DPi Departamento de Proyectos de Ingeniería

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

**DEPARTAMENTO DE PROYECTOS DE INGENIERÍA**

## MÁSTER UNIVERSITARIO EN COOPERACIÓN AL DESARROLLO

### GESTIÓN DE PROYECTOS Y PROCESOS DE DESARROLLO

# TRABAJO FIN DE MÁSTER

# ANÁLISIS DE LA RELACIÓN ENTRE EL CONSUMO ALIMENTARIO Y LA SALUD EN LA POBLACIÓN ESPAÑOLA MEDIANTE DATA SCIENCE

**AUTOR/A:**

HERNÁNDEZ BALINOT, ROSARIO

**DIRECTOR/A:**

ÁLVARO FERNÁNDEZ-BALDOR MARTÍNEZ

Fecha: 10/09/2021

Número de palabras: 14181

# ÍNDICE GENERAL

## ÍNDICE DE TABLAS



## ÍNDICE DE IMÁGENES

## RESUMEN

En una sociedad en la que los datos creados y almacenados sobre, prácticamente, cualquier actividad realizada por la población crecen de forma exponencial, el análisis de datos, concretamente de cantidades masivas de datos a través de técnicas de data science, se ha convertido en una de las grandes herramientas con la que cuentan las grandes empresas para aumentar sus ganancias. Sin embargo, no solo estas grandes empresas son las que pueden beneficiarse de las diferentes técnicas de data science.

Por otro lado, con una población en la que aumenta el interés por conocer diferentes pautas para cuidar mejor de su salud, especialmente a la hora de alimentarse, se ve necesario que estas personas puedan tener información clara y segura que responda a sus preguntas sobre alimentación.

El presente trabajo, realizado como proyecto de prácticas del máster en Cooperación al Desarrollo, pretende unir ambas situaciones. Así, mientras aprovecha los datos ya recolectados y almacenados por parte de una gran empresa de análisis de datos, Nielsen, busca información que pueda ayudar a la población a la hora de tomar decisiones sobre su alimentación.

Para el desarrollo de este proyecto, se han empleado dos bases de datos, una primera con información del gasto en alimentación de los españoles, cedida por Nielsen y, otra extraída del Instituto Nacional de Estadística (INE), con datos sobre el porcentaje de casos de enfermedades crónicas (colesterol, hipertensión, diabetes y obesidad) en España.

Para llevar a cabo el análisis, se han seguido diferentes etapas de trabajo que incluyen el estudio de proyectos similares y búsqueda de la base de datos de enfermedades, el análisis separado de cada base de datos, el análisis conjunto de las mismas y, finalmente, el análisis de los resultados obtenidos. La fase de análisis de datos conjunta se ha realizado con el empleo de técnicas de correlación lineal y de regresión lineal, gracias a las cuales se han podido determinar aquellas categorías de alimentos que guardan una relación significativa, positiva o negativa con las diferentes enfermedades crónicas analizadas.

Se han encontrado diferentes limitaciones a la hora de desarrollar el análisis, principalmente la escasez de muestra con la que se ha trabajado. Sin embargo, los resultados, que corroboran la opinión de los expertos sobre la influencia de la alimentación en el desarrollo de enfermedades crónicas, demuestran que es posible utilizar estas técnicas para llegar a resultados válidos y se ven como un primer acercamiento a esta problemática. Finalmente, se exponen una serie de conclusiones y recomendaciones de cara a futuros proyectos que ahonden en esta problemática.

## INTRODUCCIÓN

En este documento se va a explicar el proceso seguido para la realización de un proyecto de análisis de datos que busca conocer la posible relación entre el consumo de determinados alimentos y su relación con diferentes enfermedades crónicas. Este proyecto ha sido realizado como parte del trabajo llevado a cabo por la ONG So Good Data y como proyecto de prácticas del máster en Cooperación al Desarrollo.

So Good Data es una ONG fundada en 2019 con el objetivo de aplicar los conocimientos de los profesionales de análisis de datos a problemas que preocupen a diferentes sectores de la sociedad. Uno de los principales objetivos con los que nació esta ONG es poner al alcance de la sociedad general el conocimiento obtenido a través de técnicas de Big Data. Los temas en los que se centra su trabajo son principalmente la contaminación, la salud, la nutrición, el transporte colaborativo, la igualdad de género y la concienciación ciudadana sobre la inteligencia artificial. La forma de trabajo que mantiene So Good Data es la colaboración entre un experto en análisis de datos con interés en un tema y, diferentes empresas o entidades que aporten datos y conocimiento sobre dicho tema.

Para este proyecto, han colaborado la empresa Nielsen (empresa dedicada a medición de compras y consumo a nivel mundial), cediendo los datos de consumo alimenticio, la alumna en prácticas, desarrollando el proyecto, y dos analistas de datos guiando y ayudando en el desarrollo del mismo. Además, se ha consultado a diferentes especialistas de nutrición especialmente en la fase de diseño del proyecto para que pudiesen ofrecer su visión y responder ciertas dudas como desde qué perspectiva realizar el análisis o en qué tipo de datos centrarlo.

La selección del tema para desarrollar el proyecto surge debido al interés mostrado tanto, por parte de la alumna como de la ONG, en estudiar la relación existente entre la alimentación y la salud. Viendo la gran variedad de alimentos que se pueden encontrar en el supermercado y, especialmente, viendo la diversidad que hay en cuanto a características de un producto (ecológico, sin lactosa/gluten, light…) surgió el interés en poder comprobar si realmente la selección de uno u otro producto tenía alguna relación directa con la salud del consumidor. Sin embargo, el proyecto ha ido cambiando, dependiendo principalmente de los datos a los que se ha tenido acceso.

El proyecto llevado a cabo, consiste en estudiar la relación existente entre el gasto que destina la población española al consumo de diferentes productos alimenticios y el porcentaje de enfermedades crónicas de esta población. De este modo, se pretende obtener información que ayude a las personas a aumentar el conocimiento que tienen sobre la influencia de la alimentación en la salud, así como ser base de futuros proyectos que profundicen más y puedan obtener conclusiones más concretas.

El desarrollo de este proyecto ha constado de una primera etapa de diseño y búsqueda de bases de datos. El objetivo de esta primera etapa era analizar la situación actual de investigaciones en este campo y encontrar fuentes de datos que se pudiesen analizar junto a los datos cedidos por Nielsen para obtener información relevante. Posteriormente, se comenzó el análisis de datos, dividido en una primera etapa de análisis de cada base de datos por separado, y, a continuación, el análisis de la relación entre ambas. Finalmente, se estudiaron los diferentes resultados que se habían ido obteniendo con el fin de poder hacer una publicación en la que mostrar este conocimiento tanto a la sociedad como a Nielsen. Además, se prepararon una serie de conclusiones sobre cómo se podrían

dar los siguientes pasos en proyectos en este ámbito y se determinaron también las principales limitaciones que habían influido en el desarrollo del proyecto.

El documento se estructura de forma que, primero se presenta la información que ha servido para plantear y dar forma al proyecto y, posteriormente se pasa a explicar y analizar el mismo. Primero se presentan dos apartados dedicados a exponer los antecedentes y la problemática general que ha podido ser analizada y que sirve como justificación para el desarrollo del proyecto. Posteriormente, se presenta otro apartado destinado a la explicación de la metodología que ha sido utilizada. A continuación, se desarrolla un apartado en el que se explica cada una de las fases que ha seguido el proyecto, así como los resultados obtenidos en el mismo. Finalmente, se han destinado dos apartados para plasmar las principales conclusiones extraídas de este proyecto, así como al desarrollo de un breve análisis crítico de todo el proceso.

## ANTECEDENTES Y CONTEXTO DE LA INTERVENCIÓN/PROYECTO

A lo largo de este apartado, se desarrollan los principales hechos que sirven como base y justificación para la creación de este proyecto. Como se explicaba en el apartado anterior, la idea de realizar este análisis nace en un contexto muy actual, como respuesta a la creciente cantidad de datos producidos por parte de la sociedad y el gran potencial de estos para dar respuestas eficaces a diferentes problemáticas sociales (Lytras, M y Visvizi, A., 2019).

Antes de centrar el documento en el problema concreto sobre el que se ha trabajado, se va a dar a conocer el contexto actual tanto en alimentación y enfermedades crónicas, como en la producción y gestión de grandes cantidades de datos. En primer lugar, se va a explicar qué son las enfermedades crónicas, cómo se están desarrollando en la sociedad española y la importancia de actuar sobre ellas. A continuación, se analiza cómo es la alimentación de los españoles y las principales tendencias en este ámbito. Finalmente, se explica qué es el Big Data y cuál es su potencial para dar respuestas no solo a empresas sino también a la población.

### Enfermedades crónicas

Según el Instituto Nacional del Cáncer (Instituto Nacional del Cáncer, s.f.), una enfermedad crónica o enfermedad no transmisible (ENT) es aquella que, por lo general, dura 3 meses o más, y es posible que empeore con el tiempo. Normalmente, se presentan en adultos y a menudo se controlan, pero no se curan.

Las enfermedades crónicas se han posicionado como uno de los principales factores de riesgo para la salud a nivel mundial (Organización Mundial de la Salud, 2021). Las ENT matan a 41 millones de personas cada año, lo que equivale al 71% de las muertes que se producen en el mundo. Las principales enfermedades que causan estas muertes son las enfermedades cardiovasculares (17.9 millones de muertes cada año), el cáncer (9 millones), las enfermedades respiratorias (3.9 millones) y la diabetes (1.6 millones) (Organización Mundial de la Salud, 2021).

Entre la población española, los tumores malignos, las enfermedades del corazón, las enfermedades cerebrovasculares, las enfermedades crónicas de las vías respiratorias bajas, el Alzheimer y la diabetes mellitus, eran ya en 2016, las primeras causas de muerte, con más del 61% de las defunciones según Mayoral Cortes, J. M., Aragonés Sanz, N. et al. (2016).

Además, no se trata solo de un problema que afecte a la mortalidad de la población. Estas enfermedades no suponen la muerte inmediata de las personas, pero sí que suponen un mal estado de su salud y una pérdida en la calidad de vida de las personas que las sufren.

A pesar de que esta información se conoce desde hace más de una década, estas enfermedades crónicas no dejan de aumentar en el mundo y en España. Un reciente artículo de The Lancet [bibliografía], afirmaba que el aumento de estas enfermedades crónicas podría llevar incluso a un punto de inflexión con respecto a la esperanza de vida.

Los factores de riesgo que más influyen en la aparición de estas enfermedades se pueden dividir en factores de comportamiento y factores metabólicos. Entre los factores de riesgo comportamentales se destacan el consumo de tabaco, la inactividad física y el uso nocivo del alcohol y las dietas malsanas. Entre los factores de riego metabólicos destacan el aumento de la tensión arterial, el sobrepeso y la obesidad, la hiperglucemia y la hiperlipidemia (Organización Mundial de la Salud, 2021).

Estas enfermedades tienen un alto impacto socioeconómico. La OMS (2021) alerta de que las enfermedades crónicas ponen en peligro el avance hacia la consecución de los Objetivos de Desarrollo Sostenible (ODS), entre los que se encuentra la reducción de las muertes prematuras por enfermedades crónicas en un 33% para 2030.

Al ser enfermedades de largo recorrido, generan un alto impacto en la economía de los diferentes países. En entornos con pocos recursos dificultan la lucha contra la pobreza ya que las familias tienen que asumir un alto gasto para hacer frente a estas enfermedades (Organización Mundial de la Salud, 2021). En España, según datos de Redacción Médica (2018), en 2017 el 40% del gasto de las Comunidades Autónomas se destinó a la salud. De este gasto sanitario, el 80% se destinó a 4 enfermedades, la hipertensión, la diabetes, la EPOC y la insuficiencia cardiaca.

Para hacer frente a estas enfermedades es muy importante centrar las actuaciones en la prevención y el diagnóstico temprano de las mismas (Organización Mundial de la Salud, 2002). En España, se cuenta con sistemas de vigilancia de la salud, que engloban la detección y seguimiento de problemas o determinantes de la salud de la población (Mayoral Cortes, J. M., Aragonés Sanz, N. et al., 2016).

Y, a pesar de que entre las áreas a vigilar se encuentran las enfermedades crónicas, no existe un sistema nacional para la vigilancia de estas enfermedades, solo algunas iniciativas a nivel de las comunidades autónomas. Esto significa, que no se cuenta con un sistema que recoja los indicadores necesarios para conocer la magnitud de estas enfermedades o los posibles cambios de comportamiento de las mismas, algo necesario para actuar en la prevención de las mismas. Poder desarrollar este sistema de vigilancia para las enfermedades crónicas es complicado debido, entre otros hechos, a los múltiples factores que influyen en su desarrollo y al gran volumen de datos con el que es necesario trabajar (Mayoral Cortes, J. M., Aragonés Sanz, N. et al, 2016).

Puesto que muchos de los factores que aumentan el riesgo de padecer una enfermedad crónica están relacionados con los hábitos de vida y la alimentación, la concienciación sobre estos hábitos debe de ser también uno de los focos en los que se centren las medidas de prevención (Organización Mundial de la Salud, 2002). En este sentido, los medios juegan un papel determinante. Los medios informativos suelen buscar episodios sensacionales o trágicos y, en cuanto a salud, se centran en las enfermedades más temidas como el VIH/SIDA. Sin embargo, no se suele prestar atención a riesgos

sanitarios más comunes crónicos y pequeños, como la exposición pasiva al humo del tabaco, el sedentarismo o la importancia de la dieta seguida (Organización Mundial de la Salud, 2002)

### Alimentación de la población española

Como se veía anteriormente, uno de los factores de riesgo debido a hábitos comportamentales era llevar una alimentación insana. Además, el aumento de la tensión arterial, la hiperglucemia, la hiperlipidemia, el sobrepeso y la obesidad, todos ellos incluidos entre los factores de riesgo metabólicos, están altamente determinados por la alimentación de la persona. Es por ello por lo que, a continuación, se van a desarrollar las principales tendencias alimenticias entre la población española.

Según un informe de la Series de Informes Técnicos de la OMS (Organización Mundial de la Salud, 2003), los cambios en la economía alimentaria a nivel mundial se han traducido en cambios en los hábitos alimenticios (más consumo de alimentos con alto nivel energético, grasas saturadas…). Estos, unidos a un aumento del modo de vida sedentario y la disminución de actividades que exijan esfuerzo físico ha propiciado un aumento de las enfermedades crónicas como la diabetes, la hipertensión e incluso algunos tipos de cáncer.

La dieta de los españoles se ha identificado tradicionalmente con la dieta mediterránea, la cual es una dieta equilibrada y saludable y que se relaciona con una menor incidencia de algunas enfermedades crónicas (bibliografía UCM). Cada año, se hacen diferentes estudios que tienen como objetivo analizar cómo es el consumo alimenticio de la población española. Para este TFM se van a mostrar las conclusiones obtenidas por Ayuso, M. (2015) a cerca del informe de la Fundación Mapfre, “Alimentación y sociedad en la España del siglo XXI” (Varela Moreiras, G., Serrano Iglesias, M. et al., 2015).

A pesar de que en esta publicación se hace referencia a distintos hábitos de los españoles en cuanto a la alimentación (horario, número de comidas al día, composición…), en este trabajo solo se hará referencia a los alimentos consumidos, puesto que es lo que se analizará posteriormente. Se aprecia que, en todas las zonas geográficas de España, la pasta es el alimento preferido, solo superado por el arroz en la zona del Levante. La ensalada, la fruta y las legumbres se encuentran en los puestos más bajos. En cuanto al pescado, se come más en el noroeste, mientras que las verduras y hortalizas se consumen más en el norte. Finalmente, algo que se destaca en la publicación, es el hecho de que hay una relación directa entre las personas que muestran mayor preocupación por su alimentación y la proporción de personas con sobrepeso.

