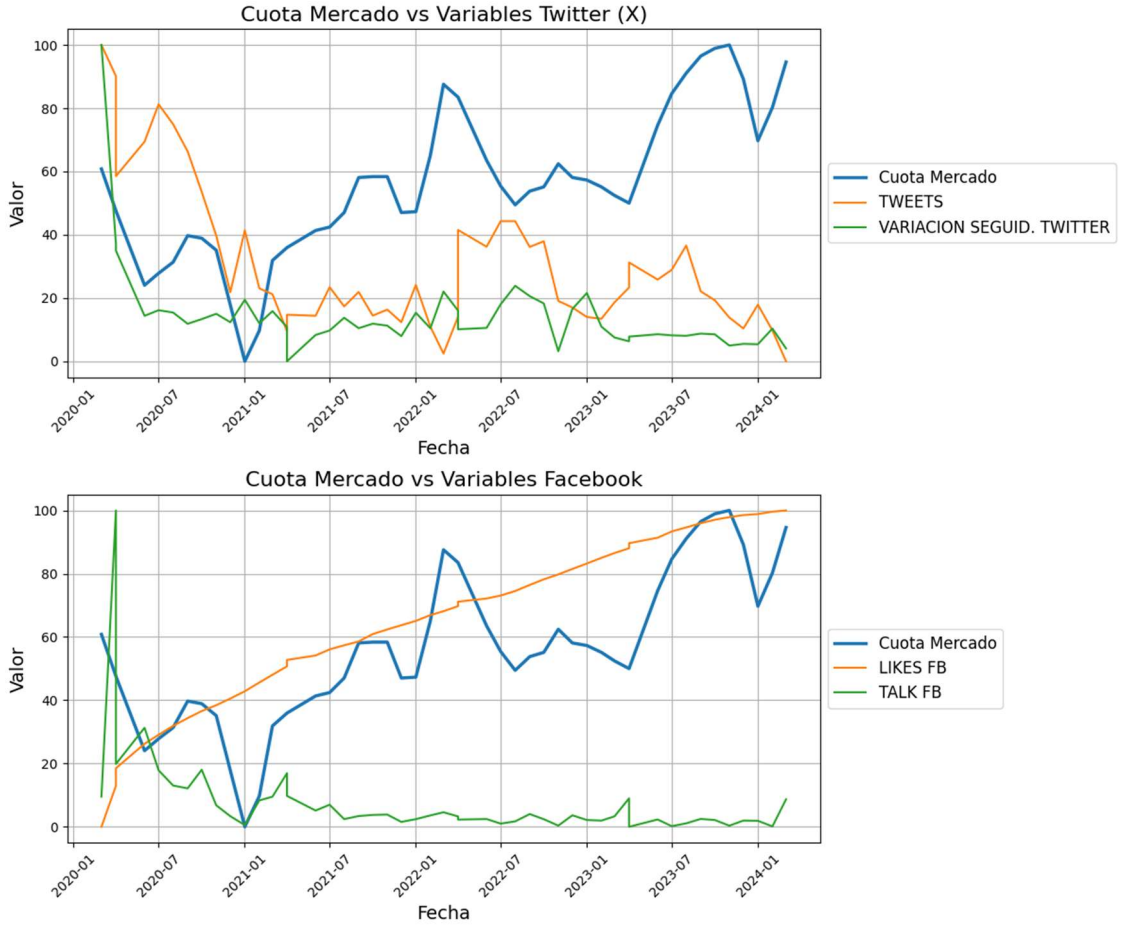


Ilustración 23. Gráfico de Línea: Cuota de Mercado vs Resto de Variables (Fuente: Propia)

Como se observan demasiadas variables, se ha desagregado las variables para mayor facilidad interpretativa:



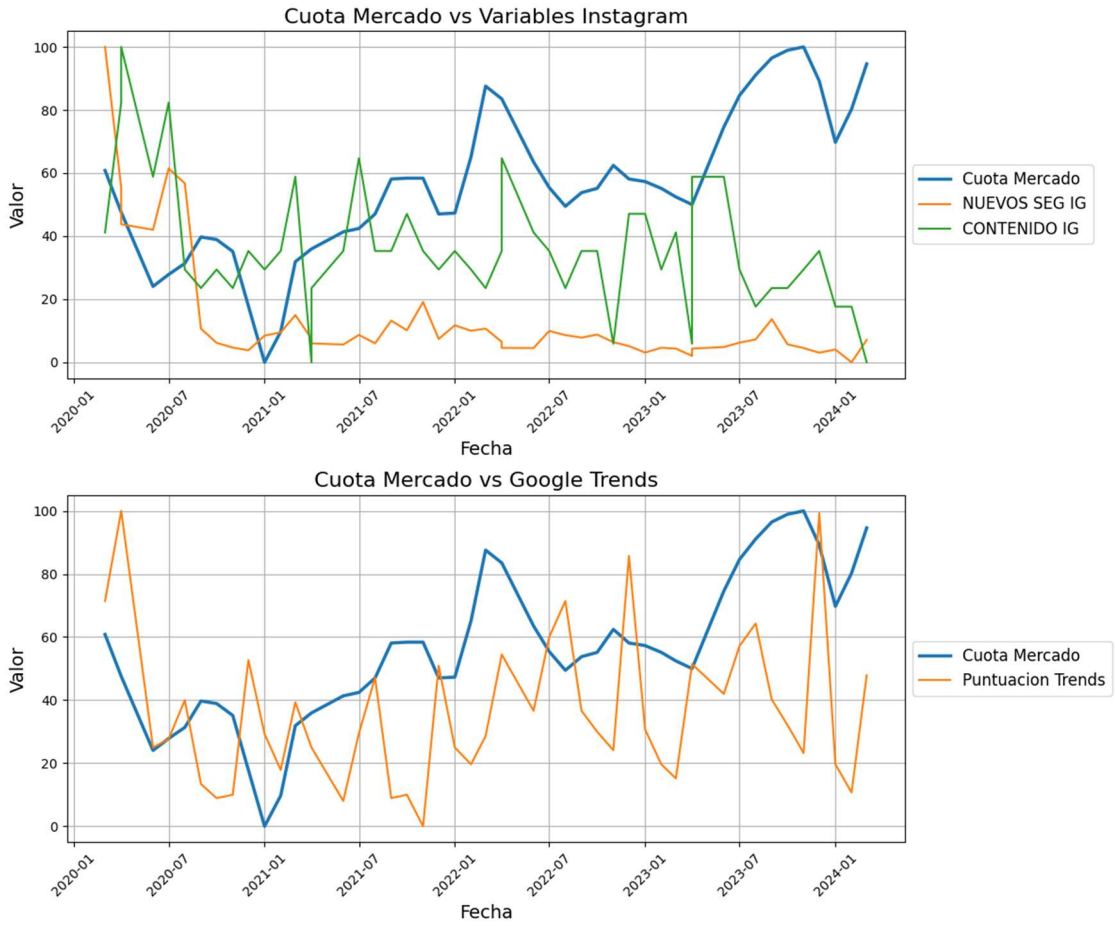


Ilustración 24. Gráfico de Líneas: Cuota de Mercado vs Resto de Variables (Desagregadas) (Fuente: Propia)

En los gráficos mostrados arriba, se puede ver ciertos patrones visuales como en la variable de contenido de Instagram, en número de Tweets o en Google Trends, pero igualmente no es sencillo de observar y no llega a ningún resultado concluyente. Por esta razón, se ha buscado otra forma de observar si alguna de las variables tiene que ver con el comportamiento de la cuota de mercado.

A partir de los datos escalados de Mercadona se ha construido un nuevo DataFrame en el que se almacenan las variaciones. Para entenderlo se observa el siguiente ejemplo: (la explicación de las variables están explicadas en el apartado de metodología)

Tabla 3. Registros de Ejemplo (Fuente: Propia)

Nombre Empresa	Cuota de Mercado	TWEETS	VARIACION SEGUIDORES TW	LIKES FB	TALK FB	NUEVOS SEG IG	CONTENIDO IG	Puntuacion Trends
Mercadona	60.810811	100.000000	100.000000	0.000000	9.527560	100.000000	41.176471	71.428571
Mercadona	47.567568	90.122768	37.222742	12.974298	100.000000	55.321534	82.352941	100.000000

Estas 2 observaciones del DataFrame original se transformarán al DataFrame de Variaciones resultando los valores de la segunda observación menos la primera, resultando:

Tabla 4. Variaciones Resultantes Tabla Anterior (Fuente: Propia)

Nombre Empresa	Cuota de Mercado	TWEETS	VARIACION SEGUIDORES TW	LIKES FB	TALK FB	NUEVOS SEG IG	CONTENIDO IG	Puntuacion Trends
Mercadona	-13.243243	-9.877232	-62.777258	12.974298	90.472440	-44.678466	41.176471	28.571429

De esta forma se obtiene un DataFrame de variaciones y se podrá visualizar de forma más concreta cuántas de las veces que la cuota de mercado de Mercadona baja bajan el resto de variables, y viceversa. Para ello y también haciendo uso de la programación en Python se ha obtenido sendas tablas que muestran para las variaciones positivas de la cuota de mercado cuántas veces alguna de las otras variables han aumentado en ese período también y para las negativas lo contrario. Cabe destacar que hay 18 variaciones negativas de la cuota de mercado entre 2 meses consecutivos y 25 variaciones positivas.