En cuanto a tendencias que sigue la población española en alimentación, se van a hacer referencia a las conclusiones aportadas por Miralles Navarro, M. (2020). En este documento se enmarcan cuatro tendencias generales: La primera tendencia es el “Consumidorcentrismo”, que hace referencia a un consumidor que quiere que lo cuiden y lo sorprendan. Además, busca cercanía, personalización, productos naturales y locales. La segunda tendencia es la responsabilidad en el consumo. Con un 23% de los consumidores priorizan la alimentación sostenible, las marcas están aumentando sus opciones ecológicas, con envases sostenibles y formas de producción diversificadas. Se podría decir que es una tendencia hacia el residuo cero. La tercera de las tendencias destacadas es el hedonismo y la salud. Aumenta la idea de comer de manera funcional, es decir, con un objetivo que puede ser sentirse mejor, aumentar el rendimiento, ir al baño… Por ello, en los próximos años aumentarán los alimentos fermentados, los probióticos, los azúcares de origen natural, los alimentos sin gluten y sin lactosa o

para paliar carencias de hierro. Además, se harán cada vez más dietas personalizadas. La cuarta y última tendencia es más vegetales, menos proteínas animales. EL veganismo es una tendencia seguida ya por buena parte de la población. Las motivaciones para seguirla son: salud 31%, fitness 12%, perder peso 4%, sostenibilidad 2%, derechos de los animales 2%, no transgénico 1% y limpieza 1%. Esta tendencia hará que se introduzcan o aumente el consumo de diferentes tipos de alimentos como levaduras, setas, algas, legumbres, frijol, semillas de cáñamo, diferentes tipos de harinas, etc.

Según un estudio realizado por Google en colaboración con Kantar Worldpanel y Lantern (Echanove, M y Puértolas, S., 2019), un 74% de los hogares españoles cree que es importante seguir una dieta sana. Además, el mismo estudio indica que un porcentaje significativo de los españoles pide cada vez más alimentos BIO, orgánicos, veggie…En cuanto al perfil de estas personas más preocupadas por la alimentación, el estudio señala a las mujeres que no son madres de entre 18 y 35 año como las que más interés tienen en la alimentación. Finalmente, otro dato que podemos extraer del artículo es el aumento en las búsquedas por móvil sobre alimentación.

### Creación de datos y Data Science

Los datos y, especialmente, su análisis se ha convertido en una de las herramientas más valoradas por las empresas y por diferentes centros de investigación. Hasta 2003 se habían generado 5 exabytes (ver equivalencias en la tabla de equivalencias del anexo) de información a lo largo de toda la historia. En 2011, ya se habían generado unos 600 exabytes (Fundación Mapfre, s.f). En la actualidad, la información generada crece de una manera exponencial. Las redes sociales son una de las grandes fuentes de datos, pero no son las únicas. La llamada huella digital hace referencia a todos los registros y rastros que dejamos cuando utilizamos internet (Ambit, 2019). Según Yubero, B. (2021), en el 2020, por cada persona en el mundo se produjeron 1.7 MB de datos cada segundo y, se prevé que cada año se duplique la cantidad de datos producidos en el año anterior.

Los términos Big Data y Data Science son cada vez más conocidos y usados especialmente en el ámbito empresarial, donde se utilizan, entre otros, con el fin de aumentar los beneficios. Por un lado, según una de las definiciones de (Maté, C., 2014), el Big Data se puede definir como “activos de información caracterizados por su volumen elevado, velocidad elevada y alta variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y la toma de decisiones en las organizaciones.”

Sin duda, el Data Science ha sido una revolución para los negocios por sus múltiples aplicaciones como, por ejemplo, en segmentación de mercados, captación de nuevos clientes, fidelización de clientes, detección de futuras necesidades del cliente, optimización de rutas de reparto… A cualquier empresa que quiera seguir creciendo, se le recomienda seguir un plan de captación de datos para poder comenzar con un proyecto de Data Science para su empresa (Universidad de Alcalá, s.f.).

Sin embargo, no solo las empresas son las que pueden beneficiarse de los datos, sino que prácticamente en cualquier campo de trabajo es posible aplicar estas técnicas para mejorar los resultados y optimizar el trabajo. Por ejemplo, en el campo de la medicina las técnicas de Big Data y Data Science facilitan la creación de programas para la ayuda a la decisión por parte de los médicos. También se está trabajando para lograr una medicina casi personalizada a partir del estudio de los genes, por ejemplo, ayudando a prevenir futuras enfermedades que, a pesar de no haber presentado

síntomas, en los genes ya aparecen y actuando antes de su aparición es posible evitarlas (Zaforas, M., 2016)

### Evolución del acceso a información por parte de la población

La información disponible para consulta de cualquier ciudadano ha aumentado de una forma exponencial en apenas 20 años. La aparición de internet, sin lugar a duda, puede contarse como el factor principal que ha propiciado este aumento en el acceso a la información. En España, según datos del INE de 2019 (Instituto Nacional de Estadística, 2019), un 90,7% de la población usó Internet. Sin embargo, hay que tener en cuenta que existe una brecha dentro de la población entre aquellos que usan internet y los que no. Esta brecha digital se produce principalmente entre generaciones, siendo un 99% de la población de 16 a 24 años la que usa internet, mientras que solo el 63% de las personas de 65 a 74 años lo usan (Instituto Nacional de Estadística, 2019).

Poco a poco, los libros, enciclopedias e incluso los medios de comunicación tradicionales, como la televisión o la radio, han ido dejando paso a las búsquedas en internet como principal medio de información. Sin embargo, ¿sabemos qué es lo que encontramos en Internet? ¿quién lo ha escrito? ¿cuál es su veracidad? ¿tenía algún interés al escribirlo?... y así multitud de preguntas que nos podemos plantear sobre la información de internet.

Como se veía anteriormente, del análisis de datos se pude obtener conocimiento sobre multitud de campos. Sin embargo, a pesar de que es la actividad de los consumidores la que aporta todos los datos con los que las empresas pueden mejorar su rendimiento y ganancias, no hay opciones para que los consumidores se puedan beneficiar de este análisis a la hora de tomar decisiones.

## PROBLEMAS A RESOLVER

Una vez que se ha visto el contexto en el que se enmarca el proyecto, este se puede centrar en un objetivo principal: uso de datos ya creados para analizar la relación entre consumo alimenticio y enfermedades crónicas, a partir de la unión de dos problemas que surgen tras el análisis del contexto. Por un lado, está el número creciente de enfermedades crónicas y su relación con la alimentación, y, por otro lado, el crecimiento de los datos creados y recolectados que no llegan a beneficiar directamente a las personas, sino al rendimiento económico de grandes empresas con capacidad para analizar estos datos.

Con una mayoría de la población urbana ajena a la fabricación de alimentos, la gran parte de la comida que se consume es comprada y en muchos casos está ya procesada de algún modo, facilitando la preparación y ahorrando tiempo al consumidor (Marti, A., Calvo, C. y Martínez, A., 2021). Esto, además de ventajas como la facilidad y rapidez a la hora de cocinar y consumir, lleva también una serie de inconvenientes asociados. Uno de los problemas es que el consumidor no llega a conocer cómo ha sido ese procesamiento del alimento antes de la compra. Por otro lado, el aumento en la cantidad de alimentos ofertados ha llevado a un cambio en la dieta de las personas, pasando a dietas altas en grasas y azúcares. La relación entre estos cambios en la dieta y el aumento en el nivel de sedentarismo de la población, son uno de los principales factores del aumento de las enfermedades crónicas en las últimas décadas (Organización Mundial de la Salud, 2003).

Lograr frenar el aumento de estas enfermedades depende de la colaboración de todos los sectores de la población, autoridades sanitarias, políticas y sociedad civil. De entre las actuaciones que se podrían hacer dentro de la sociedad civil se encuentra la sensibilización tanto al resto de sociedad civil como a la clase política, exigiendo medidas eficaces. Sin embargo, para poder lograr este interés en la población, es necesario que la población realmente conozca la importancia de prevenir las enfermedades crónicas y la relación de los hábitos de vida con ellas. Este será, por tanto, el primer problema al que pretende dar soluciones este proyecto.

¿Cómo es posible sensibilizar a la población? ¿Cuáles son las preocupaciones actuales de la población en el ámbito de la alimentación y la salud? Como se veía en el contexto, entre la población española, se ve cada vez más preocupación por la alimentación y diferentes hábitos de consumo en busca de mejorar la calidad de la alimentación y la salud. Sin embargo, también se ha visto que esto no se está traduciendo en menores tasas de enfermedades crónicas, sino que estas siguen en aumento, especialmente la obesidad. ¿Qué es lo que está fallando? Esta cuestión puede tener varias respuestas y no habrá un único factor que esté dando lugar al problema. Tras realizar diferentes búsquedas internet es posible apreciar la gran cantidad de investigaciones y opiniones que hay respecto a este tema (¿son los alimentos ecológicos más sanos o solo márquetin?, ¿cuántas grasa/ azúcares… puedo tomar sin afectar a mi salud?, ¿dieta para bajar el colesterol /hipertensión…?).

De aquí surge el segundo problema en el que se ha centrado el proyecto. La gran cantidad de información disponible que hay en la web y, especialmente, la gran cantidad de opiniones, consejos y opciones se pueden encontrar sin poder saber cuál es mejor o si hay intereses detrás de esa opinión. Cada vez se encuentran más fraudes en determinados alimentos que se vendían como saludables y realmente no lo son y se está trabajando, por ejemplo, en un etiquetado más claro que facilite la elección de productos saludables por parte de la población (López-Cano, L., Restrepo-Mesa, S. et al., 2014). Sin embargo, la pregunta sobre qué alimentos son más sanos, qué dieta debo seguir o si hay alimentos específicos para aumentar o disminuir la tasa de hipertensión, colesterol o diabetes siguen estando sin respuesta.

Sabiendo que sería imposible establecer una única respuesta válida para el conjunto de la población el proyecto se ha centrado en analizar el conjunto de datos de consumo de España y establecer posibles relaciones que se estén creando entre el consumo de unos alimentos y las enfermedades crónicas. De este modo, no se trata de dar las causas o seleccionar unos alimentos como mejores o peores que otros, sino dar la información sobre lo que está pasando para que tanto la sociedad civil como expertos puedan verlos y establecer su propia conclusión.

## METODOLOGÍA

En este apartado se desarrolla el proceso seguido para llevar a cabo el análisis de datos. Se explica el tipo de metodología que se ha seguido, cómo se ha establecido la muestra con la que trabajar, las diferentes etapas en las que se ha dividido el proyecto, las técnicas que se han empleado para realizar el análisis y las limitaciones que se han tenido durante el desarrollo de todo este proceso. Pero antes de todo ello, se van a desarrollar algunos de los conceptos de análisis de datos a los que se hace referencia a lo largo de la explicación con el fin de facilitar la comprensión de este.

**Definición de conceptos:**

- Población: Conjunto formado por todos los elementos sobre los que se va a realizar el estudio. Normalmente, demasiado amplia para poder abarcar.
- Muestra: Subconjunto de la población que es seleccionado para realizar el estudio
- Variable: Característica de la muestra que va a ser estudiada. Puede ser cuantitativa, si es medible numéricamente, o cualitativa, si no se puede expresar mediante números.
- Análisis univariante: Análisis que solo estudia una variable de la población. No se preocupa de las relaciones que puedan existir entre estas y otras variables y principalmente, se centra en describir estos datos (Bastis Consultores, 2021).
- Análisis multivariante: Análisis en el que se emplean técnicas que permiten procesar simultáneamente diferentes variables estadísticas y estudiar la relación existente entre ellas (García, M., 2021).
- Correlación lineal simple: Método estadístico que estudia la relación lineal que puede existir entre dos variables. La correlación cuantifica cómo de relacionadas están ambas variables. Existen diferentes formas de calcular el coeficiente de correlación. La empleada en el documento ha sido la correlación de Pearson (Amat, J., 2016).
- Coeficiente de Pearson: El coeficiente de correlación de Pearson es la covarianza estandarizada, y su ecuación difiere dependiendo de si se aplica a una muestra, Coeficiente de Pearson muestral (r), o si se aplica la población, Coeficiente de Pearson poblacional (ρ). Toma valores entre [-1, +1], siendo +1 una correlación lineal positiva perfecta y -1 una correlación lineal negativa perfecta (Amat, J., 2016).
- P-valor: Además del valor obtenido para el coeficiente, es necesario calcular su significancia. Solo si el p-value es significativo se puede aceptar que existe correlación y esta será de la magnitud que indique el coeficiente. Por muy cercano que sea el valor del coeficiente de correlación a +1 o -1, si no es significativo, se ha de interpretar que la correlación de ambas variables es 0 ya que el valor observado se puede deber al azar (Amat, J., 2016).
- Regresión lineal simple: La regresión lineal simple consiste en generar un modelo de regresión (ecuación de una recta) que permita explicar la relación lineal que existe entre dos variables. A la variable dependiente o respuesta se le identifica como Y, y a la variable predictora o independiente como X.

  El modelo de regresión lineal simple se describe de acuerdo con la ecuación:

$$Y=\beta 0+\beta 1X1+\epsilon$$

  Siendo β0 la ordenada en el origen, β1 la pendiente y ϵ el error aleatorio (Amat, J., 2016).

**Tipo de metodología**

Puesto que uno de los objetivos principales con los que nace este proyecto es emplear técnicas propias de Data science en la resolución de diferentes preguntas relacionadas con la alimentación y la salud, la metodología seleccionada ha sido de índole cuantitativa. Además, dado que una de las bases

de datos con las que se ha trabajado ya estaba seleccionada (cedida por Nielsen), se ha tenido en cuenta antes del diseño del proyecto y la selección de técnicas, que las unidades con las que se trabajarían serían euros, número de personas y porcentajes (todas unidades numéricas).

### Muestreo

El muestreo consiste en la división de la población en los grupos que van a ser analizados posteriormente. Para este caso, la población era el conjunto de todos los españoles y se tenía que decidir la manera de dividirla (territorialmente, por edades…). Al contrario que en un análisis en el que se van a recoger los datos después de haber realizado el muestreo y, puesto que uno de los objetivos del trabajo era emplear los datos ya generados, el muestreo ha estado limitado por los datos con los que se iba a trabajar. Por lo tanto, el muestreo se realizó después de haber seleccionado las diferentes bases de datos con las que se iba a trabajar, puesto que era necesario comprobar cuánto se podían disgregar las mismas. Tanto en la base de datos de alimentación como las diferentes bases de datos de salud que se podían emplear la división de la población solo estaba hecha por comunidades autónomas. Después de contactar con los encargados de ambas bases de datos, se vio que no era crear grupos más pequeños como provincias. Esto limita la muestra y la convierte en 17 comunidades autónomas y 1 ciudad autónoma compuesta por Ceuta y Melilla. Un total de 18 unidades muestrales, lo cual da una idea de lo limitado que resultará el proyecto a la hora de establecer conclusiones. Sin embargo, al ser una muestra comprendida por datos que abarcan a toda la población de España continúa siendo un proyecto interesante como primera aproximación y base para posteriores investigaciones con más nivel de detalle y profundidad.

### Fases y cronograma

Para la realización del trabajo, se establecieron una serie de etapas que comprendían todos los pasos que se veían necesarios para cumplir el objetivo inicial marcado (poder dar información relevante acerca de la influencia del consumo de unos u otros alimentos en el desarrollo de enfermedades crónicas). En estas etapas estaba incluido desde el período de contacto con la empresa Nielsen y es estudio de la base de datos cedida, hasta la obtención de conclusiones y devolución de resultados. Estas se han desarrollado principalmente en el período de duración de las prácticas del máster (04/05/2019 al 29/07/2019), aunque, las fases iniciales de comunicación con Nielsen y búsqueda de las bases de datos se realizaron previamente a dicho periodo.

La primera etapa del trabajo, como se citaba anteriormente, fue el análisis de la base de datos de alimentación. Puesto que esta base de datos estaba fijada desde el comienzo del proyecto, era necesario conocerla y comprenderla bien para poder determinar junto a qué otras bases de datos se podían analizar obteniendo la información buscada. El período fue, por tanto, el análisis del consumo de la población española, en el que se analizó la base de datos que contenía el consumo de la población sin ninguna división territorial. Lo que se pretendía con este primer análisis era conocer información básica como la estructura de la base de datos, las categorías en las que se podían dividir los alimentos, las categorías de productos que se debían eliminar (por no ser alimentos), la división temporal que se podía realizar, la división de los alimentos que se podía establecer, las unidades con las que se podía trabajar… Además, también se pidió a la empresa información sobre alimentación ecológica, con el objetivo de comprobar si podía resultar relevante para el posterior análisis con las enfermedades crónicas.

De esta primera etapa, finalmente, se obtuvo una imagen general sobre el consumo de los españoles, así como sobre las características de las bases de datos que se debían de buscar. También se vieron algunas características estructurales de la base de datos que limitarían el posterior análisis y los resultados. Estas limitaciones se debían principalmente a la imposibilidad de dividir a la población más allá de las comunidades autónomas. Además, tampoco se podían dividir las categorías de alimentos y esto, como se verá posteriormente, es algo que hubiese sido necesario para determinar algunas conclusiones (por ejemplo, a la hora de determinar la relación entre la población que toma yogures y las enfermedades crónicas, no se puede tomar de igual forma un yogur natural o uno con azúcar).

La segunda etapa, consistió en realizar un estudio del estado del arte de proyectos similares al que se iba a realizar, así como contactar con diferentes expertos en alimentación. Con el estudio del estado del arte, se buscaba obtener información acerca de si había ya proyectos que estuviesen persiguiendo objetivos similares, cómo lo estaban haciendo, qué tipo de datos o técnicas de análisis estaban empleando, etc. Por otro lado, se contactó con diferentes expertos que estuviesen trabajando en proyectos que relacionasen enfermedades crónicas con la alimentación, con el objetivo de conocer su opinión sobre cómo diseñar el análisis con los datos de partida que se tenían. Así, se contactó con una profesora de la universidad, con un pediatra y un nutricionista que aportaron información y artículos para el estudio del estado del arte y dieron ciertas recomendaciones sobre las enfermedades a analizar. Además, la profesora vio relevante profundizar en el análisis de los alimentos de producción ecológica puesto era experta en nutrición ecológica y veía que se podían encontrar resultados interesantes en ese aspecto. Tras esta etapa, se tenía ya una idea más clara de las enfermedades que se iban a analizar (diabetes, colesterol, hipertensión, obesidad, cáncer…) y se podía empezar a buscar bases de datos con estas enfermedades que cumpliesen las características necesarias para poder analizarlas junto a los datos de alimentación.

La búsqueda de bases de datos que tuviesen información de estas enfermedades con las características que se habían establecido en el análisis de la base de datos de alimentación, fue la tercera etapa. Se había determinado que las enfermedades que más interés podían tener para el análisis eran enfermedades crónicas como el colesterol, la diabetes y la hipertensión. La obesidad, vista como enfermedad o como factor de riesgo que hace aumentar la posibilidad de sufrir otras enfermedades crónicas, también se tenía como relevante para el análisis. Finalmente, el cáncer era otra de las enfermedades que tras, en la etapa anterior haber visto que había numerosas investigaciones que analizaban su relación con la alimentación, se había seleccionado como posible enfermedad a analizar. Esta búsqueda se realizó en distintas páginas oficiales de datos, tanto autonómicas como nacionales. Se descartaron algunas debido a que los datos no estaban suficientemente actualizados y no concordaban con el período de tiempo que se poseía en la base de datos de alimentación. Además, algunas incluían datos de provincias o municipios, pero no de comunidades autónomas. Se decidió no analizar el cáncer puesto que era una enfermedad muy amplia y no había una base de datos que tuviese las características que se buscaban. Finalmente, se seleccionaron los datos del Instituto Nacional de Estadística (INE) puesto que eran los que poseían unas características que permitían el análisis junto a los datos de alimentación.

La cuarta etapa, fue el diseño del análisis que se iba a llevar a cabo, así como la selección de las técnicas que se iban a emplear. Durante el diseño del análisis que se iba a realizar, se contó con la ayuda de expertos en nutrición, tal y como ocurrió en el proceso de búsqueda de la base de datos de

enfermedades. Con una visión más clara de la información que se tenía y de las principales limitaciones con las que se iniciaba el proceso, se pudo diseñar un análisis formado por dos etapas. Una primera etapa en la cual se analizaría la relación existente entre el porcentaje de gasto destinado a diferentes categorías de alimentos por cada comunidad y el porcentaje de personas con una enfermedad crónica o con obesidad que había en la misma comunidad en ese período de tiempo. Una vez establecidas las categorías de alimentos que tenían relación con alguna de las enfermedades analizadas, se pasaría a la segunda etapa en la que analizar esa relación más detalladamente.

La última etapa de este proyecto fue el análisis de los resultados obtenidos y la creación de un documento en el que mostrasen las principales conclusiones como devolución a la Nielsen y para poder cumplir con el objetivo de devolver la información a la población y ayudar a mejorar el conocimiento general en este ámbito. Además, tras analizar también las limitaciones que han caracterizado este proceso, se determinaron también una serie de características que se deberían buscar para poder hacer futuros proyectos en este sentido más detallados y con resultados más profundos.

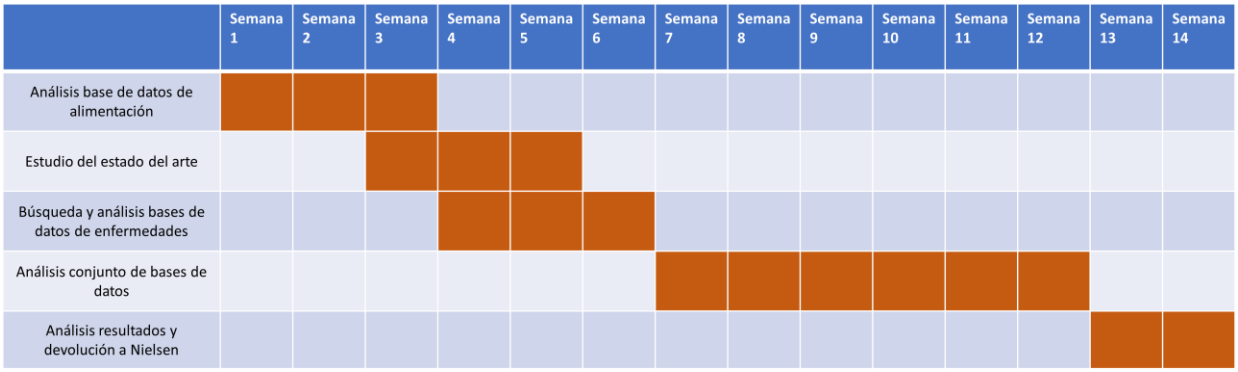


**Imagen 1: Cronograma del proyecto. Fuente: elaboración propia**

### Selección de técnicas

Las técnicas que han sido empleadas en este proyecto han estado limitadas por la escasa muestra con la que se contaba. Como se ha visto anteriormente, solo se contaba con una muestra formada por 18 comunidades autónomas y el período temporal que se podía analizar era solo el dato de un año, puesto que era el único en el que ambas bases de datos coincidían. Esta escasez muestral llevaba a resultados poco fiables e imposibilitaban el uso de algunas técnicas. Por ello, se fue probando con diferentes opciones y comprobando si los resultados eran válidos.

Las técnicas empleadas han sido tanto descriptivas como inferenciales. Las técnicas descriptivas se han empleado principalmente para comprender cada base de datos por separado, primero en la base de datos de alimentación y, posteriormente en la de cada enfermedad. Estas técnicas han estado dirigidas a facilitar la comprensión de cada base de datos, así como a ordenar y permitir visualizar los datos de interés de una forma clara.

Una vez que se conocían ambas bases de datos, se emplearon diferentes técnicas de análisis inferencial. En la primera fase del análisis se emplearon técnicas de correlación lineal. De esta manera

se esperaba comprobar si el hecho de que una comunidad autónoma dedicase mayor o menor porcentaje del gasto a una categoría de alimentos estaba relacionado con un mayor o menor porcentaje de alguna enfermedad crónica en esa comunidad. Para cuantificar esa relación, se calculó, además, el coeficiente de Pearson, el cual oscila entre los valores -1 y +1. Un valor -1 significa una relación lineal perfecta negativa (el aumento de una variable se traduce en una disminución en la otra), 0 una relación nula y +1 una relación lineal perfecta positiva (un aumento en una variable es un aumento en la otra).

A continuación, se trataron de determinar relaciones entre más de dos variables (comprobar si había, por ejemplo, determinadas categorías de alimentos, que en conjunto tenían relación con aquellas comunidades donde había un mayor o menor porcentaje de personas con cierta enfermedad). Para ello, se realizó una correlación multivariante. Sin embargo, como se explicaba anteriormente, el hecho de no poder dividir a la población en más muestras y no poder ver una evolución temporal más completa, ha impedido usar técnicas más avanzadas como las multivariantes dado que no se obtenían buenos resultados. Tener solamente 18 puntos en el análisis (el valor de cada comunidad autónoma medido en 2017) y más del doble de variables (cada categoría de alimentos) impedía realizar un análisis correcto multivariante.

Lo siguiente que se hizo fue seleccionar aquellas categorías para las que el coeficiente de Pearson indicaba que existía una relación lineal fuerte con alguna categoría y tratar de explicar esa relación a través de un modelo de regresión lineal en el cual la obtención de un p_valor menor de 0.05 sería tomado como significativo y verificaría la existencia de esa relación entre el consumo de un alimento y la prevalencia de una enfermedad crónica determinada.

Finalmente, se intentó emplear la técnica de los árboles de decisión, pero debido a la escasez de variables con las que se contaba no se pudo llevar a cabo.

### Limitaciones

Las principales limitaciones con las que se ha tenido que lidiar en el desarrollo de este trabajo han estado relacionadas con la escasez muestral. Como ya se ha visto, solo se contaba con 18 valores de cada variable (categorías de alimentos o enfermedades). Esto se ha debido, por un lado, a la imposibilidad de dividir a la población en sectores más pequeños y, por el otro, a no poder analizar más períodos de tiempo (ya fuesen de menor duración temporal o fuese un mayor número de años con el que trabajar).

Además, al ser una empresa la que cedía los datos, no podía dar información concreta sobre los productos, como el nombre de la marca, por si salía algún tipo de información que pudiese resultar perjudicial para esa marca.

Por otro lado, las enfermedades crónicas son enfermedades sobre las que afectan múltiples factores, por lo que el hecho de que solamente se hayan analizado datos de alimentación va a proporcionar un conocimiento limitado. Sería necesario analizar otros factores como la actividad física o incluso factores económicos de la población. Sin embargo, al haber trabajado con comunidades autónomas concretas, encontrar estos datos para cada comunidad no ha sido posible.

Finalmente, otra limitación que hay que tener en cuenta es el hecho de que la base de datos de alimentación se mida en gasto, ya que dentro de una categoría no todos los productos tienen el mismo

precio. Por ejemplo, si se tuviesen dos comunidades autónomas, una con un gasto de100 en yogures y otra con un gasto de 150. El hecho de que el gasto de una sea mayor que el de la no tiene que deberse solamente a la cantidad comprada, puede ser también que en una comunidad el producto sea más caro. Esta limitación se debe tener en cuenta a la hora de interpretar los datos y se solventa analizando el porcentaje del gasto destinado a cada categoría (estando la posibilidad de que se deba a su mayor precio o a la decisión de esa comunidad de consumirlo más).

## DESCRIPCIÓN DEL PROYECTO

En este apartado se va a explicar el proceso de análisis de datos que se ha realizado siguiendo las etapas y la metodología desarrolladas anteriormente y se irán mostrando los resultados obtenidos en dichas etapas. En primer lugar, se expondrán las características de la base de datos de alimentación y el análisis previo realizado sobre la misma. A continuación, se verá el estudio del estado del arte que se realizó y las principales conclusiones extraídas del mismo, a partir de las cuales, se hizo la búsqueda de la base de datos de enfermedades crónicas. Después, se verá el análisis los datos de las enfermedades crónicas con los que se trabajaron. Seguidamente, se explicará todo el proceso de análisis que se ha llevado a cabo y se expondrán los resultados que se han obtenido. Para finalizar, se propondrán una serie de resultados o conclusiones del estudio realizado.

### Base de datos de alimentación

Como se veía anteriormente, esta base de datos fue cedida por Nielsen y es la base de datos con la que se contaba para iniciar el proyecto. Se ha trabajado con varios formatos de esta base de datos. Se recibió una primera base de datos del conjunto de la población española, sin divisiones territoriales y 18 bases de datos con la misma información de consumo, pero con la información de cada comunidad autónoma por separado. A la hora de realizar este primer análisis, se utilizó la base de datos general y, una vez que se comprendía esta, se pasó todo ese conocimiento y la limpieza y preparación de los datos establecida a las distintas bases de datos de las comunidades autónomas. En el análisis de la base de datos general, primero se observaron las características generales esta. A continuación, se hizo un análisis inicial del consumo de la población española.

#### *Características de la base de datos*

Las características de esta base de datos son las siguientes:

- Recoge el gasto económico destinado a diferentes categorías de alimentos.
- Los alimentos están divididos en diferentes categorías dependiendo del tipo de alimento y de su procesamiento. Las categorías se pueden observar en las diferentes columnas que se muestran en la imagen 2.
- Los datos abarcan más del 90% del gasto en alimentación de España.
- No incluyen el gasto en alimentos frescos ni el gasto realizado en pequeños comercios, solo en supermercados e hipermercados.
- Hay una división dentro de cada categoría de alimento para distinguir si ha sido un producto ecológico o no.