Tabla 5. Variación de Variables Entre Periodos cuando la Variación de Cuota de Mercado es Positiva (Fuente: Propia)

	Positivos	Negativos	%Positivos
TWEETS	7	18	28.00%
VARIACION SEGUID. TWITTER	11	14	44.00%
LIKES FB	25	0	100.00%
TALK FB	14	11	56.00%
NUEVOS SEG IG	13	12	52.00%
CONTENIDO IG	10	10	40.00%
Puntuación Trends	10	15	40.00%

Tabla 6. Variación de Variables Entre Periodos cuando la Variación de Cuota de Mercado es Negativa (Fuente: Propia)

	Positivos	Negativos	%Positivos
TWEETS	7	10	55.56%
VARIACION SEGUID. TWITTER	9	9	50.00%
LIKES FB	18	0	0.00%
TALK FB	9	9	50.00%
NUEVOS SEG IG	4	14	77.78%
CONTENIDO IG	7	10	55.56%
Puntuación Trends	10	8	44.44%

Se obtienen varias conclusiones observando las tablas anteriores. Primero de todo, la variable LIKES FB es una variable que desde 2020 ha ido en aumento y nunca ha disminuido, por lo que se puede decir que Mercadona ha ido obteniendo más “me gustas” en esa red social, pero eso no ha tenido nada que ver en su ocupación de cuota de mercado. Variables como “NUEVOS SEG IG”, es decir, los nuevos seguidores de instagram, sí parece explicar algo más el comportamiento de la cuota de mercado, pues de las 18 variaciones negativas, en 14 de ellas

ha habido variación negativa también en los nuevos seguidores en esta red social, y más del 50% de veces que ha habido variación positiva también ha habido variación positiva en esta variable. Por otra parte, las impresiones en Facebook también explican más del 50% de las subidas y la subida de contenido a Instagram también cabe indicar que es negativo para la cuota de mercado de Mercadona.

En conclusión, para la empresa “Mercadona” no se observa ninguna variable concluyente a la hora de decir que tenga un gran peso a la hora de variar la cuota de mercado de un período a otro. Se observa algún detalle, como se ha comentado anteriormente, pero no hay ninguna variable que lo muestre de una forma clara y directa.

En el Anexo 5 se puede observar el resto de tablas de variaciones y su relación con el resto de variables para “Grupo Dia”, “Lidl”, “Grupo Eroski”

4.2 Análisis de regresión sobre la cuota de mercado

Pese a que se ha observado que las variables de métricas de redes sociales no tienen gran relación directa con la cuota de mercado, quizá haya algún tipo de patrón entre esas variables que hagan que, unidas, se traduzcan en una mayor o menor cuota de mercado. Un objetivo de este estudio es analizar cuál sería la cuota de mercado para una empresa dados los valores de impacto social asociados a dicha empresa. Por ello se va a utilizar un método de regresión en el que, analizando los valores de impacto de redes sociales para las distintas empresas, predigan el valor de cuota de mercado que tendrán asociado. Para este fin, se utilizará el DataFrame de datos extraídos en Kantar WorldPanel, junto al DataFrame de los datos extraídos de SocialBlade, buscando así la relación que tienen las métricas de redes sociales con respecto a la cuota de mercado de las empresas del sector.

Por tanto, se tratará de predecir el valor de la cuota de mercado a partir de las otras variables relacionadas con las métricas de las redes sociales. La variable respuesta o variable a predecir será la cuota de mercado, mientras que el resto de variables serán las variables explicativas o predictoras.

4.2.1 Resultados del modelo y evaluación

- *4.2.1.1 Selección de parámetros, resultados y lanzamiento*

Para escoger el mejor método de regresión que se ajustase más a los datos de los que se dispone, se realizan varios métodos distintos, basándose en la teoría de las regresiones explicadas en el apartado de metodología y buscando maximizar las distintas métricas, tanto el MSE como el R2.

Para comenzar, en cada uno de los casos, se añade el método de perturbaciones aleatorias, para la que se define una función augment_row en el que se aumentan tantas filas como se desee, agregando un factor de ruido que se define por un número aleatorio dentro de una desviación estándar que se ha escogido. Se generan nuevas filas para cada una de las instancias iniciales sea cual sea la empresa. Dicho de otra forma, para cada una de las empresas y cada uno de los periodos de tiempo (mensuales) se generan tantas filas como se desee. Para cada uno de los casos se evalúan el rendimiento de cada una de las regresiones, aumentando 40, 60, 80, 100 y 150 filas (por cada una de las instancias “reales”) y desviaciones estándar de

0.1, 0.5 y 1. Se realizan las distintas regresiones y se analiza cuál de ellas proporciona un resultado más acorde a lo que se busca obtener.

También indicar que para cada una de las regresiones se destina el 20% de los datos al test y el 80% al train.

```
El mejor resultado es:
Menor MSE: std_dev: 0.1 n_new_instances: 100 MSE: 13.423867671347939
Mayor R^2: std_dev: 0.1 n_new_instances: 100 R^2: 0.837556904362134
```

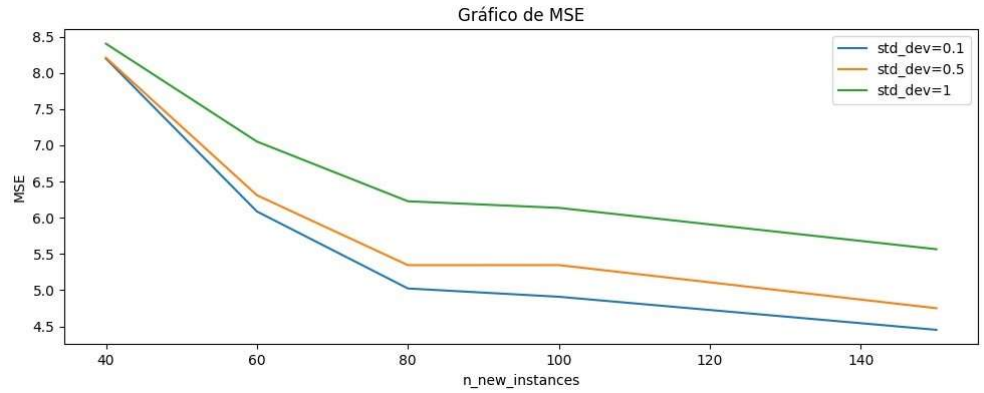
```
El mejor resultado para SVR es:
Menor MSE: std_dev: 0.1 n_new_instances: 150 MSE: 4.453508599047769
Mayor R^2: std_dev: 0.1 n_new_instances: 150 R^2: 0.9460433048040003
```

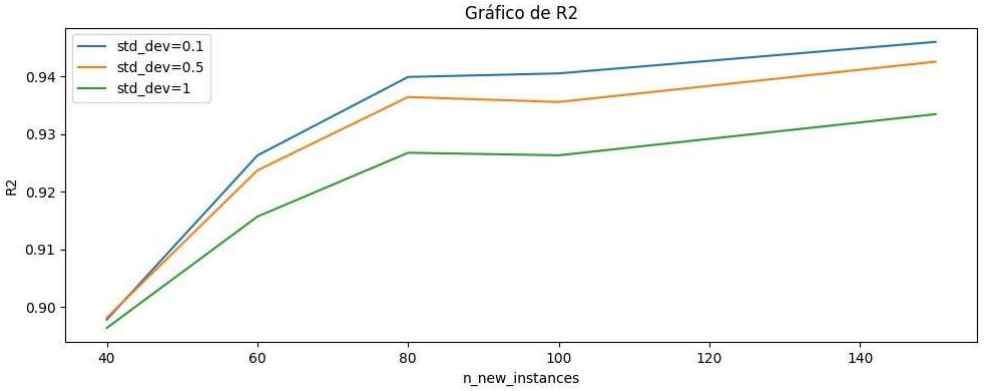
```
El mejor resultado para PLSR es:
Menor MSE: std_dev: 0.1 n_new_instances: 40 MSE: 16.564452177684572
Mayor R^2: std_dev: 0.1 n_new_instances: 40 R^2: 0.7934097988009907
```

Ilustración 25. Métricas Evaluación Regresiones (Fuente: Propia)

Una forma rápida de observar qué regresión puede ser más adecuada para nuestro estudio es observar los valores de MSE y de R^2 que tiene la mejor de las combinaciones de cada tipo de regresión. Se observa a simple vista que la regresión SVR o Support Vector Regression tiene valores mucho mejores que las otras regresiones, por lo que continuará el estudio basándose en la regresión mediante vectores soporte.