- Los productos ecológicos se recogen con una etiqueta BIO y son aquellos marcados en el supermercado con el sello de certificación ecológica de la unión europea (hoja verde).
- La información está segmentada, territorialmente por comunidades autónomas y temporalmente, en semanas desde la semana 8 de 2016 a la semana 12 de 2020.
- No se puede segmentar la información hasta llegar productos concretos ni a marcas de estos productos

*Limpieza de la base de datos*

La base de datos de alimentación estaba dividida en diferentes categorías, tal y como se muestran en la siguiente imagen.

| SECTOR | SECCIÓN | CATEGORÍA | FAMILIA | SEGMENTO BIO |
|---|---|---|---|---|
| Alimentación y bebidas | Alimentación seca | - Aceite<br>- Aditivos de cocina<br>- Alimentos infantiles<br>- Alimentos mascotas<br>- Aperitivos<br>- Arroz<br>- Azúcar y edulcorantes<br>- Bollería industrial<br>- Cacao<br>- Cafés<br>- Cereales desayuno<br>- Chocolates<br>- Dietéticos<br>- Dulces navideños<br>- Frutos secos<br>- Galletas<br>- Golosinas<br>- Infusiones<br>- Legumbres secas<br>- Panadería industrial<br>- Pastas<br>- Reportaría<br>- Salsas<br>- Sopas y deshidratados | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Alimentación y bebidas | Bebidas | - Agua<br>- Bebidas alcohólicas<br>- Bebidas refrescantes<br>- Cerveza<br>- Espumosos<br>- Vinos<br>- Zumos | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Alimentación y bebidas | Conservas | - Aceitunas y encurtidos<br>- Conservas de carne y patés<br>- Conservas de frutas y dulces<br>- Conservas de pescado<br>- Conservas vegetales<br>- Platos preparados conserva | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Alimentación y bebidas | Leche y batidos | - Batidos<br>- Horchata<br>- Leche líquida y bebidas vegetales<br>- Leche no líquida | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Droguería y perfumería | Droguería y limpieza | - Ambientadores<br>- Celulosas hogar<br>- Complementos cocina<br>- Complementos lavado<br>- Detergentes ropa<br>- Insecticidas<br>- Lavavajillas<br>- Lejías y desinfectantes<br>- Limpiadores del hogar<br>- Limpieza calzado<br>- Suavizantes ropa<br>- Útiles limpieza | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |

| SECTOR | SECCIÓN | CATEGORÍA | FAMILIA | SEGMENTO BIO |
|---|---|---|---|---|
| Droguería y perfumería | Perfumería e higiene | - Afeitado<br>- Cuidado capilar<br>- Cuidado corporal<br>- Fragancias<br>- Higiene bucal<br>- Higiene corporal<br>- Higiene femenina<br>- Maquillaje<br>- Otros higiene<br>- Pañales<br>- Parafarmacias gran consumo<br>- Protección solar<br>- Tratamiento facial | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Productos frescos | Charcutería | - Ahumados<br>- Charcutería<br>- Pates refrigerados<br>- Salazones<br>- Salchichas refrigeradas<br>- Sobrasada | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Productos frescos | Congelados | - Helados<br>- Otros congelados<br>- Pescado congelado<br>- Pescado preparado congelado<br>- Platos preparados congelados<br>- Verduras congeladas | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Productos frescos | Derivados lácteos | - Mantequilla<br>- Margarina<br>- Nata<br>- Postres preparados<br>- Queso tipo petit<br>- Yogures | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Productos frescos | Quesos | - Queso azul<br>- Queso bola<br>- Queso fresco<br>- Queso fundido<br>- Queso pasta blanda<br>- Queso rallado<br>- Queso tradicional<br>- Quesos blancos pasteurizados<br>- Requesón<br>- Resto quesos naturales | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |
| Productos frescos | Platos cocinados y precocinados | - Platos preparados refrigerados | Categoría desglosada en productos concretos | Cada producto se divide en BIO o NO BIO |

**Imagen 2: Segmentación inicial base de datos de alimentación. Fuente: elaboración propia**

Se puede observar que hay 5 columnas que corresponden a las diferentes categorías delimitadas por Nielsen. La primera categoría es el Sector, la cual divide a los productos según correspondan a alimentación y bebida (no fresca), droguería (eliminada para el posterior análisis) o productos frescos.

La segunda categoría establece subdivisiones dentro de cada sector, la sección. En esta se dividen los productos según pertenezcan a productos secos, bebidas, conservas, batidos y leche (hasta aquí pertenecientes al sector alimentación y bebida), derivados lácteos, quesos, charcutería, congelados o platos precocinados (estos pertenecientes al sector productor frescos).

A continuación, se establece una subdivisión más, la categoría. Dentro de cada sección los productos son agrupados y separados en 95 categorías (incluyendo droguería) según ya se trate de una clase de productos u otros. Por ejemplo, en la sección de derivados lácteos se pueden encontrar 6 categorías diferentes, mantequilla, margarina, nata, postres preparados, queso tipo petit y nata.

Una división más es la familia. En esta, cada categoría de alimentos es de nuevo dividida atendiendo a diferentes características del producto. Por ejemplo, en la categoría yogures, se encuentran familias según el tipo de yogur (desnatado, con bífidus, de sabores, natillas…). Esta división no se muestra debido a que son demasiadas y no se ha trabajado con ellas.

Finalmente, la última columna muestra una subdivisión dentro de cada producto según corresponda a un producto con certificación ecológica o no.

En cuanto a la segmentación temporal, la base de datos está dividida por semanas, por lo que se seleccionaron las semanas pertenecientes al año 2017.

Para hacer la limpieza de datos se eliminaron aquellas categorías que incluían productos de droguería, perfumería y alimentación para mascotas. Además, no se seleccionaron los productos pertenecientes a la columna ‘familia” pese a poder obtenerse mucha información de ella, debido a la escasa muestra con la que se podía trabajar. De este modo, se seleccionaron 69 categorías de alimentos y se realizó la media del gasto durante 2017 en cada una de estas categorías.

*Análisis Consumo alimenticio de la población española*

Este primer análisis permitía tener una imagen completa y amplia sobre lo que se iba a trabajar, pero dado que no es uno de los objetivos del trabajo, no se mostrarán todos los resultados sobre el consumo de alimentos en España, sino aquellos vistos como más relevantes para ayudar a la comprensión del resto del proceso. A continuación, se muestra una gráfica en la que aparece el porcentaje del gasto en el supermercado que la población española destinada a cada categoría.

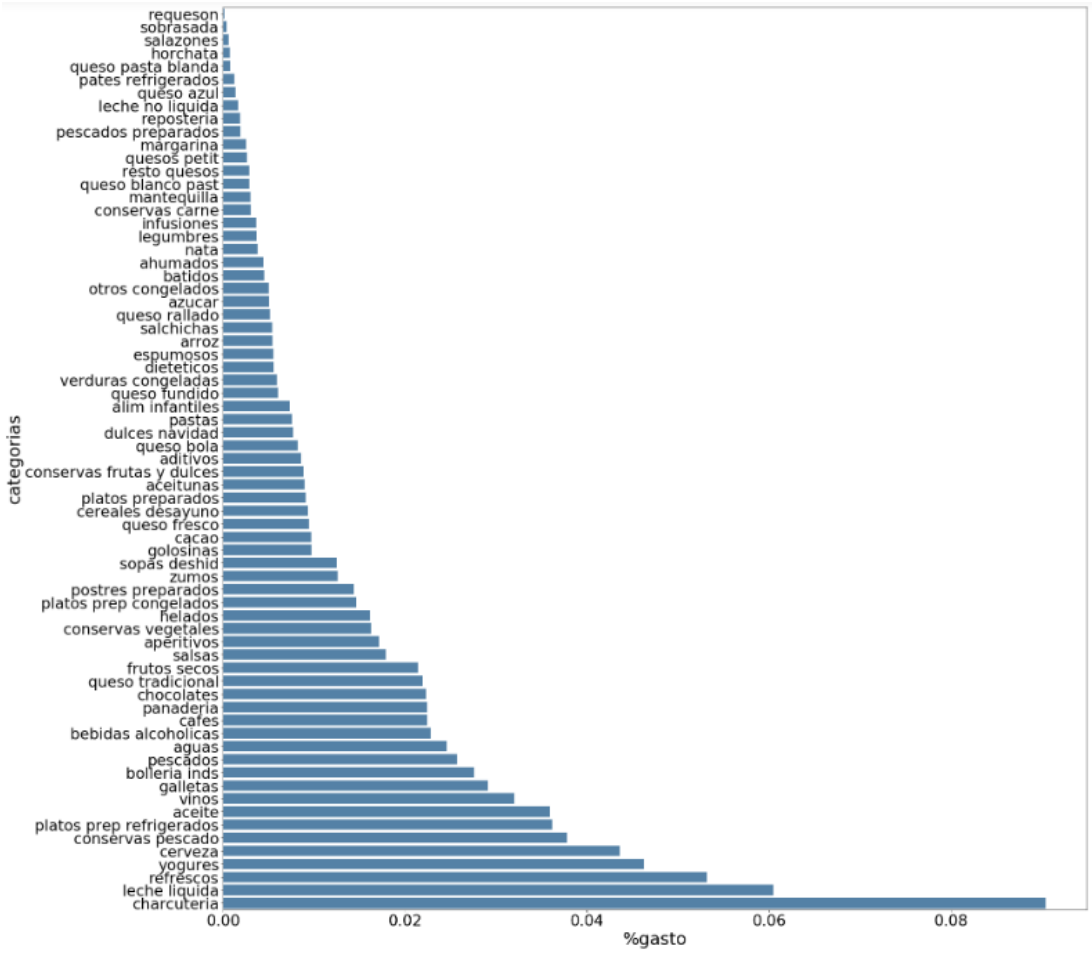


**Imagen 3: Distribución gasto por categorías en España. Fuente: elaboración propia**

Se han estudiado 69 categorías de alimentos, pero no se tuvieron en cuenta las cuatro categorías que representaban un porcentaje menor ya que el porcentaje no llegaba al 1% y, además, al ver el consumo por comunidades, se apreciaba que los consumidores de estas categorías estaban muy localizados. Esto podía alterar los resultados de posteriores análisis dando lugar a correlaciones erróneas. Las categorías eliminadas fueron: requesón, sobrasada, salazones y horchata.

Las 5 categorías a las que se destina mayor porcentaje de gasto son: charcutería, leche líquida, refrescos, yogures y cerveza.

De este gráfico resulta relevante comprobar que, si se realiza el mismo análisis dividiendo la población en comunidades autónomas, se puede apreciar que el orden en el que se distribuye el gasto es similar en todas ellas.

Por otro lado, se analizó el porcentaje del gasto que la población española destinaba a productos ecológicos. En la siguiente imagen, se establece la división geográfica de este porcentaje de gasto ecológico.

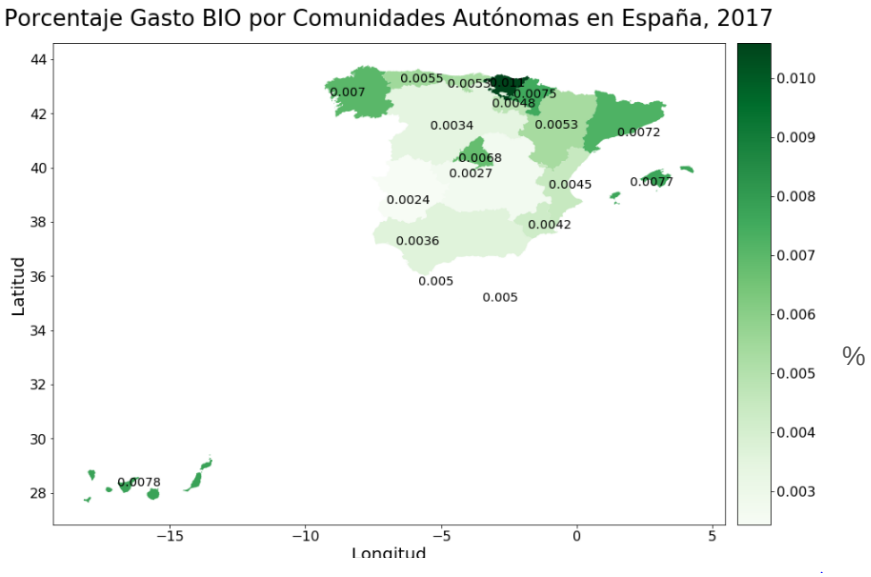


**Imagen 4. Distribución del gasto en productos BIO por CC.AA. Fuente: elaboración propia**

El porcentaje del gasto destinado por los españoles a alimentación BIO no llega a ser el 1% del gasto total en ninguna comunidad autónoma excepto en País Vasco donde alcanza el 1.1%. En general, País Vasco, Islas Canarias, Baleares, Navarra y Cataluña son las comunidades que mayor porcentaje del gasto invierten en alimentos productos ecológicos. En este análisis no se muestra, pero si se observa el gasto por semanas se puede ver que este va aumentando con el tiempo, aunque sigue siendo una parte despreciable del gasto en el conjunto de la población.

### Estado del arte

Previamente a definir el proceso de análisis se realizó un estudio del estado del arte. De esta manera, se esperaba conocer qué otros proyectos con objetivos similares se están realizando y conocer también, cuál es el uso que se está haciendo de las técnicas de Big Data en este sector. Este estudio se realizó de forma muy amplia puesto que no se tenían aún definidas las bases de datos con las que se trabajarían. Por lo tanto, se comenzó desde preguntas muy abiertas que se fueron centrando en diferentes preguntas a las que se podrían ir dando respuesta con el proyecto.

La primera búsqueda de proyectos que se realizó iba dirigida a conocer el uso de técnicas de Big Data en el sector de la alimentación, así como conocer quiénes eran los que lo estaban usando. Realizando búsquedas en inglés y español en Google Scholar, World Wide Science y The Lancet, principalmente, se observó que los principales resultados iban dirigidos a la industria. En estos artículos, se explicaba cómo ciertas empresas ya estaban empezando a emplear técnicas de Data Science en su producción para optimizar toda la cadena. También se encontraron múltiples estudios de cómo usarlo en márquetin sobre alimentación. Por último, otro de los resultados recurrentes, eran proyectos que trazaban un posible futuro en el cual, con técnicas de análisis de datos se pudiese optimizar todo el sector y poder alimentar un mayor número de personas (Kersting, K., Bauckhage, C. et al., 2016), ya sea mejorando los alimentos o mejorando la producción de los mismos.

Si en la búsqueda se incluían términos sobre alimentación ecológica, los resultados más recurrentes eran ahora, sobre efectos del cambio climático en la salud en general y en diferentes enfermedades como el cáncer o enfermedades respiratorias. También había resultados sobre cómo se podría tener una alimentación ecológica en el futuro o sobre agricultura más sostenible o, incluso, resultados sobre la influencia de los sistemas alimentarios en los ODS.

Por último, se realizaron búsquedas en las que ya se mezclaba el uso de Big Data, la alimentación, la alimentación ecológica y las enfermedades crónicas. Como resultado de esta búsqueda sí que se encontraban diferentes estudios en los que se analizaba la alimentación seguida por una población concreta y se obtenían diferentes conclusiones relacionadas con la salud. Por ejemplo, relación entre alimentación ecológica y el nivel de pesticidas en la orina (Hyland, C., Bradman, A. et al., 2018) o un artículo de The Lancet (Swinburn, B., Kraak, V et al., 2019) sobre la obesidad, la desnutrición y el cambio climático.

Tras esta etapa, se pudo ver que en España no había proyectos de estas características, pero en el extranjero sí que había artículos con análisis similares. Sin embargo, todos tenían una diferencia con el que se iba a llevar a cabo, la población no era tan amplia, sino que se trataba de personas que habían sido seleccionada y seguidas a lo largo de un periodo de tiempo determinado.

### Búsqueda y análisis de bases de datos de enfermedades

Como se explica en el apartado de metodología, la búsqueda de la base de datos de enfermedades crónicas se realizó con una serie de condiciones. Estas eran que se pudiera segmentar la muestra en comunidades autónomas, que comprendiera el mismo período de tiempo que la base de datos de alimentación y que los datos estuviesen medidos en las mismas unidades o se pudiesen calcular porcentajes respecto a la población de la misma forma (porcentaje de gasto respecto al gasto total de la comunidad y porcentaje de enfermos respecto al total de la población de la comunidad).

Los datos finalmente provienen de la Encuesta Nacional de Salud de 2017 puesto que eran los datos que mejor se podían analizar de forma conjunta con los datos de alimentación. Las características de estos datos son las siguientes:

- Las enfermedades crónicas analizadas han sido obesidad, colesterol, hipertensión y diabetes.
- La información está segmentada, territorialmente por comunidades autónomas y temporalmente son datos pertenecientes a 2017.
- La unidad en la que se muestran los datos es en porcentaje dentro de cada comunidad autónoma.

#### *Análisis base de datos de enfermedades crónicas*

A continuación, se muestran una serie de gráficos en los que se puede observar la división geográfica del porcentaje de personas que sufren cada enfermedad crónica en España.

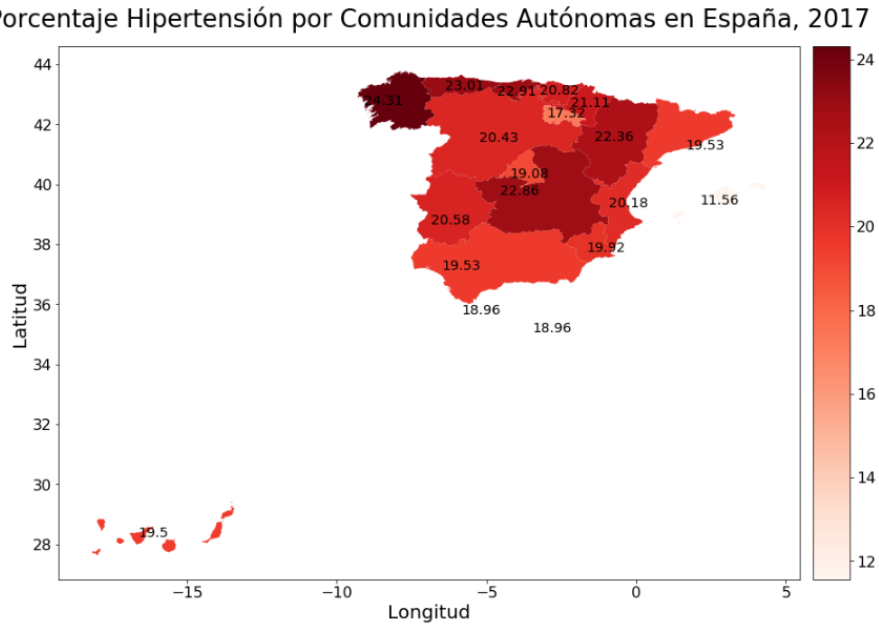


**Imagen 5. Distribución porcentaje de hipertensión por comunidades**

Se observa que las comunidades con mayor porcentaje de hipertensión son Galicia, Asturias, Cantabria, Castilla la Mancha y Aragón, mientras que aquellas con menor porcentaje de hipertensión son Islas Baleares, La Rioja, Ceuta y Melilla, Madrid y las Islas Canarias.

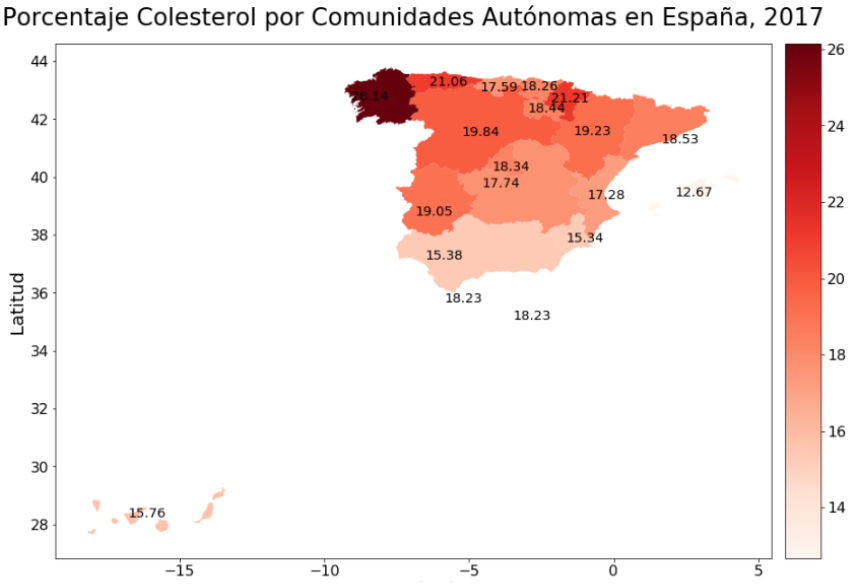


**Imagen 6. Distribución porcentaje de colesterol por comunidades autónomas Fuente: elaboración propia**

Se observa que las comunidades con mayor porcentaje de colesterol son Galicia, Navarra, Asturias, Castilla y León y Aragón, mientras que aquellas con menor porcentaje de colesterol son: Islas Baleares, Andalucía, Murcia, Islas Canarias y Comunidad Valenciana.

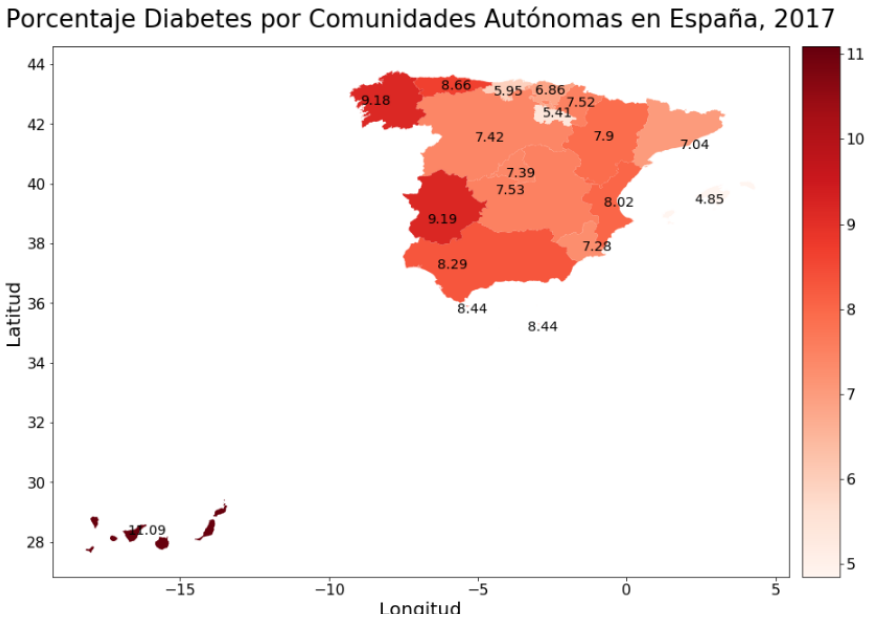


**Imagen 7. Distribución porcentaje de diabetes por comunidades autónomas. Fuente: elaboración propia**

Se observa que las comunidades con mayor porcentaje de diabetes son Islas Canarias, Extremadura, Galicia, Asturias y Ceuta y Melilla, mientras que con menor porcentaje de diabetes son Islas Baleares, La Rioja, Cantabria, País Vasco y Cataluña.

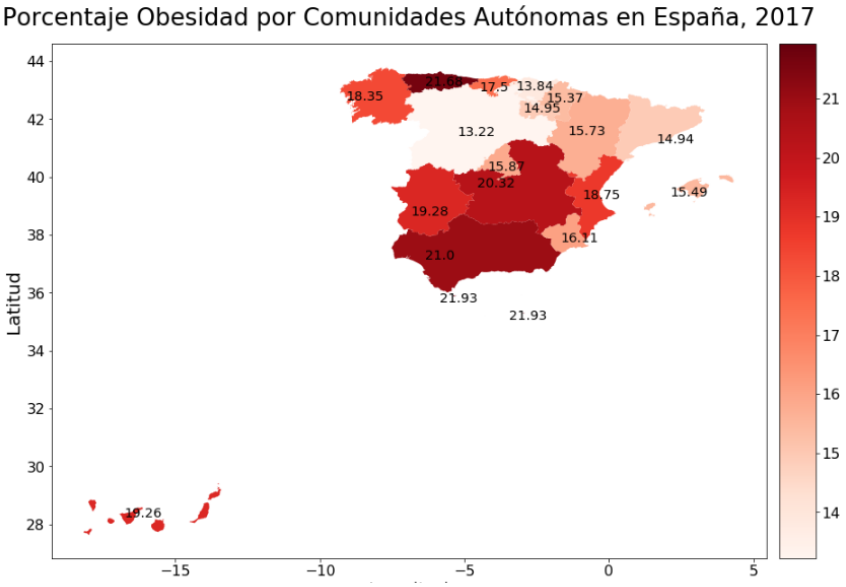


**Imagen 8. Distribución porcentaje de obesidad por comunidades autónomas. Fuente: elaboración propia**

Las comunidades con mayor porcentaje de obesidad son Ceuta y melilla, Asturias, Andalucía, Castilla La Mancha y Extremadura, mientras que aquellas con menor porcentaje de obesidad son Castilla y León, País Vasco, Cataluña, La Rioja, Navarra e Islas Baleares.

### Análisis de datos conjunto

Tras tener una imagen general de ambas bases de datos, el siguiente paso era analizar las posibles relaciones que existían entre ambas. Para ello, como se ha ido viendo anteriormente, se estableció una muestra de estudio formada por 18 comunidades autónomas y solo se obtuvo un valor para cada comunidad autónoma, la media del porcentaje del gasto en alimentación en 2017 o el porcentaje de personas con cada enfermedad crónica en 2017.

Tal y como se explicó en la metodología, el análisis se ha dividido en dos partes. Primero se realizaron diferentes correlaciones lineales entre las distintas categorías de alimentos y cada enfermedad crónica para establecer aquellas categorías que poseían una relación lineal fuerte. A continuación, se seleccionaron dichas categorías y se creó un modelo de regresión lineal para cada una de ellas.

Sin embargo, antes de este análisis, se muestra un primer análisis entre el % de gasto destinado a productos ecológicos y el % de enfermos crónicos en una comunidad autónoma. Este análisis no se ha podido profundizar más, puesto que el porcentaje de gasto en productos ecológicos es muy pequeño y al separarlo por categorías los resultados no eran concluyentes.

#### *Correlaciones entre % gasto BIO y enfermedades*

Si se vuelven a observar los mapas anteriormente mostrados (los cuatro de las enfermedades crónicas y el del consumo de productos BIO), ya visualmente, se puede apreciar que existe cierto grado de correlación negativa entre la obesidad y el % de gasto bio. Para poder analizarlo con claridad, se vuelven a mostrar ambos mapas.

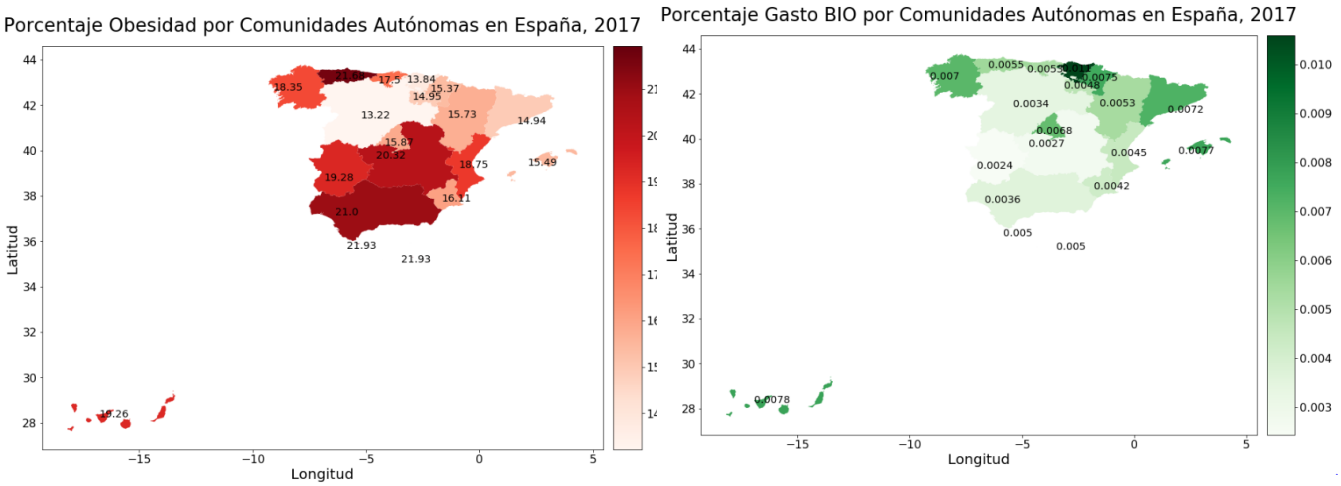


**Imagen 9. Comparación distribución territorial gasto en productos BIO y porcentaje de obesidad. Fuente: elaboración propia**

Es posible apreciar que, por ejemplo, el País Vasco, en el mapa de la obesidad está pintado de color blanco, que corresponde a los porcentajes más bajos de obesidad (anteriormente se ha visto que es la segunda comunidad autónoma con menor porcentaje de obesidad), mientras que en el mapa de productos ecológicos es la comunidad pintada de verde más oscuro, que corresponde a un mayor porcentaje de gasto en estos productos. Otros casos similares se pueden observar en Madrid o en las Islas Baleares.

Un caso contrario, lo podemos apreciar en Andalucía, la cual está pintada de rojo oscuro en el mapa de la obesidad (alto porcentaje), mientras que aparece en un verde muy claro en el mapa de alimentos ecológicos (bajo porcentaje). Otros casos similares se pueden observar en Castilla la Mancha o Extremadura.

En el siguiente gráfico de barras se muestra el valor de la pendiente de la recta del modelo de regresión lineal creado para analizar si realmente existía una relación entre el gasto destinado a alimentación ecológica y el porcentaje de alguna de las enfermedades. En este gráfico, en el eje Y, cada barra representa a una de las enfermedades analizadas y, en el eje X, se muestran los valores alcanzados por dicha pendiente. Los valores negativos significan que la relación es negativa (a mayor porcentaje de gasto, menor porcentaje de enfermos y viceversa) y, los valores positivos indican que la relación existente es positiva.

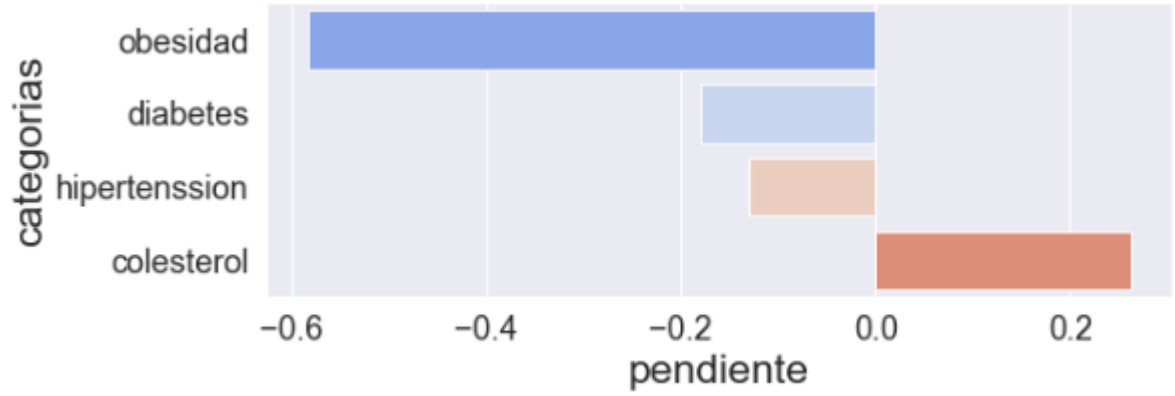


**Imagen 10. Pendientes modelos de regresión entre cada enfermedad y el porcentaje de gasto BIO. Fuente: elaboración propia**

Se puede observar que el valor de la pendiente de la recta de regresión entre obesidad y porcentaje de gasto bio es cercano al -0.6. Es el único valor que puede ser considerado como relación alta. Esta relación es negativa, lo que significa que a mayor porcentaje de gasto destinado a productos bio en una comunidad, menor porcentaje de obesidad hay en esa comunidad.

Para poder confiar en este dato, es necesario conocer el valor del p_valor obtenido en el modelo de regresión, el cual, como se explicó anteriormente, será significativo si su valor es inferior a 0.05. En este caso, el valor del coeficiente fue 0.023, lo cual es un valor significativo.

*Correlaciones entre % gasto general y enfermedades crónicas*

Las correlaciones lineales que se han realizado muestran la existencia o no de una relación directa o inversa entre el porcentaje de gasto que una comunidad destina a una categoría de alimentos y el porcentaje que hay en esa comunidad de una de las enfermedades crónicas analizadas.

Para mostrar los resultados de estas correlaciones, las categorías se han dividido según la sección a la que corresponden en la base de datos. Para cada grupo, se va a mostrar una matriz de correlación entre las categorías de esa sección y las cuatro enfermedades crónicas. En las matrices de correlación los colores de cada celda señalan el tipo de relación que existe entre las variables que forman dicha

celda. De este modo, cuanto más oscuro es el color, significa mayor correlación negativa, mientras que cuanto más clara es la celda mayor correlación positiva hay entre ambas variables. El color rosa fucsia indica una relación baja o nula. Dentro de cada celda aparece un número, el cual es el valor del coeficiente de Pearson para las variables a las que pertenece dicha celda.