Para seleccionar ahora qué valores de filas agregadas y desviación estándar es la óptima para este estudio, se va a realizar una comparación de los distintos valores para observar de forma gráfica cuál será la combinación de parámetros escogida:





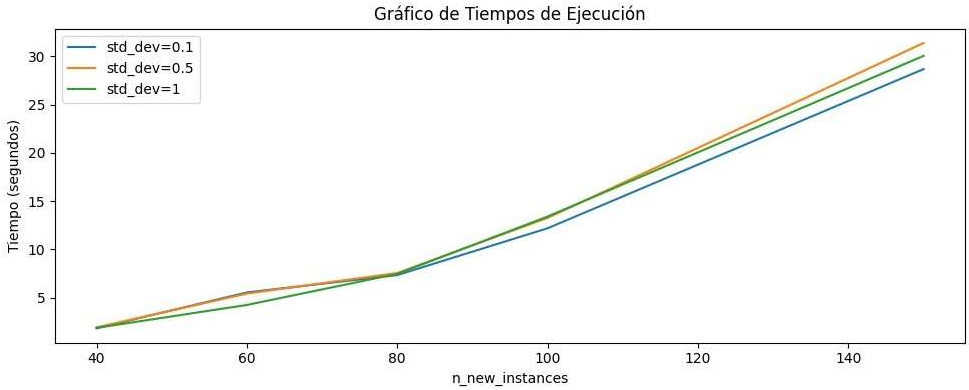


Ilustración 26. Gráficos para la elección de hiperparámetros SVR (Fuente: Propia)

En los gráficos superiores se observa en el eje X el número de nuevas instancias por cada instancia que se añade nueva, mientras que en el eje Y se muestra el MSE, el R^2 o el tiempo, respectivamente. En la leyenda se observa que cada una de las líneas es un número de desviación estándar distinta.

De los gráficos de MSE y R^2 se extrae la conclusión de que la desviación estándar de 0.1 muestra mejores resultados independientemente de la cantidad de nuevas instancias introducidas al DataFrame. Por otra parte, cuando el número de instancias es 80 en ambos gráficos se observa una especie de codo en el que los valores de las métricas dejan de aumentar o disminuir tan drásticamente para seguir con menos pendiente.

Ya se ha concluido que la desviación estándar de 0.1 es la óptima. Pero falta por obtener cuál es el número de instancias óptimo. Se observa que a mayor número de instancias mejor es el rendimiento de modelo, pero en el gráfico del tiempo se observa que a partir de 80 nuevas instancias el tiempo comienza a crecer de una forma muy abrupta, mientras que en las métricas no se observa tal diferencia entre ese número de nuevas instancias y los siguientes.

Por lo tanto, la conclusión basándose en los resultados obtenidos es que la regresión que mejores resultados obtiene es la regresión por vectores soporte (SVR), habiendo generado 80 nuevas instancias por cada instancia real con una desviación estándar de 0.1.

Ya se conoce cuál será el modelo base desde el que se partirá para buscar el modelo final. Se aplicará regresión por vectores soporte. Con los parámetros básicos se observa un 5.024 de Error Cuádratico Medio y un coeficiente de determinación de 0.9399, un 93.99%. Ahora el objetivo es encontrar nuevos parámetros dentro de la función de SVR para mejorar las métricas y alcanzar el modelo final.

Para realizar este tipo de regresión se utiliza la función SVR de la librería sklearn.svm. En un principio el único argumento que utiliza la función es kernel=”rbf”, indicando de esta forma que el kernel es radial. Pero, ¿hay alguna forma de mejorar su rendimiento? Para ver si hubiese alguna opción de mejorar el rendimiento utilizando algún hiper parámetro se realiza la prueba con los distintos tipos de kernel (poly, linear o sigmoid) y se observa que las métricas devuelven valores inferiores, por lo tanto, el modelo desprecia estos cambios y se mantiene con el kernel radial. Se puede probar a cambiar otros parámetros más internos de la función, pero visto el buen rendimiento del modelo con el kernel radial no vale la pena indagar en esa cuestión. Se continuará hacia adelante con el kernel radial utilizando Support Vector Regression.

En las próximas figuras se observa cómo aproxima el modelo cada uno de los valores de cuota de mercado reales de los supermercados:

Cabe destacar que la empresa “Grupo Carrefour” no está disponible, pues no se tienen suficientes valores de métricas de Redes Sociales como para tenerlas en cuenta en el modelo. Por otro lado, “Grupo Eroski” tiene valores incompletos a partir de julio de 2023.

Tabla 7. Comparación Cuota Mercado Original con la Predicha (Fuente Propia)

	Nombre Empresa	Año	Periodo	Cuota Mercado Orig	Cuota Mercado Predicha
0	Mercadona	2020	3	25.55	25.123559
1	Lidl	2020	3	4.76	3.199489
2	Grupo Dia	2020	3	5.55	10.554303
3	Mercadona	2020	4	25.06	24.875568
4	Mercadona	2020	4	25.06	25.035164
..	...	...	...	...	...
180	Grupo Dia	2024	2	3.58	1.340478
181	Mercadona	2024	3	26.80	23.862494
182	Lidl	2024	3	6.36	6.553036
183	Grupo Eroski	2024	3	4.13	8.992562
184	Grupo Dia	2024	3	3.68	1.341037

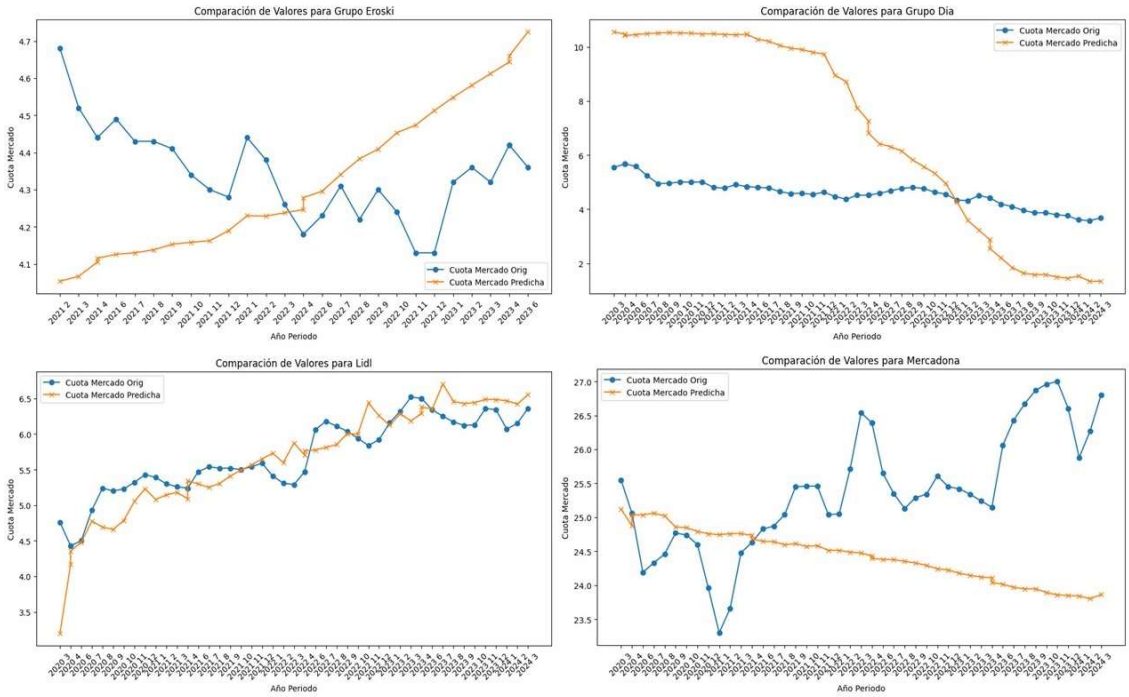


Ilustración 27. Comparación de Cuota de Mercado (Original-Predicha) por Mes por Empresa (Fuente: Propia)

Se puede llegar a varias conclusiones observando las distintas figuras y es que, en tres de los cuatro casos la aproximación se podría decir que es una muy buena aproximación, teniendo en cuenta las modificaciones que se le han tenido que realizar a los datos. Excepto los valores de “Grupo Dia”, en el resto de empresas se predice un valor de mercado similar al real (en “Mercadona” hay más diferencia, pero al tener un valor mayor la diferencia porcentual es similar al resto). Destacar la aproximación que se hace en “Lidl” donde, exceptuando el primer mes, se hace una predicción muy similar a la real. En el caso de “Grupo Eroski” los valores también son muy parecidos.