Tras mostrar el gráfico para cada sección, se determinarán aquellas categorías con mayor relación. Esta se ha establecido a partir de un valor del coeficiente de Pearson de 0.4.

- **Correlaciones en las categorías de alimentación seca**

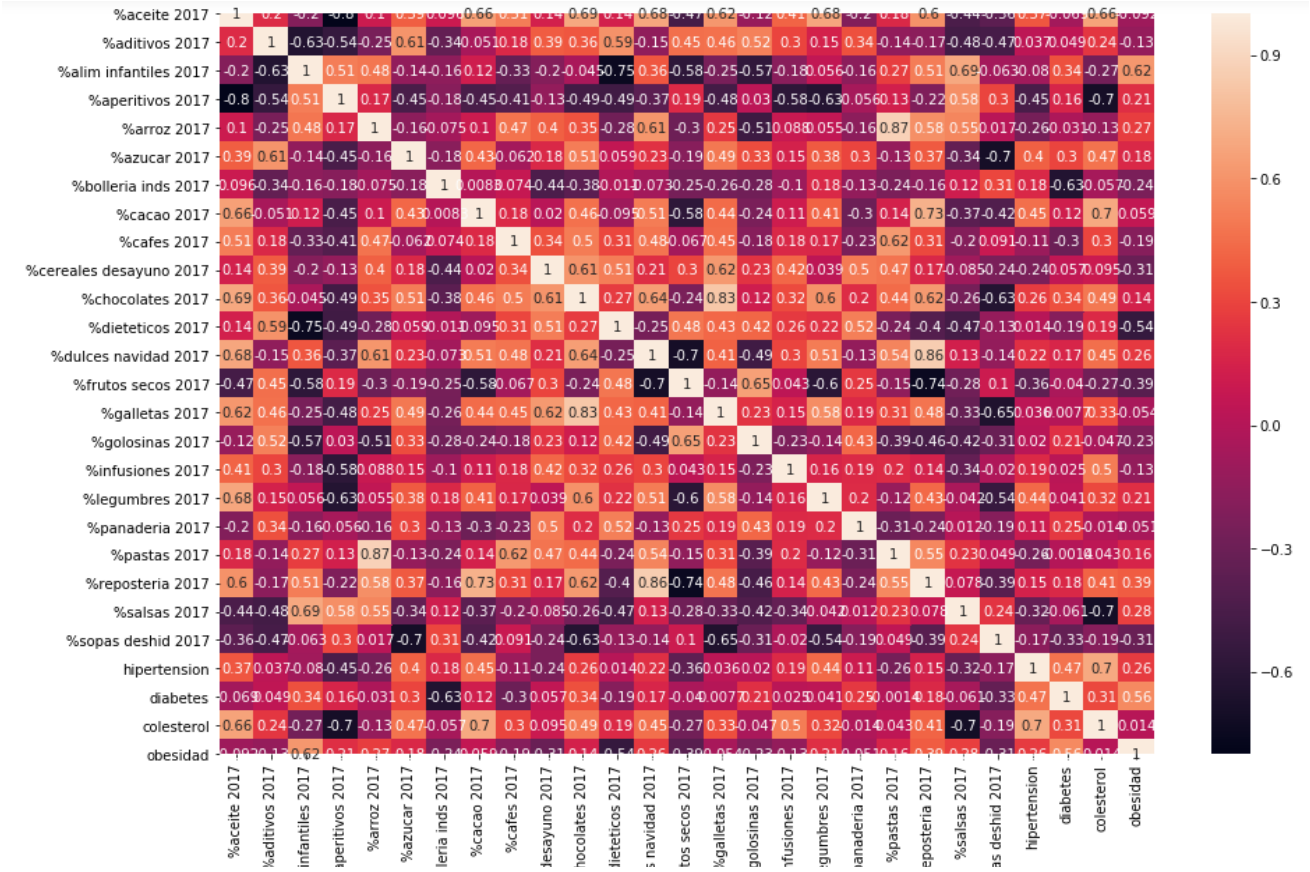


**Imagen 11. Tabla correlaciones entre enfermedades y categorías de la sección alimentación seca. Fuente: elaboración propia**

Se observa que las categorías con valores de correlación más altos para la hipertensión son aperitivos (-0.45), cacao (0.45), legumbres (0.44), azúcar (0.4) y aceite (0.36). Para el colesterol**,** las categorías con un coeficiente superior al 0.4 son salsas (-0.7), aperitivos (-0.7), cacao (0.7), aceite (0.66), infusiones (0.5) y chocolates (0.49). En el caso de la diabetes, solo se encuentra la bollería industrial (-0.63). Finalmente, para la obesidad, las categorías seleccionadas son alimentación infantil (0.62) y dietéticos (-0.54).

- **Correlaciones en las categorías de bebidas**


**Imagen 12. Tabla correlaciones entre enfermedades y categorías de la sección bebidas. Fuente: elaboración propia**

Seleccionando ahora las categorías un coeficiente de Pearson superior a 0.4, para la hipertensión se obtienen agua (-0.57), zumos (-0.53) y espumosos (0.4). Para el colesterol, se seleccionan agua (-0.55), refrescos (-0.5), bebidas alcohólicas (-0.5) y zumos (-0.44). Para la obesidad se seleccionan refrescos (0.44) y espumosos (-0.4). Finalmente, para la diabetes no hay ninguna categoría que cumpla esta condición.

- **Correlaciones en las categorías de conservas**

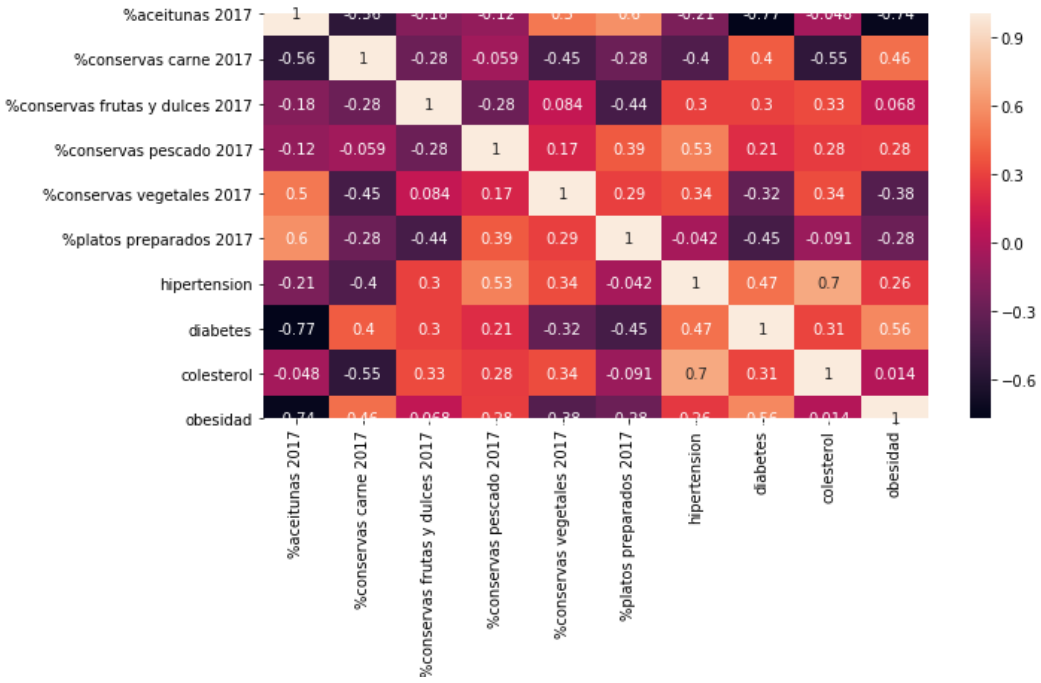


**Imagen 13. Tabla correlaciones entre enfermedades y categorías de la sección conservas. Fuente: elaboración propia**

Si se seleccionan aquellas categorías con un coeficiente de Pearson superior a 0.4 para cada enfermedad, para hipertensión se obtienen las categorías conservas de pescado (0.53) y conservas de carne (-0.4). Para el colesterol**,** la categoría conservas de carne (-0.55). Para la obesidad las categorías aceitunas (-0.74) y conservas de carne (0.46). Por último, para la diabetes, las categorías aceitunas (-0.77), platos preparados (-0.45) y conservas carne (0.4).

- **Correlaciones en las categorías de leche y batidos**

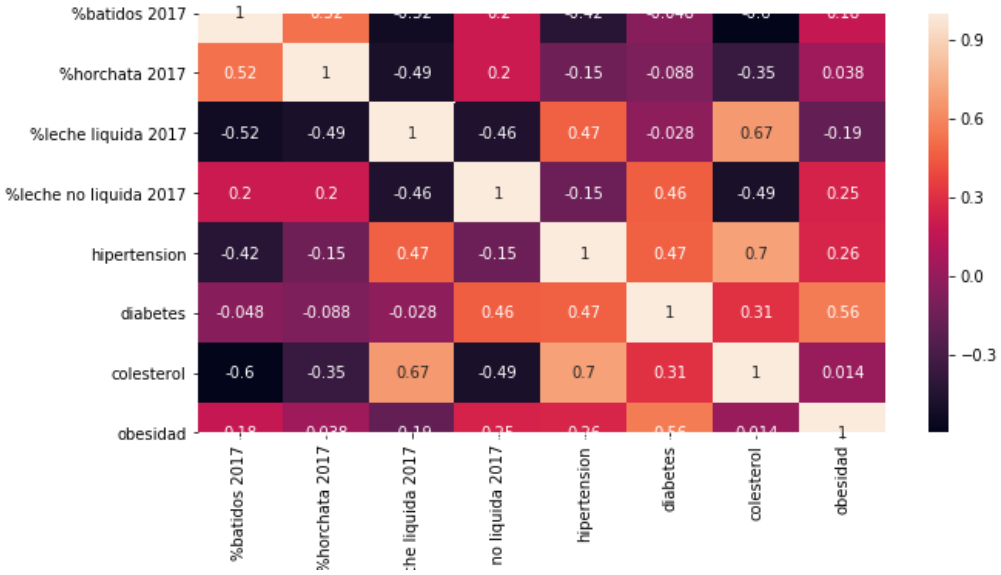




**Imagen 14. Tabla correlaciones entre enfermedades y categorías de la sección leche y batidos. Fuente: elaboración propia**

Si se seleccionan las categorías con coeficiente de Pearson superior a 0.4, se obtienen los siguientes resultados. Las categorías seleccionadas para hipertensión son leche líquida (0.47) y batidos (-0.42).Las seleccionadas para el colesterol son leche líquida (0.67), batidos (-0.6) y leche no líquida (-0.49). Para la diabetes solo se puede seleccionar leche no líquida (0.46). Para la obesidad no se encuentra ninguna que cumpla la condición impuesta.

- **Correlaciones en las categorías de charcutería**

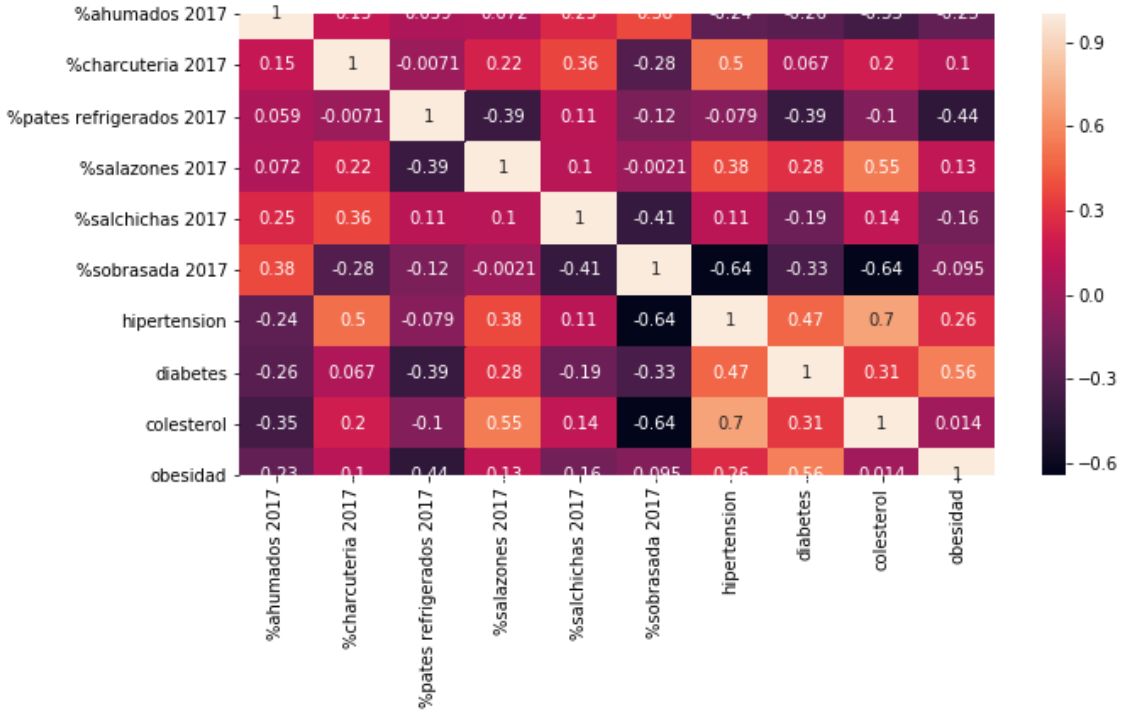


**Imagen 15. Tabla correlaciones entre enfermedades y categorías de la sección charcutería. Fuente: elaboración propia**

Si se seleccionan las categorías con un coeficiente de Pearson superior a 0.4, se obtienen los siguientes resultados. Para la hipertensión se seleccionan las categorías, sobrasada (-0.64) y charcutería (0.5). Para el colesterol, son seleccionadas las categorías, sobrasada (-0.64) y salazones (0.55). Para la obesidad se selecciona la categoría pates refrigerados (-0.44). Finalmente, para la diabetes no se puede seleccionar ninguna categoría.

- **Correlaciones en las categorías de congelados**

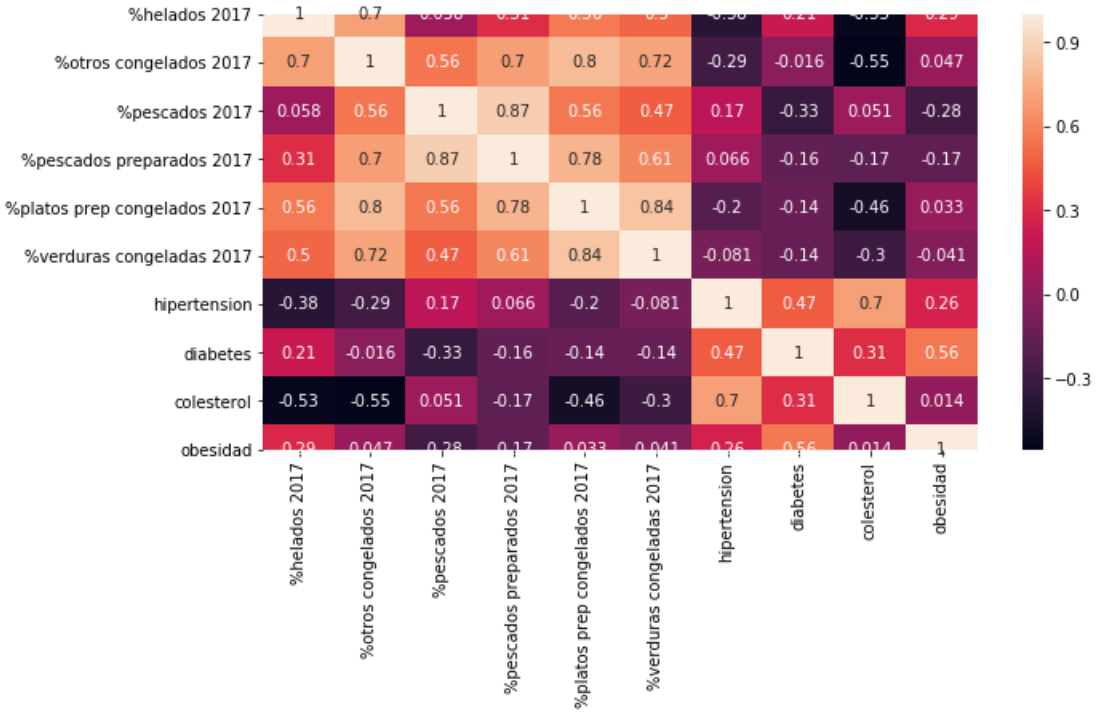


**Imagen 16. Tabla correlaciones entre enfermedades y categorías de la sección congelados.**

**Fuente: elaboración propia**

Los valores del coeficiente de Pearson superiores a 0.4 se obtiene solamente para el colesterol. Las categorías con estos valores son otros congelados (-0.55), helados (-0.53) y platos preparados (-0.46) con el colesterol, respectivamente.