4.2.2 Prueba de introducción de una nueva empresa al sector

Lo que se va a buscar en este apartado es generar una situación ficticia en la que un empresario del sector agroalimentario quiere lanzar en España su empresa. Este empresario conoce los valores que tiene en cuanto a redes sociales, los nuevos seguidores que prevé obtener en el mes actual, la cantidad de posts que va a subir a las redes, etc. y quiere conocer cuál sería su situación en el mercado español teniendo los datos que tiene actuales.

Para este objetivo también se ha generado un script en Python en el que el usuario escoge entre múltiples opciones los valores de métricas de redes sociales de la empresa que desea implantar. Estas opciones no han sido escogidas al azar, sino que se ha realizado un análisis exploratorio en el que se ha analizado cuáles han sido los valores máximos y mínimos de cada una de las métricas y se han adaptado a esos valores. Las opciones son las siguientes:

Opciones para 'TWEETS' (tweets posteados por la empresa un mes concreto):

- (Opción 1): Entre 0 y 1.000
- (Opción 2): Entre 1.000 y 2.000
- (Opción 3): Entre 2.000 y 3.000
- (Opción 4): Entre 3.000 y 4.000
- (Opción 5): Más de 4.000

Opciones para 'VARIACION SEGUID. TWITTER' (variación mensual de seguidores de la empresa en twitter):

- (Opción 1): Entre -2.500 y 0
- (Opción 2): Entre 0 y 2.500
- (Opción 3): Entre 2.500 y 5.000
- (Opción 4): Entre 5.000 y 10.000
- (Opción 5): Entre 10.000 y 20.000
- (Opción 6): Más de 20.000

Opciones para 'LIKES FB' (número de me gustas recibidos en Facebook):

- (Opción 1): Entre 100.000 y 250.000
- (Opción 2): Entre 250.000 y 750.000
- (Opción 3): Entre 750.000 y 1.500.000
- (Opción 4): Entre 1.500.000 y 3.000.000
- (Opción 5): Más de 3.000.000

Opciones para 'TALK FB' (número de impresiones en Facebook):

- (Opción 1): Entre 1.000 y 10.000
- (Opción 2): Entre 10.000 y 50.000
- (Opción 3): Entre 50.000 y 100.000
- (Opción 4): Entre 100.000 y 200.000
- (Opción 5): Más de 200.000

Opciones para 'NUEVOS SEG IG' (número de nuevos seguidores en Instagram):

- (Opción 1): Entre 0 y 2.500
- (Opción 2): Entre 2.500 y 10.000
- (Opción 3): Entre 10.000 y 35.000
- (Opción 4): Entre 35.000 y 75.000
- (Opción 5): Más de 75.000

Opciones para 'CONTENIDO IG' (cantidad de contenido posteado por la empresa en instagram en un mes concreto):

- (Opción 1): Entre 0 y 50
- (Opción 2): Entre 50 y 100
- (Opción 3): Entre 100 y 250
- (Opción 4): Entre 250 y 400
- (Opción 5): Más de 400

Una vez conocidas las opciones, al usuario se le plantea las distintas opciones para que escoja los valores que prevé que su empresa tendrá en este mes que busca la cuota de mercado. Una vez el usuario escoge los rangos que creen que más se adecúen a su situación se utiliza el modelo de regresión creado anteriormente para predecir cuál será el valor de cuota de mercado en función de los valores que se ha escogido. Los distintos valores de las variables predictoras (como tweets, o variación de seguidores, etc.) se generan mediante un número aleatorio dentro del rango que sea escogido. Es decir, si el usuario decide que en "CONTENIDO IG" va a tener un valor entre 50 y 100, pues el valor que se utilizará para la predicción será un número aleatorio entre esos 2 valores. En el caso de que el usuario seleccione la opción de "Más de…" el valor se seleccionará aleatoriamente entre el valor máximo y 200% de su valor: por ejemplo, de nuevo en el caso de "CONTENIDO IG", si el usuario selecciona "Más de 400", el modelo predecirá utilizando un valor entre 400 y 1200.

A continuación, se muestra una ejecución utilizando los valores que se observan en la figura siguiente:

```
Elige una opción para 'TWEETS' (1-5): 5
Elige una opción para 'VARIACION SEGUID. TWITTER' (1-6): 3
Elige una opción para 'LIKES FB' (1-5): 3
Elige una opción para 'TALK FB' (1-5): 2
Elige una opción para 'NUEVOS SEG IG' (1-5): 4
Elige una opción para 'CONTENIDO IG' (1-5): 2
Nueva fila con valores seleccionados por el usuario:
{'TWEETS': 12471, 'VARIACION SEGUID. TWITTER': 4969, 'LIKES FB': 979643, 'TALK FB': 39589, 'NUEVOS SEG IG': 42063, 'CONTENIDO IG': 76}
La predicción del 'Valor Empresa' para los nuevos datos es: 19.236656620926187
```

Ilustración 28. Ejecución del regresor (Fuente: Propia)

En el ejemplo observado, se aprecia que el usuario ha seleccionado los valores: Para la variable 'TWEETS': Opción 5: Más de 4.000. Para 'VARIACION SEGUID. TWITTER': Opción 3: Entre 2.500 y 5.000. Para la variable 'LIKES FB' de nuevo ppción 3: Entre 750.000 y 1.500.000. Para 'TALK FB': Opción 2: Entre 10.000 y 50.000. Para 'NUEVOS SEG IG': Opción 4: Entre 35.000 y 75.000. Por último para 'CONTENIDO IG': Opción 2: Entre 50 y 100.

El modelo, como se ha comentado en su funcionamiento selecciona un número aleatorio dentro del rango que el usuario indica: los valores escogidos se pueden observar en la penúltima línea de la figura superior:

{'TWEETS': 12471, 'VARIACION SEGUID. TWITTER': 4969, 'LIKES FB': 979643, 'TALK FB': 39589, 'NUEVOS SEG IG': 42063, 'CONTENIDO IG': 76}

A partir de estos datos y con el modelo de regresión por vectores soporte que se ha utilizado anteriormente, el modelo genera la predicción de la cuota de mercado que esta empresa ficticia ocuparía en el caso de entrar al mercado del sector con estos datos de métricas de redes sociales.

5. CONCLUSIONES

En este capítulo se abordará cuál o cuáles han sido las conclusiones obtenidas a través del estudio, así como su relación con los estudios cursados y el legado y proyecto a futuro que puede tener este Trabajo de Fin de Grado.

5.1 Conclusiones generales

Como ya se ha visto a lo largo del estudio, en este trabajo se analiza la relación entre la popularidad online de grandes distribuidores del sector agroalimentario junto a su cuota de mercado utilizando herramientas como Google Trends o redes sociales para ver esas relaciones; y Kantar Worldpanel para analizar la cuota de mercado a lo largo del tiempo. Estas son las principales conclusiones extraídas del estudio:

En primer lugar, con respecto a la **relación entre visibilidad online y cuota de mercado** se ha observado que existe una correlación significativa entre la popularidad online (medida a través de Google Trends) y la cuota de mercado de una empresa. Las empresas con mayores cuotas de mercado, como Mercadona y Carrefour en este caso, también obtienen puntuaciones más altas en la métrica de popularidad que nos arroja Google Trends. Esto sugiere que una fuerte presencia online puede contribuir positivamente a la posición en el mercado, ya que las empresas que son más visibles y mencionadas en las plataformas digitales tienden a atraer más atención de los consumidores, lo que se traduce en mayores ventas y conocimiento de la marca.

En cuanto a la **influencia de las redes sociales**, se ha observado que tienen cierta importancia, pero que no son un factor clave. A lo largo del estudio se ha podido observar que empresas con unas métricas en redes sociales superiores a las de sus competidores no traducen esa superioridad en mejor cuota de mercado. La única red social que tiene un ligero poder en la subida o bajada de cuota de mercado de las empresas estudiadas es Twitter, aunque no tan significativo como para tenerlo en cuenta.

Por tanto, si se pudiera dividir el concepto de popularidad online en dos “subconceptos” entre los que estarían: búsquedas online de las empresas del sector (al que se asocia Google Trends) y métricas de redes sociales de cada una de las empresas, se podría concluir que es la primera razón la que justifica más esa diferencia entre cuota de mercado de las distintas empresas. Las redes sociales tienen influencia, pero ni se acercan a la influencia online.

Otro objetivo que se planteaba a la hora de afrontar el estudio era la creación de un **modelo predictivo** que, a partir de todas las variables relacionadas con la popularidad online, predijera el valor de cuota de mercado que iba a tener una empresa. Los modelos predictivos utilizados, como los bosques aleatorios (Random Forest) y los métodos de regresión, han demostrado ser eficaces para predecir la cuota de mercado basándose en datos de popularidad online.

En general, los pronósticos están muy cerca de los valores reales, lo que valida la capacidad de estos modelos para capturar la dinámica del mercado. Especialmente en el caso de Lidl, la diferencia entre los valores previstos y los reales es pequeña, ajustándose mes a mes al valor real. Esta precisión demuestra que el modelo puede adaptarse eficazmente a las tendencias de popularidad en línea y reflejarlas en las cuotas de mercado previstas. Este modelo predictivo puede llegar a ser muy útil ya que, dada la capacidad de anticipar la cuota de mercado, permite a las empresas ajustar sus estrategias de marketing y producción de manera más ajustada y adaptada a su público. Si se prevé un aumento en la cuota de mercado, una empresa puede aumentar su producción para satisfacer la demanda esperada, y de esta forma evitar problemas de desabastecimiento. O al contrario, si se prevé una disminución, la empresa puede

reducir la producción y ajustar su estrategia de marketing para intentar contrarrestar esta tendencia.

Por otro lado, este modelo predictivo también se ha utilizado para crear una especie de escenario ficticio en el que un usuario escoge los valores de métricas de redes sociales que tiene su empresa ficticia y a partir de estos valores introducidos el modelo predice qué cuota de mercado ocuparía una empresa con dichas características. Esto es realmente útil en muchas situaciones y es que, si un usuario tiene empresas del sector agroalimentario fuera de España y quiere ver, con los datos que ya tiene, como funcionaría su empresa en España, podría aplicar este modelo regresivo y se le muestra el valor de cuota de mercado. A partir de ahí, el empresario podrá decidir si hace o no la inversión acorde con la cuota de mercado que va a ocupar dentro de este nuevo escenario, ahorrando así tiempo y dinero en hacerlo "a ciegas".

En resumen, este estudio ha demostrado que la popularidad online es un indicador relevante y que se relaciona con la cuota de mercado de las empresas del sector agroalimentario. Las empresas que invierten en estrategias de marketing digital y gestión de redes sociales pueden obtener ventajas competitivas significativas en términos de popularidad y cuota de mercado. Además, la capacidad de predecir cuotas de mercado a partir de datos online proporciona una herramienta valiosa para la planificación estratégica y la toma de decisiones en el ámbito empresarial.

5.2 Relación del trabajo con los estudios cursados

Este Trabajo de Fin de Grado tiene mucha relación con toda la carga de estudio realizada durante el Grado en Ciencia de Datos en la Universidad Politécnica de Valencia (UPV). Tal y como está definido desde la propia web de la universidad: "Los datos son la base del conocimiento que tenemos del mundo: desde los movimientos de vehículos hasta las temperaturas en un hospital. La Ciencia de Datos genera profesionales capaces de crear conocimiento extraído a partir de los datos. Los/las profesionales formados/as en Ciencia de Datos son capaces de diseñar la obtención de los datos de cualquier entorno (industrial, sociológico, económico, político, empresarial, etc.), y pueden procesar, analizar y combinar datos provenientes de distintas fuentes, para extraer el conocimiento y comunicar de manera efectiva cómo gestionar la toma de decisiones estratégicas.". Dentro de este Grado se conocen y se indaga en el uso de datos para fines específicos, utilizando conocimientos desde informática hasta matemáticas, pasando por estadística o interpretación. A continuación, se detallan los aspectos más relevantes de la relación entre los estudios cursados y este trabajo:

- Estudios de Mercado y Redes Sociales

Una parte importante del trabajo se resume en el conocimiento y el estudio del mercado del sector agroalimentario, así como el estudio y la importancia de las redes sociales en la actualidad. Durante los años de estudio del Grado, en varias asignaturas se han tratado aspectos similares: En la asignatura de Comportamiento Económico y Social se estudió las bases del mercado y el funcionamiento de este, así como una aproximación a la importancia de las redes sociales. Por otra parte, en la asignatura de Economía Digital se trató en profundidad las estructuras de mercado.

- Uso de Técnicas de Análisis de Datos

Durante el desarrollo del TFG, se han aplicado varias técnicas de análisis de datos,desde modelos predictivos de regresión hasta los bosques aleatorios o Random Forest. Estas técnicas

se han aprendido y perfeccionado (además de muchas otras) en diversas asignaturas del grado,. La capacidad de predecir la cuota de mercado a partir de datos de popularidad online, utilizando estas metodologías ya conocidas, demuestra los conocimientos teóricos adquiridos y su aplicación a la práctica.

- Implementación de Herramientas y Lenguajes de Programación

Los distintos scripts de Python para acceder a los datos de las diversas fuentes, así como la programación de modelos o la visualización de gráficas son claros ejemplos de la aplicación de las habilidades de programación adquiridas en el grado. Durante los estudios se ha enseñado a programar en distintos lenguajes, aplicando la forma más eficiente de hacerlo y en estos ejemplos se ha puesto de manifiesto los estudios realizados sobre ello.

- Visualización y Presentación de Datos

Dentro de todos los aspectos tratados durante los años de estudio, un aspecto fundamental es la presentación de los datos, para que cualquier persona interesada pueda comprender los estudios, la visualización de gráficas elegidas y la forma de hacerlo han de ser lo más claras y comprensibles posibles. Hay que saber elegir la forma en la que se mostrarán los datos así como el formato (ya sea mediante gráficos, tablas, u otras opciones). En este TFG se ha plasmado los conocimientos adquiridos sobre ello y se ha buscado mostrar toda la información de la manera más clara posible.

En resumen, durante la elaboración del Trabajo de Fin de Grado se aplicaron varias técnicas de análisis de datos, como modelos predictivos (regresión y Random Forest), las cuales fueron aprendidas y perfeccionadas a lo largo de los años de estudio. Además, se desarrollaron y utilizaron scripts en Python para la recolección y procesamiento de datos provenientes de diversas fuentes, en los cuales se aplicaron las habilidades adquiridas en programación durante el grado. Además, la presentación de los resultados se llevó a cabo de manera efectiva y comprensible, utilizando gráficos y tablas seleccionados cuidadosamente para garantizar la claridad y la comprensión de los estudios realizados.

5.3 Legado

Al realizar este Trabajo de Fin de Grado, se han utilizado diversas fuentes de datos para conseguir múltiples conclusiones. El dataset utilizado junto a la obtención de datos de las diversas fuentes podrían ser de utilidad para estudios futuros similares, en donde se necesiten obtener datos e información de dichas fuentes. De esta forma, cualquier usuario que quiera realizar un estudio de popularidad dentro de cualquier sector similar al sector agroalimentario puede hacer uso de la base obtenida a lo largo de este Trabajo de Fin de Grado y utilizar tantas conclusiones y herramientas utilizadas como necesite.

5.4 Trabajo futuro

Dado que se trata de un trabajo acerca de un sector en concreto, el trabajo futuro a realizar puede ser tan extenso como se quiera. Desde la ampliación del estudio a más sectores de la sociedad hasta la incorporación de métricas de más redes sociales o más empresas al sector, se puede escalar este estudio hasta donde se quiera. Para acotar ligeramente el trabajo futuro propuesto, se centrará este punto en el trabajo futuro a proponer a este sector concreto, sin entrar en detalle en más sectores de la sociedad.

Una posible ampliación que resultaría útil para globalizar el estudio y que fuese de utilidad allá donde se necesite sería **ampliar el análisis geográfico**: obteniendo datos de otros países disponibles se podría observar si el comportamiento en el extranjero ocurre de la misma forma que ocurre en España, pudiendo extrapolar los resultados y haciendo el estudio mucho más global.

Otro futurible sería la **incorporación de nuevas fuentes de datos**. Para comenzar, se podrían obtener datos de muchas otras redes sociales minoritarias, pero que igualmente podrían tener impacto en el estudio y no se están estudiando actualmente. Las redes sociales emergentes podrían darnos la clave del comportamiento de muchas empresas y tampoco se están estudiando en este estudio, en donde se tienen en cuenta las redes sociales más utilizadas en territorio español. Además, y dado el continuo cambio y expansión del sector en sí, se podría crear una línea de análisis en donde se estudie concretamente el comercio electrónico (dentro del sector agroalimentario) para ver cuánta gente ha dejado de consumir productos del supermercado concreto a pedirlos por Internet a plataformas como eBay o Amazon. Esto enriquecerá el análisis al proporcionar una perspectiva más completa y general del mercado que se busca comprender y analizar.