- **Correlaciones en las categorías de derivados lácteos**

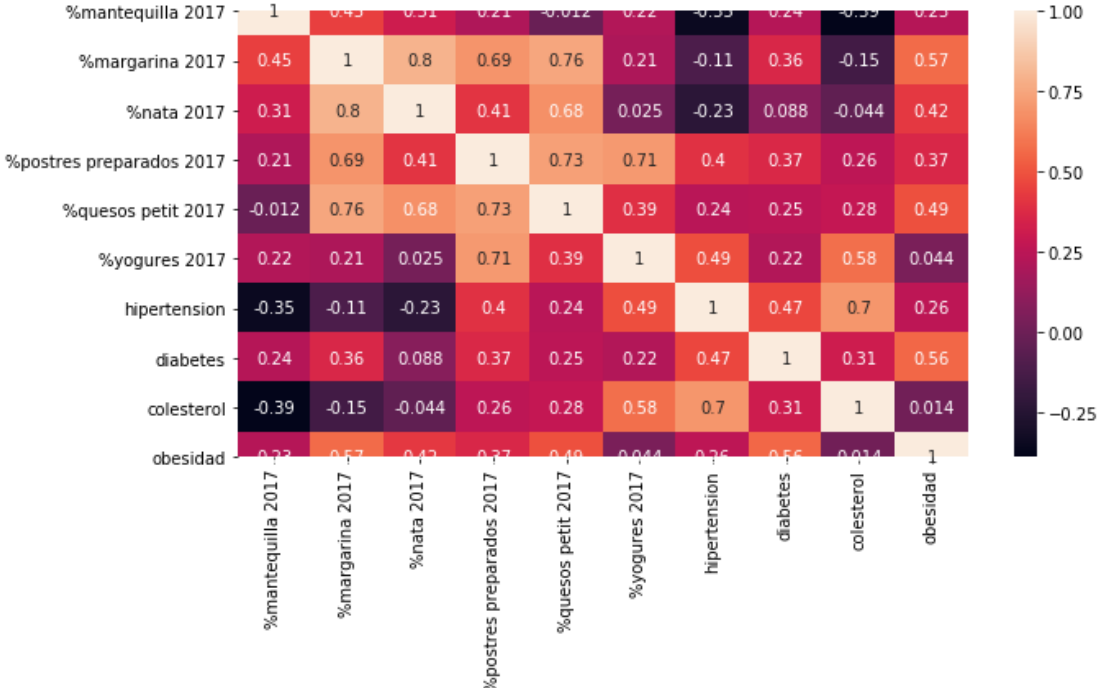




**Imagen 17. Tabla correlaciones entre enfermedades y categorías de la sección derivados lácteos.**

**Fuente: elaboración propia**

De esta tabla se pueden seleccionar las siguientes categorías. Para la hipertensión, yogures (0.49) y postres preparados (0.4). Para el colesterol, yogures (0.58). Y para la obesidad margarina (0.57)) y quesos petit (0.49).

- **Correlaciones en las categorías de platos preparados**

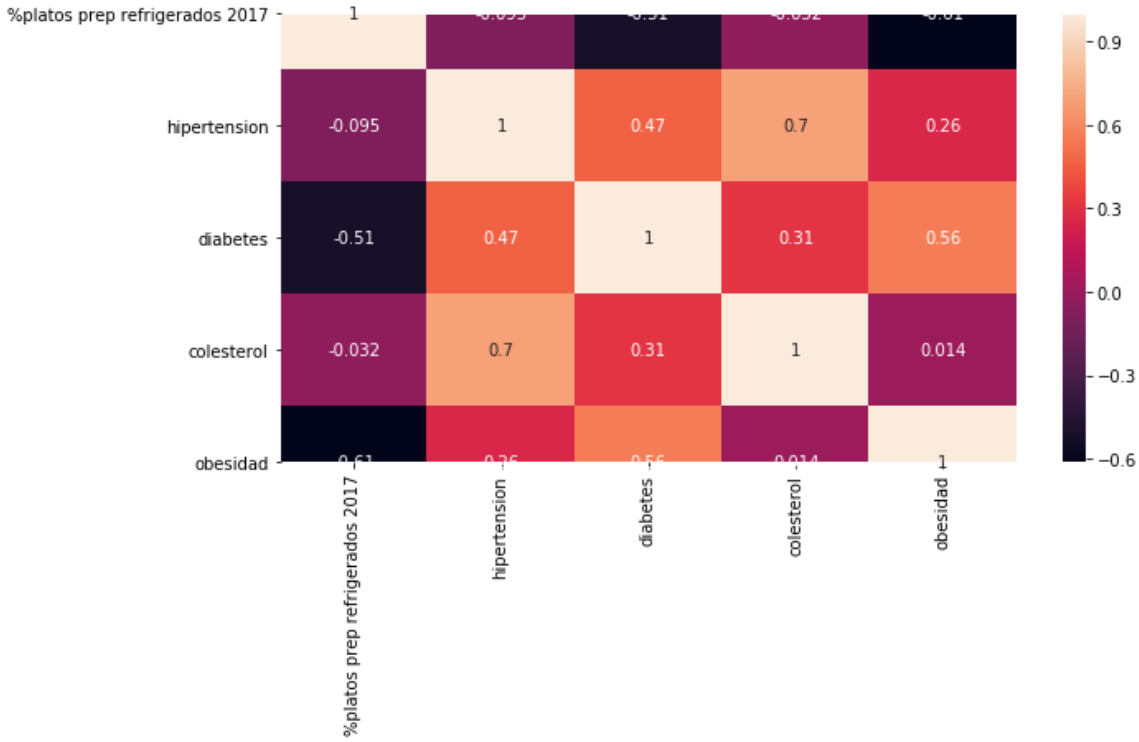


**Imagen 18. Tabla correlaciones entre enfermedades y categorías de la sección platos preparados. Fuente: elaboración propia**

Solamente hay una categoría en este sector y tiene un valor de Pearson superior a 0.4 en las correlaciones con diabetes (-0.51) y obesidad (-0.61).

- **Correlaciones en las categorías de quesos**

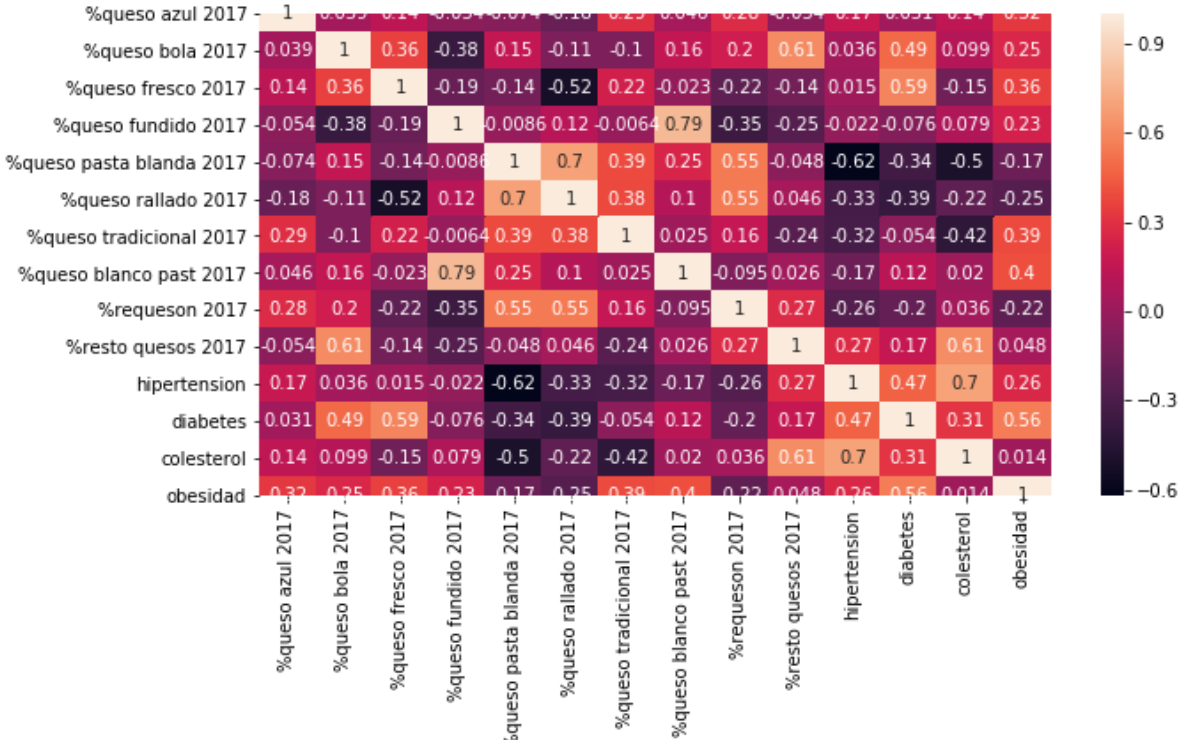


**Imagen 19. Tabla correlaciones entre enfermedades y categorías de la sección quesos. Fuente: elaboración propia**

De esta tabla se pueden seleccionar las siguientes categorías. Para la hipertensión, queso pasta blanda (-0.62). Para el colesterol, queso pasta blanda (-0.5), queso tradicional (-0.42) y resto de quesos (0.61). Y para la diabetes, queso de bola (0.49) y queso fresco (0.59).

A continuación, se muestra una tabla resumen donde aparecen todas las categorías que se han ido seleccionando. Las columnas corresponden a cada enfermedad crónica, las filas separan las categorías de cada sector y el color en el que se escribe la categoría hace referencia a si se trata de una relación positiva o negativa, siendo el verde el color elegido para las correlaciones negativas y el rojo el elegido para las positivas.

| Hipertensión | Colesterol | Diabetes | Obesidad |
|---|---|---|---|
| Aperitivos (-)<br>Cacao (+)<br>Legumbres (+)<br>Azúcar (+) | Aceite (+)<br>Cacao (+)<br>Infusiones (+)<br>Chocolates (+)<br>Aperitivos (-)<br>Salsas (-) | bollería industrial (-) | alimentación infantil (+)<br>dietéticos (-) |
| Aguas (-)<br>Espumosos (-)<br>Zumos (-) | Aguas (-)<br>Refrescos (-)<br>Bebidas alcohólicas (-)<br>Zumos (-) | | Refrescos (+)<br>Espumosos (-) |
| Conservas Pescado (+)<br>Conservas Carne (-) | Conservas Carne (-) | Aceitunas (-)<br>Conservas Carne (+) | Aceitunas (-)<br>Platos preparados (-)<br>Conservas Carne (+) |
| Leche líquida (+)<br>Batidos (-) | Batidos (-)<br>Leche líquida (+)<br>Leche No Líquida (-) | Leche no líquida (+) | |
| Charcutería (+)<br>Sobrasada (-) | Sobrasada (-)<br>Salazones (+) | | Patés refrigerados (-) |
| | Helados (-)<br>Otros congelados (-) | | |
| Yogures (+)<br>Postres preparados (+) | Yogures (+) | | Margarina (+)<br>Queso petit (+) |
| | | Platos prep refr (-) | Platos prep refr (-) |
| Queso pasta blanda (-) | Queso pasta blanda (-)<br>Resto quesos (+) | Queso de bola (+)<br>Queso fresco (+) | |

**Tabla 1. Tabla resumen de las correlaciones entre categorías de alimentos y las enfermedades. Fuente: elaboración propia**

*Regresiones lineales entre el % de gasto y las enfermedades crónicas*

Los modelos de regresión se han creado para comprobar esa relación, positiva o negativa, que se apreciaba entre determinadas categorías y las cuatro enfermedades estudiadas. Puesto que la población entre unas comunidades autónomas y otras varía mucho, estas regresiones se han calculado teniendo en cuenta de forma diferente el peso de cada comunidad según su población.

Para no mostrar cada uno de los modelos de regresión creados, a continuación, se muestra una gráfica a modo de ejemplo.

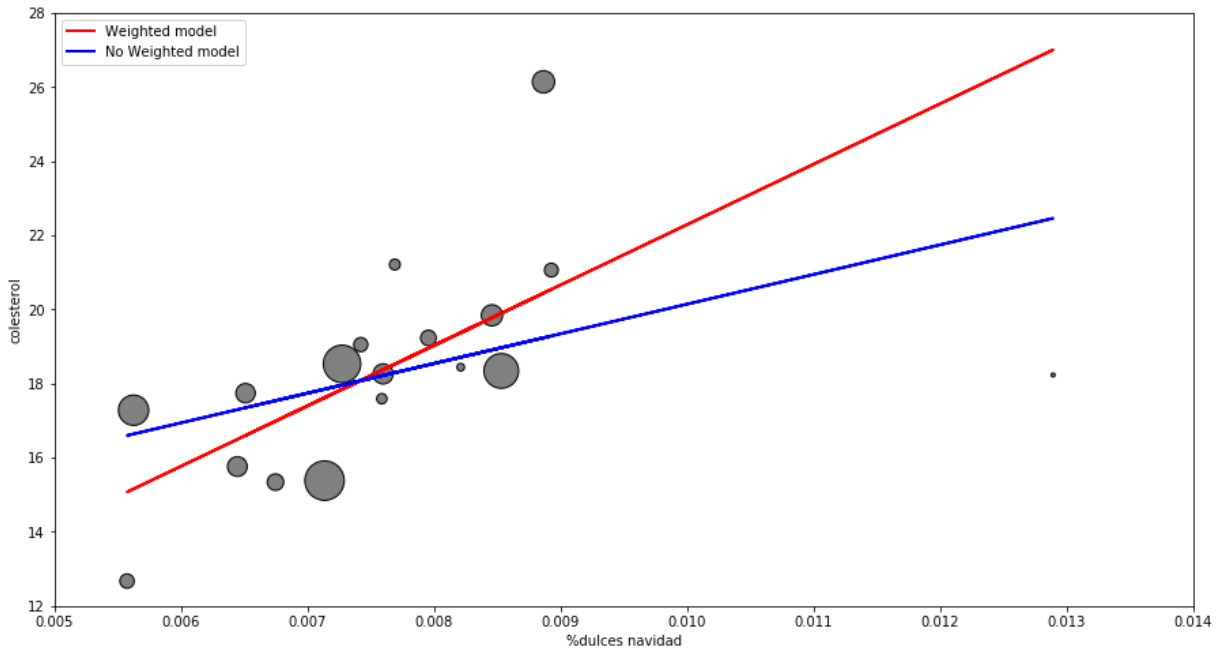


**Imagen 20. Modelo de regresión lineal entre el gasto en dulces de navidad y el colesterol. Fuente: elaboración propia**

Cada uno de los puntos del gráfico corresponde a una comunidad autónoma. Aquellos puntos de tamaño mayor representan las comunidades con mayor población y al contrario ocurre con aquellos puntos más pequeños. En este gráfico aparecen tanto la recta de regresión que aparecería sin tener en cuenta el peso de la población (azul), como la recta cuando sí que se tiene en cuenta (roja). Se puede observar que, en este caso, la recta con pesos tiene mayor pendiente positiva porque en la parte superior hay comunidades con más peso y la comunidad más desplazada a la derecha, que queda por debajo de la recta de regresión, tiene un peso muy pequeño.