Dentro de la incorporación de nuevas fuentes de datos y más en concreto dentro del análisis de redes sociales se podrían incorporar nuevas variables asociadas al **análisis de sentimientos** de comentarios en redes sociales. Cuantificar y analizar el feedback de los usuarios acerca de cada empresa y las impresiones de los mismos proporcionará nuevas variables de estudio, las cuales pueden explicar las variaciones de la cuota de mercado. Para el análisis de sentimientos se obtendrían tweets desde la API de Twitter o comentarios desde Facebook y se almacenarían en un dataset donde, aplicando diversos modelos (por ejemplo, de Hugging Face) de diversas librerías podríamos cuantificar cuántos de esos comentarios son positivos, negativos, agresivos, etc. Utilizando esta información se podría de nuevo hacer un análisis predictivo en donde las variables predictivas fueran estas (o estas sumadas a las ya conocidas del dataset utilizado) y se podría buscar entre estos datos si los comentarios de los usuarios influyen en la cuota de mercado del mes en concreto. De esta forma se respondería a preguntas como: ¿El hecho de que en un mes haya un pico de comentarios negativos acerca de la empresa Mercadona influye en que en el próximo mes la cuota de mercado de Mercadona descienda? Y viceversa, si los comentarios son positivos, ¿la cuota de mercado ascenderá? Y también en general, ¿las empresas con mayor porcentaje de comentarios positivos son las empresas que ocupan el mayor porcentaje de cuota de mercado? Estas cuestiones y muchas otras similares se pueden llegar a resolver utilizando un análisis de sentimientos de redes sociales.

En conclusión, una posible ampliación sería el análisis geográfico, obteniendo datos de otros países para observar si el comportamiento en el resto de países es similar al de España, pudiendo extrapolar el estudio y globalizando el mismo. Además, incorporar nuevas fuentes de datos, como redes sociales minoritarias y emergentes podría enriquecer el análisis. También se podría analizar el comercio electrónico para entender la migración de consumidores hacia plataformas online. Asimismo, integrar el análisis de sentimientos de comentarios en redes sociales permitiría cuantificar el feedback de los usuarios y explorar su impacto en la cuota de mercado, proporcionando una perspectiva más completa y detallada del sector agroalimentario en España.

6. Bibliografía

Ayedee, N., & Kumar, A. (2020). Social Media Tools For Business Growth of SMES. *Journal of Mananement (Conference)*.

Boyd, D., & Ellison, N. (2008). Social Network Sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 210-230.

Cebrián, E., & Domenech, J. (2023). Is Google Trends a quality data source? *Applied Economics Letters*, 811-815.

Cheung, C., & Thadani, D. (2012). The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems*, 461-470.

Dash, R., Nguyen, T., & Cengiz, K. (2023). Fine-tuned support vector regression model for stock predictions. *Neural Comput & Applic*, 23295-23309.

Fahrudin, T., Asniar, & Faizul Ula, M. (2022). Comparison The New Computer Sciences Study Programs in Indonesia using PyTrends. *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 379-383.

Fonseca, J. (2020). PyTrends (Version 4.9.1).

Ghose, A. (2019). Understanding User Behavior Using Google Trends. *Foundations and Trends in Intormation Retrieval*, 1-100.

Gong, S., Zhang, J., Zhao, P., & Jiang, X. (2017). Tweeting as a marketing tool: A field experiment in the TV industry. *Journal of Marketing Research*, 833-850.

Gonzalez, S., Garcia, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 205-237.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. *In Advances in neural information processing systems*, 2672-2680.

Green, T. (2015). *Understanding Kantar Worldpanel: a practical guide to understanding, using and maximising the value of Kantar Worldpanel's consumer panel data.*

Grover, P., Kar, A., & Dwivedi, Y. (2022). The evolution of social media influence - A literature review and research agenda. *International Journal Of Information Management Data Insight*, 100-116.

Guo, B., Wang, Y., Zhang, H., Liang, C., Feng, Y., & Hu, F. (2023). Impact of the digital economy on high-quality urban economic development: Evidence from Chinese cities. *Economic Modelling*, 120.

Hill, S., Troshani, I., & Chandrasekar, D. (2017). Signalling Effects of Vlogger Popularity on Online Consumers. *Journal of Computer Information Systems*, 76-84.

James, G., Witten, D., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. *New York: Springer*.

Kamboj, U., Paramita, G., & Mishra, S. (2022). Comparison of PLSR, MLR, SVM regression methods for determination of crude protein and carbohydrate content in stored wheat using near Infrared spectroscopy. *materialstoday*, 576-282.

Kaplan, A., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 59-68.

Kosior, K. (2018). Digital Transformation in the Agri-Food Sector – Opportunities and Challenges. *AgEcon Search*, 98-104.

Mavragani, A., Ochoa, G., & Tsagarakis, K. (2018). Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review. *JMIR Publications*.

Meire, M., Ballings, M., & Van den Poel, D. (2017). The added value of social media data in B2B customer acquisition systems: A real-life experiment. *Decision Support Systems*, 26-37.

Ministerio de Agricultura, Pesca y Alimentación, Gobierno de España. (14 de Diciembre de 2023). *Informe Anual de la Industrial Alimentaria Española*. Obtenido de https://www.mapa.gob.es/es/alimentacion/temas/industria-agroalimentaria/20240126informeanualindustria2022-20234t23ok_tcm30-659567.pdf

Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to Linear Regression Analysis.* John Wiley & Sons.

Montoriol Garriga, J. (4 de Octubre de 2019). *La industria agroalimentaria española: estructura empresarial y productividad*. Obtenido de https://www.caixabankresearch.com/es/analisis-sectorial/agroalimentario/industria-agroalimentaria-espanola-estructura-empresarial-y

Ngo, G., Beard, R., & Chandra, R. (2022). Evolutionary bagging for ensemble learning. *Neurocomputing*, 1-14.

Pourkhani, A., Abdipour, K., & Baher, B. (2019). The impact of social media in business growth and performance: A scientometrics analysis. *International Journal of Data and Network Science*, 223-244.

Qualman, E. (2019). *Socialnomics: How Social Media Transform the Way We Live and Do Business?* Wiley.

Salmond, & EdN, S. (2007). Box and Whisker Plots: Displaying Mean, Interquartile Range and Range. *Orthopaedic Nursing*, 33.

Shorten, C., & Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 1-48.

Smith, A., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook and Twitter? *Journal of Interactive Marketing*, 102-113.

Vignesh, V., Pavithra, D., Dinakaran, K., & Thirumalai, C. (2017). Data analysis using box-and-whisker plot for stationary shop analysis. *2017 International Conference on Trends in Electronics and Informatics*, 1072-1076.

WorldPanel, K. (s.f.). *Kantar WorldPanel*. Obtenido de https://www.kantarworldpanel.com/global

Xia, L., Baghaie, S., & Sajadi, S. (2024). The digital economy: Challenges and opportunities in the new era of technology and electronic communications. *Ain Shams Engineering Journal*.

Yoon, J. (2021). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. *Comput Econ*, 247-265.

Zhang, F., & O'Donnell, L. (2020). *Machine Learning* (Vol. 7).

Zou, R., & Schonlau, M. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 3-29.

7. ANEXOS

7.1 Anexo 1: Objetivos de Desarrollo Sostenible (ODS).

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

Tabla 8. Objetivos de Desarrollo Sostenible

Objetivo	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		

En cuanto al impacto en los Objetivos de Desarrollo Sostenible, se ha considerado que los escogidos en la tabla superior son los que más impacto tienen en el desarrollo de este Trabajo de Fin de Grado. Se va a desarrollar los ODS que se consideran más importantes en relación con este TFG:

- **ODS 8. Trabajo decente y crecimiento económico**: Este TFG se centra en el supuesto de la posible relación entre el auge de redes sociales y el crecimiento en popularidad online y la cuota de mercado que una empresa ocupa. Por tanto, las conclusiones sacadas a lo largo del mismo afectan en gran medida en esta ODS, ya que estas conclusiones se pueden utilizar por empresas para invertir más o menos en ciertos departamentos y personal, lo que puede generar puestos de trabajo y mejorar las condiciones laborales de los empleados.
- **ODS 9. Industria, innovación e infraestructuras**. Utilizando tecnologías avanzadas se ha podido valorar relaciones entre métricas de popularidad online y cuota de mercado de las empresas, lo que supone un avance innovador y que, en un trabajo futuro, podría

extenderse a muchas ramas y ser una herramienta de gran utilidad. Esto mismo fomenta el desarrollo de infraestructuras más avanzadas con la finalidad de conseguirlo.

- **ODS 12. Producción y consumo responsables**. Al conocer una previsión o como va a variar la cuota de mercado para una empresa concreta del sector, se puede conseguir no generar excedente o hacer corto de producto, lo que reduciría el gasto de la empresa en producto y, a mayores, reduciría el desperdicio de material (en este caso alimenticio) evitando sobreproducción. Esta conclusión también podría asemejarse con el **ODS 13: Acción por el clima**; pues la no sobreproducción de productos alimentarios haría reducir emisiones y menor variación de los ecosistemas, optimizando así tanto la parte económica de la empresa como la parte ecológica del planeta.

8.1 Anexo 2: Pytrends.

Durante el desarrollo de este Trabajo Final de Grado se ha utilizado la API no oficial de Pytrends para automatizar el acceso a los datos de Google Trends utilizando Python. Pytrends proporciona una interfaz sencilla para interactuar con Google Trends, lo que permite descargar informes y extraer información relevante para analizar las tendencias de búsqueda en línea.

Google Trends es una herramienta valiosa para comprender el interés del público en determinados temas a lo largo del tiempo y en diferentes regiones geográficas. Sin embargo, estos datos no son accesibles de forma directa. Aquí es donde entra en juego Pytrends, que proporciona una interfaz de programación sencilla para interactuar con los datos de Google Trends y descargarlos al gusto del usuario.

Pytrends tiene múltiples usos a la hora de la elaboración del TFG. En el caso de este reporte, simplemente se ha utilizado la zona geográfica que hemos acotado a España; y el filtro de interest_over_time, que permite extraer los datos de búsqueda de ciertos términos a lo largo de un periodo. Pero además de estas utilidades que se han trabajado a lo largo de la extracción de datos, también se tiene otros muchos servicios a los que se podría acceder:

- Análisis Geográfico: Se ha utilizado Pytrends para filtrar las búsquedas dentro de España, pero también se podría analizar el interés por regiones geográficas, buscando información valiosa sobre la distribución geográfica del interés en determinados temas.
- Exploración de temas relacionados: Pytrends facilita la exploración de temas relacionados y consultas asociadas a un término de búsqueda específico, lo que serviría para extender el análisis y comprender las tendencias de búsqueda.
- Seguimiento de Búsquedas Tendenciales: En relación a la exploración de temas similares; Pytrends permite acceder a las últimas búsquedas tendenciales, lo que permitiría, en caso de necesitarlo, estar al tanto de los temas populares en tiempo real.

Pytrends es una estrategia efectiva para acceder y analizar datos de Google Trends de manera automatizada. Su programación y su utilización son fáciles de usar para quien está familiarizado con el entorno de la programación. Esto es muy interesante, ya que permite al usuario centrarse en el análisis de datos en lugar de preocuparse por la extracción de datos. Pero hay que tener en cuenta también sus limitaciones y es que no es una API oficial de Google Trends, por lo que puede estar sujeta a cambios. Además, su uso excesivo puede ser limitado. Para futuros trabajos, es recomendable monitorear las actualizaciones de Pytrends y estar al tanto de cualquier cambio en las políticas o restricciones de uso de Google Trends para garantizar la integridad y la eficacia de la investigación. (Fonseca, 2020)

8.2 Anexo 3: Análisis completo por región de Google Trends.

Gracias a Google Trends puede realizarse un estudio pormenorizado centrado en las regiones de España. En el caso de este proyecto, no se ha realizado este estudio, pues los datos extraídos por región no eran de utilidad para el estudio, debido a que los valores de cuota de mercado no estaban desagregados por región. Por esa razón, en este anexo se indagará dentro de cuál es el supermercado preferido de cada una de las Comunidades Autónomas y se compararán buscando cuáles de estas comunidades tienen unas búsquedas similares o, si al contrario, hay diferencias entre alguna de ellas. Este estudio puede realizarse a través de la propia interfaz de Google Trends o conectándose a Pytrends, como se ha hecho anteriormente.

Desde la interfaz de Google Trends se observa de forma muy intuitiva cuál es el supermercado favorito de cada CCAA. Como se ha observado a lo largo del estudio, los dos supermercados favoritos por los españoles son Carrefour y Mercadona. Ahora se puede observar de manera desagregada por región cuál ha sido el supermercado favorito de cada una durante este último año.

Tabla 9. Puntuaciones asignadas a grandes supermercados por CCAA (Fuente: Propia, con datos de Google Trends)

geoName	mercadona	carrefour	lidl	eroski	dia supermercado
Andalucía	44	36	18	1	1
Aragón	48	28	18	6	0
Canarias	54	22	24	0	0
Cantabria	31	45	19	5	0
Castilla y León	40	38	19	2	1
Castilla-La Mancha	46	32	20	1	1
Cataluña	47	30	22	1	0
Ceuta	41	42	14	3	0
Comunidad Valenciana	45	38	17	0	0
Comunidad de Madrid	45	38	15	1	1
Extremadura	37	46	16	0	1
Galicia	34	30	21	15	0
Islas Baleares	40	23	22	15	0
La Rioja	45	26	18	11	0
Melilla	51	25	24	0	0
Navarra	35	22	17	26	0
País Vasco	31	20	17	32	0
Principado de Asturias	39	40	19	2	0
Región de Murcia	48	34	17	0	1

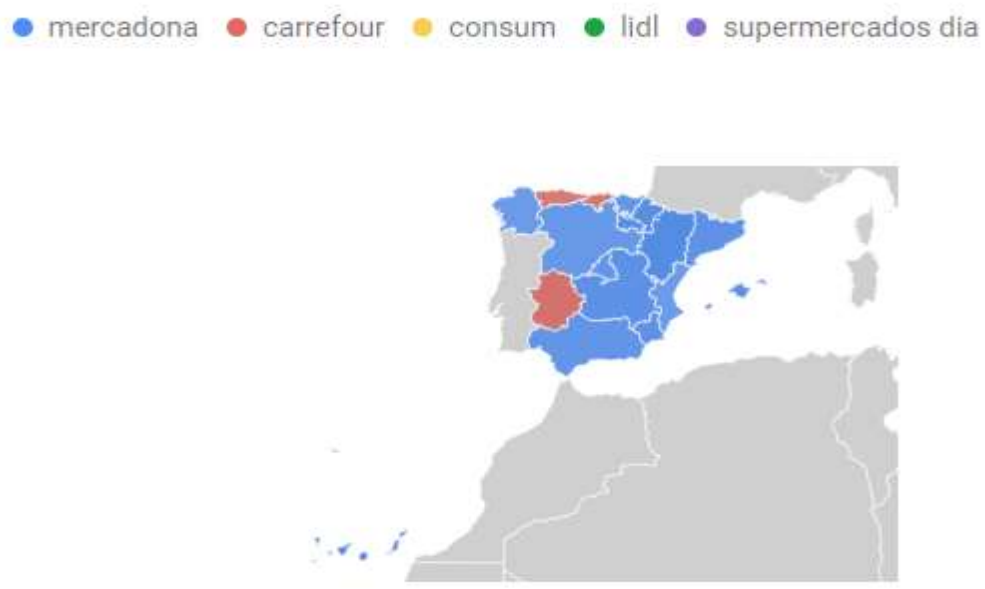


Ilustración 29. Desglose comparativo por regiones (Fuente: Google Trends)

Se observa una grandísima preferencia de los españoles por el supermercado Mercadona, mientras que solamente 3 CCAA (Extremadura, Asturias y Cantabria) tienen más búsquedas de Carrefour. Otro de los KPIs que ofrece Google Trends es el top N regiones que realizan más búsquedas de un supermercado. Además, se puede situar el cursor sobre una de las regiones y te mostrará de forma completa el porcentaje de búsquedas de cada uno de los términos en esa región durante este último año, como se observa a continuación:

Ilustración 30. Intereses por supermercados (descendiente por el interés en "Mercadona") (Fuente: Google Trends)

También la herramienta muestra las tendencias de búsqueda relacionadas con cualquier término, por ejemplo, para el caso de mercadona encuentra como tendencias:

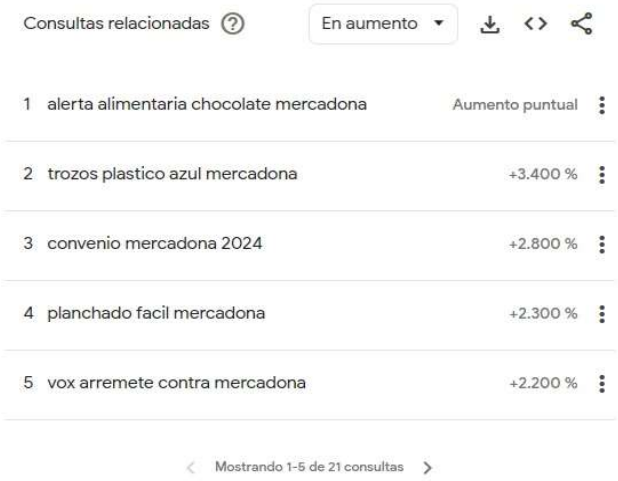


Ilustración 31. Consultas relacionadas con la búsqueda inicial (Fuente: Google Trends)

Otra curiosidad y que puede venir bien para estudios futuros es notar como Google Trends indica claramente la presencia de supermercados regionales (como Consum en la zona del Levante o HiperDino en las Islas Canarias). Introduciendo el término de búsqueda de esos supermercados denota con gran claridad la presencia de búsquedas en sus zonas con mayor afluencia de público, y prácticamente nulas en el resto de regiones (sobre todo para el caso de Hiperdino).

Ilustración 32. Interés por Supermercados Consum (Fuente: Google Trends)

Ilustración 33. Interés por Hiperdino (Fuente: Google Trends)

8.3 Anexo 4: Funcionamiento del diagrama de caja (Box & Whisker diagram).

El diagrama de caja y bigotes, también conocido como diagrama de caja, es una herramienta gráfica utilizada en múltiples campos de la estadística para representar la distribución de un conjunto de datos de forma visual. Este diagrama proporciona información sobre la mediana, los cuartiles, los valores atípicos y la dispersión de los datos.

Componentes del Diagrama:

1. **Caja (Box):** La caja en el centro del diagrama representa el rango intercuartílico (IQR: Interquartile Range en inglés), que abarca el 50% central de los datos. La parte inferior de la caja marca el primer cuartil (Q1) y la parte superior marca el tercer cuartil (Q3). La longitud de la caja indica la dispersión de los datos dentro de este rango intercuartílico.
2. **Bigotes (Whiskers):** Los bigotes se extienden desde la caja y representan la variabilidad fuera del rango intercuartílico. La longitud de los bigotes se extiende hasta 1.5 veces el rango intercuartílico desde los cuartiles Q1 y Q3. Por otra parte, si existen valores atípicos, se muestran como puntos individuales más arriba o debajo de los bigotes.
3. **Mediana**: Una línea dentro de la caja indica la mediana de los datos, que es el punto medio del conjunto de datos cuando se ordenan de menor a mayor.

Para entender bien lo que se puede aprovechar de estos gráficos, se debe conocer de la mejor forma posible su interpretación:

- La posición de la mediana en el diagrama indica la tendencia central de los datos.
- La longitud de la caja proporciona una medida de la dispersión de los datos dentro del rango intercuartílico. A mayor longitud de caja, mayor dispersión, y al contrario.
- La longitud de los bigotes sugiere la variabilidad de los datos más allá del rango intercuartílico.

Estos gráficos tienen una gran cantidad de usos dentro de la estadística descriptiva y el análisis de datos. Entre ellos se pueden destacar la comparación entre la distribución de diferentes grupos de datos. Además, son de gran utilidad para identificar la presencia de valores atípicos en un conjunto de datos, por su sencilla interpretación. También son útiles para visualizar la dispersión y la tendencia central de los datos de manera concisa y efectiva. Por otra parte hay que tener presentes también sus limitaciones o desventajas como que estos diagramas son realmente útiles para datos numéricos y simétricos, pero normalmente no son apropiados para datos categóricos o sesgados. Además hay que tener cierta precaución con la interpretación de los valores atípicos considerando el contexto del problema. (Salmond & EdN, 2007)

8.4 Anexo 5: Bagging.

Bagging es el método que se utiliza en el algoritmo de Random Forest. En este anexo, se conocerá más en profundidad cuál es el funcionamiento, así como sus ventajas e inconvenientes.

El bagging es un método de aprendizaje por conjuntos diseñado para reducir la varianza en conjuntos de datos ruidosos. Este método implica tres pasos clave:

1. **Bootstrapping**: En el primer paso, se crean múltiples muestras del conjunto de datos de entrenamiento mediante muestreo aleatorio con reemplazo, lo que permite múltiples selecciones de puntos de datos.
2. **Entrenamiento Paralelo**: Estas muestras generadas se entrenan de manera independiente utilizando modelos débiles (para evitar sobre ajustar el modelo).
3. **Agregación**: Este es el último paso del bagging: las predicciones de los modelos entrenados se combinan y se utiliza el promedio para regresión o la mayoría de votos para clasificación, con el objetivo primordial de mejorar la precisión de la estimación final.

El concepto de aprendizaje por conjuntos en el que se basa el bagging se basa en la "sabiduría de las multitudes", en donde múltiples modelos colaboran para mejorar las predicciones. Los métodos de aprendizaje por conjuntos como bagging ayudan a equilibrar el sesgo y la varianza, e impiden o al menos contrarrestan problemas de overfitting o underfitting presentes en modelos individuales.

El Bagging ofrece ventajas como la facilidad de implementación, ya que utilizando herramientas de programación como scikit-learn se facilita la combinación de predicciones para mejorar el rendimiento del modelo. Además, como principal ventaja es la capacidad reducción de varianza, la cual es especialmente útil para datos de alta dimensión, ya que gracias a esta reducción de varianza, se disminuye la probabilidad de sobreajuste y mejora la generalización.

Por otra parte también encuentra inconvenientes como la pérdida de interpretabilidad: debido a que al agregar predicciones se puede dificultar la extracción de información precisa para negocios. Además tiene un elevado coste computacional: requiere considerable capacidad de procesamiento, lo que puede ralentizar el proceso hasta niveles, en ocasiones, extremos. Por último, es menos flexible y funciona mejor con algoritmos menos estables; por lo que no ofrece tanto beneficio con modelos altamente estables o sesgados. (Gonzalez, Garcia, Del Ser, Rokach, & Herrera, 2020) (Ngo, Beard, & Chandra, 2022)

8.5 Anexo 6: Variaciones en la cuota de mercado y el resto de las variables para Grupo Dia, Grupo Eroski y Lidl.

- **GRUPO DÍA:**

Para ”Grupo Dia” se tiene un total de 14 variaciones positivas y 28 variaciones negativas. Las tablas de resumen son las siguientes:

	Positivos	Negativos	% Positivos
TWEETS	11	3	78.57%
VARIACION SEGUID. TWITTER	8	6	57.14%
LIKES FB	12	2	85.71%
TALK FB	6	8	42.86%
NUEVOS SEG IG	10	4	71.43%
CONTENIDO IG	8	4	57.14%
Puntuacion Trends	5	5	35.71%

	Positivos	Negativos	% Negativos
TWEETS	13	15	53.57%
VARIACION SEGUID. TWITTER	12	16	57.14%
LIKES FB	23	5	17.86%
TALK FB	15	13	46.43%
NUEVOS SEG IG	10	18	64.29%
CONTENIDO IG	14	14	50.00%
Puntuacion Trends	8	13	46.43%

Ilustración 34. Tablas variaciones de métricas con respecto a variación de cuota de mercado para Grupo Día (Fuente: Propia)

- **GRUPO EROSKI**:

Para ”Grupo Eroski” se tiene un total de 11 variaciones positivas y 21 variaciones negativas. Las tablas de resumen son las siguientes:

	Positivos	Negativos	% Positivos
TWEETS	4	7	36.36%
VARIACION SEGUID. TWITTER	3	8	27.27%
LIKES FB	11	0	100.00%
TALK FB	4	7	36.36%
NUEVOS SEG IG	6	5	54.55%
CONTENIDO IG	3	6	27.27%
Puntuacion Trends	4	6	36.36%

	Positivos	Negativos	% Negativos
TWEETS	11	10	47.62%
VARIACION SEGUID. TWITTER	11	10	47.62%
LIKES FB	21	0	0.00%
TALK FB	11	10	47.62%
NUEVOS SEG IG	10	11	52.38%
CONTENIDO IG	9	9	42.86%
Puntuacion Trends	13	8	38.10%

Ilustración 35. Tablas variaciones de métricas con respecto a variación de cuota de mercado para Grupo Eroski (Fuente: Propia)

- **LIDL**:

Para ”Lidl” se tiene un total de 21 variaciones positivas y 22 variaciones negativas. Las tablas de resumen son las siguientes:

	Positivos	Negativos	% Positivos
TWEETS	11	10	52.38%
VARIACION SEGUID. TWITTER	8	13	38.10%
LIKES FB	21	0	100.00%
TALK FB	13	8	61.90%
NUEVOS SEG IG	6	15	28.57%
CONTENIDO IG	12	8	57.14%
Puntuacion Trends	12	9	57.14%

	Positivos	Negativos	% Negativos
TWEETS	9	13	59.09%
VARIACION SEGUID. TWITTER	13	9	40.91%
LIKES FB	22	0	0.00%
TALK FB	12	10	45.45%
NUEVOS SEG IG	12	10	45.45%
CONTENIDO IG	7	12	54.55%
Puntuacion Trends	7	14	63.64%

Ilustración 36. Tablas variaciones de métricas con respecto a variación de cuota de mercado para Lidl (Fuente: Propia)