A continuación, se van a mostrar una serie de gráficos que recogen los resultados de la realización de las diferentes regresiones lineales realizadas entre cada categoría de alimentos y cada enfermedad. Se trata de gráficos de barras horizontales centradas en el 0. El hecho de que la barra crezca hacia la izquierda significará que la relación entre esa categoría y la enfermedad estudiada es negativa y, al contrario, en el caso de que se genere una barra hacia la derecha. Además, están pintadas de forma diferente, en azul las correlaciones negativas y en rojo las positivas

En el eje Y, cada barra corresponde a una categoría de alimentos diferente. En el eje X se muestran los valores de la pendiente de la recta de regresión. Cuanto mayor sea la pendiente de la recta, mayor será la relación entre ambas variables.

Se mostrará un gráfico para cada enfermedad. En este únicamente aparecen las categorías de alimentos en cuyo modelo de regresión, el P-valor ha sido significativo, es decir, menor de 0.01.

- **Resultados modelos de regresión con la hipertensión**

En el siguiente gráfico se muestran las categorías con mayores valores de correlación, tanto positiva como negativa

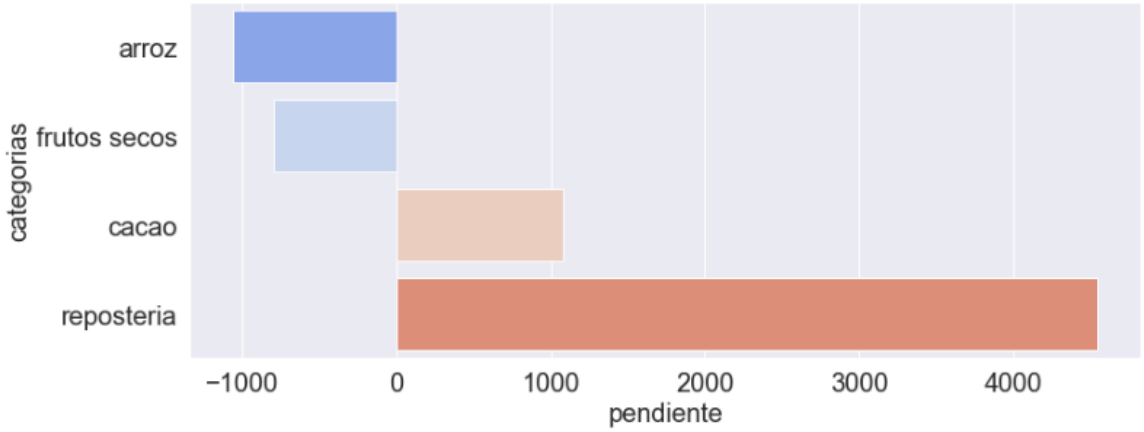


**Imagen 21. Categorías con p_valor significativo para la hipertensión. Fuente: elaboración propia**

Las categorías que tienen relación negativa con la hipertensión son el arroz y los frutos secos. Estas son dos categorías de alimentos que se podrían asociar a cualquier dieta equilibrada. Sin embargo, las categorías cuya relación es positiva son el cacao y la repostería, ambas con alto contenido en grasas y azúcares.

- **Resultados modelos de regresión con el colesterol**

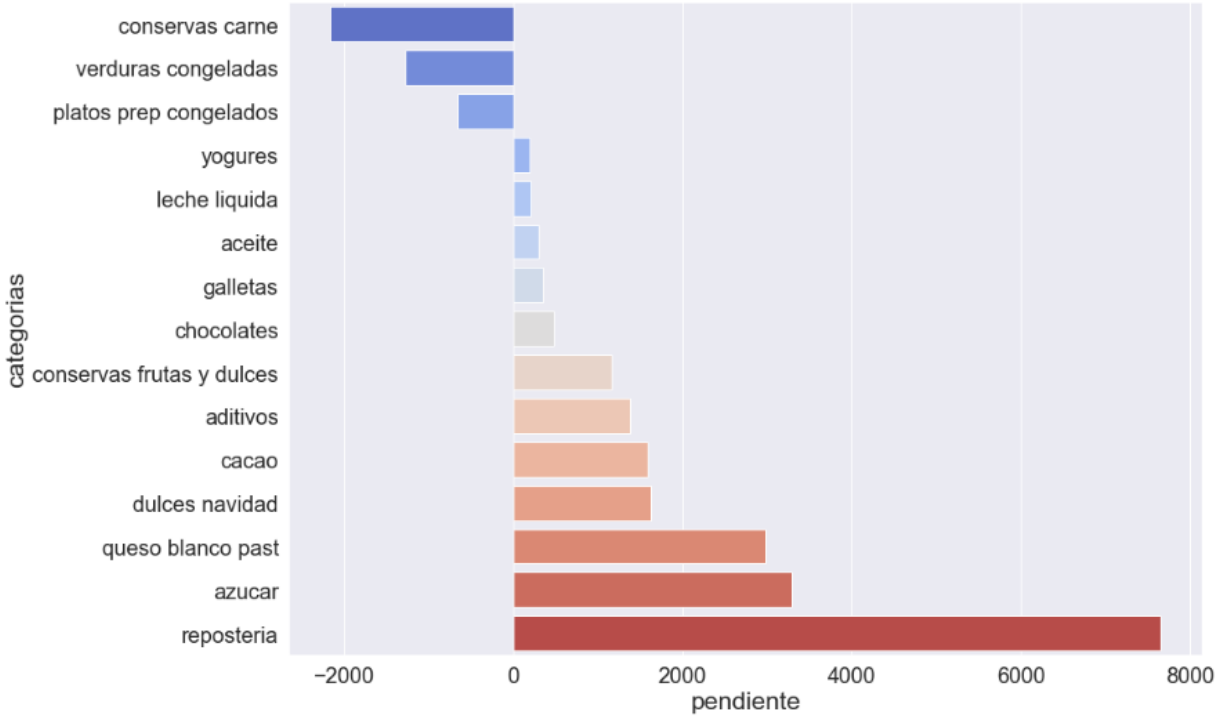


**Imagen 22. Categorías con p_valor significativo para el colesterol. Fuente: elaboración propia**

Los alimentos con correlación negativa son conservas de carne, verduras congeladas y platos preparados congelados. Son categorías con alimentos propios de una dieta equilibrada. Por otro lado, los alimentos con correlación positiva se pueden agrupar en alimentos dulces y de repostería, con alto contenido en grasas y azúcares; lácteos y aceite, ambos con alto contenido en grasas (aunque no sean grasas perjudiciales).

- **Resultados modelos de regresión con la diabetes**

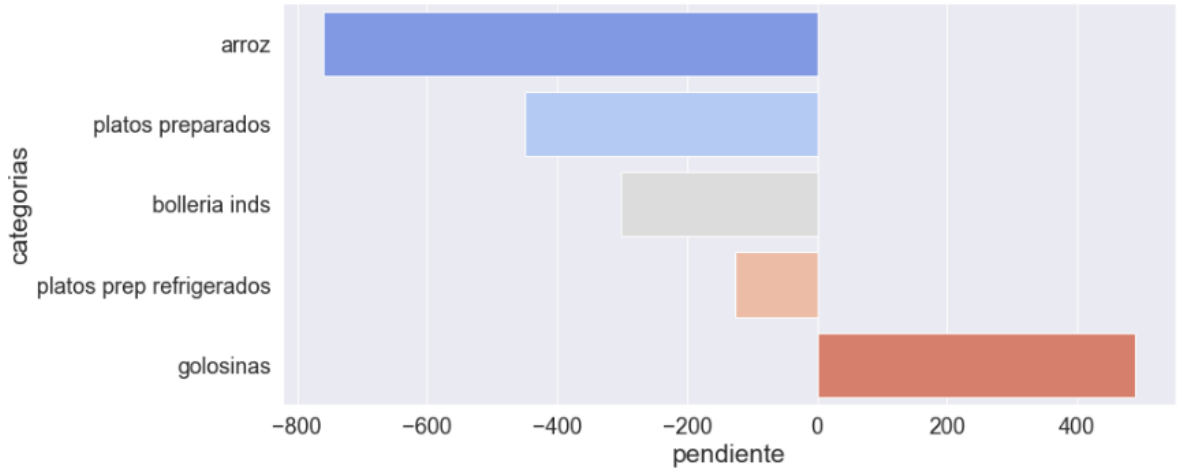


**Imagen 23. Categorías con p_valor significativo para la diabetes. Fuente: elaboración propia**

Las categorías con correlación negativa son el arroz, platos preparados, bollería industrial y platos preparados refrigerados. De estos, excepto la bollería industrial, se puede decir que son alimentos que se incluyen en la mayoría de las dietas equilibradas. En el caso de la bollería industrial, es más difícil saber por qué tiene una correlación negativa significativa con la diabetes sabiendo que es perjudicial para la misma. Se podría dar la explicación de que aquellas comunidades donde hay más porcentaje de personas diabéticas invierten menos gasto en estos productos puesto que no los pueden consumir. Sin embargo, sería necesario un estudio sobre una población más concreta, así como un mejor análisis causa-efecto para poder determinar si tal relación es cierta.

En cuanto a las categorías con correlación positiva se encuentran las golosinas. El hecho de que bollería industrial sea correlación negativa y golosinas positiva parece algo contradictorio. Sin embargo, en la categoría de golosinas se incluyen también chicles y caramelos con posibilidad de que sean sin azúcar. Como no se pueden separar en cada producto, no es posible saber cuál de los alimentos dentro de la categoría de golosinas es el que tiene esta correlación positiva con la diabetes.

- **Resultados modelos de regresión con la obesidad**

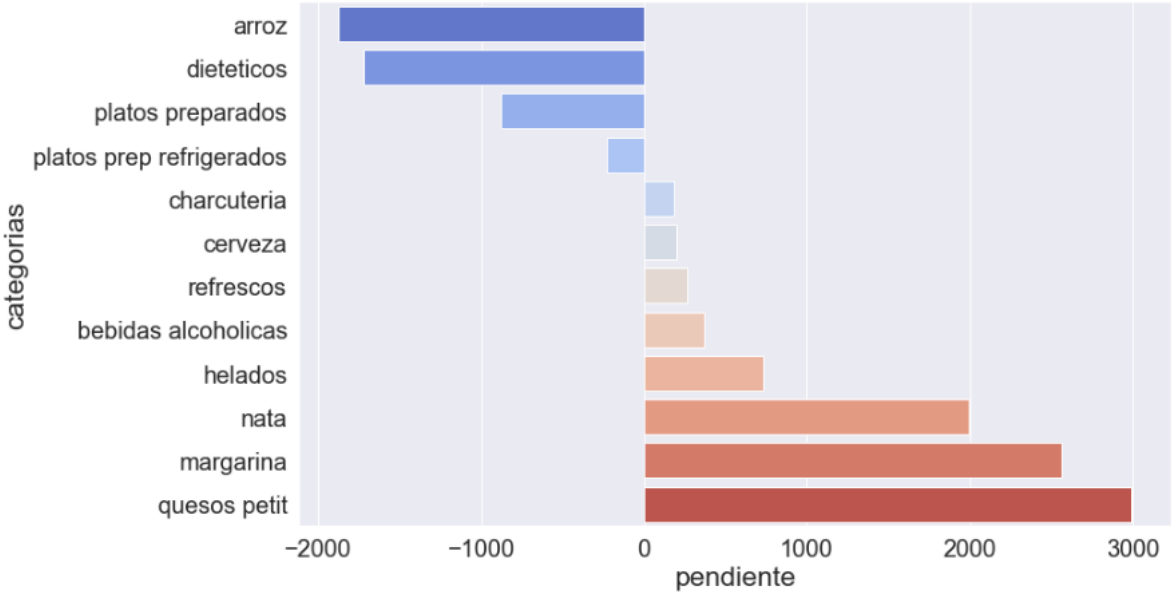


**Imagen 24. Categorías con p_valor significativo para la obesidad. Fuente: elaboración propia**

Las categorías con correlación significativa negativa son: arroz, dietéticos, platos preparados y platos preparados refrigerados. Se observa que excepto los dietéticos, son categorías que ya habían salido en otras correlaciones negativas con enfermedades y que son adecuados para una dieta equilibrada. En el caso de los dietéticos se podría pensar que aquellas categorías que invierten más dinero en alimentos dietéticos son aquellas más preocupadas con el peso y, por lo tanto, con menor tasa de obesidad.

En cuanto a los alimentos con correlación positiva se encuentran tanto alimentos como bebidas que no dormán parte de las comidas principales, sino que se les puede considerar como alimentos de picoteo o de entre horas. Son alimentos con alto contenido en alcohol, grasas y azúcar.

### Resultados finales

A continuación, se exponen algunos resultados generales tras ver las correlaciones entre cada enfermedad y el gasto destinado a las diferentes categorías de alimentos por comunidad. Se ha ido observando que en general, muchas de las categorías se han ido repitiendo en las distintas enfermedades.

En primer lugar, como resultado del análisis del gasto destinado a productos BIO, se observa que la proporción del gasto en el supermercado destinado a alimentos ecológicos muestra una correlación significativa y negativa con el porcentaje de personas con obesidad en una comunidad autónoma.

En segundo lugar, en cuanto a las categorías que presentan una correlación positiva con las enfermedades en general, se puede observar que son alimentos con alto contenido en grasas e hidratos de carbono (azúcar). Abundaban los alimentos destinados a desayunos, meriendas y picoteo, así como bebidas alcohólicas o golosinas, ambos alimentos no recomendados en una dieta saludable.

Por último, en cuanto a las categorías de alimentos con correlación negativa con las enfermedades, se ha observado que había más variedad de alimentos, incluyendo verduras y platos pertenecientes a una de las comidas principales del día. Estos alimentos contienen, entre otros, proteínas y vitaminas, algo que no se ha visto entre las categorías con correlación positiva.

## CONCLUSIONES Y RECOMENDACIONES

En este apartado se muestran las conclusiones obtenidas tras la realización del proyecto. En primer lugar, se van a detallar las conclusiones que se pueden extraer de los resultados que se han ido obteniendo en las diferentes etapas del análisis. A continuación, se expondrán las conclusiones que se han podido obtener de todo el proceso, desde el planteamiento del proyecto hasta la obtención de los resultados. Por otro lado, también se darán una serie de recomendaciones de cara, tanto a posibles proyectos con objetivos similares al planteado, como a un posible proyecto que se realice como continuación de este (podría ser por parte de la propia ONG).

### Conclusiones del análisis

Como se ha visto en el apartado anterior, tras todo el proceso de análisis, se han obtenido tres resultados principales. Estos resultados reafirman lo dicho por los expertos acerca de una alimentación saludable y, es que, se ha podido comprobar que en las categorías en las que abundan alimentos con altas cantidades de grasas y azúcares son aquellas que muestran correlaciones positivas con alguna de las enfermedades crónicas estudiadas. Además, como se ve en el primer resultado, se ha podido comprobar que la alimentación ecológica también influye en el desarrollo de enfermedades, concretamente en el aumento de personas con obesidad.

Sin embargo, al determinar concretamente la relación que guardan las enfermedades y el gasto destinado a las diferentes categorías de alimentos, surge la siguiente pregunta, ¿es la buena alimentación la causa de que estas personas no tengan enfermedades crónicas o es el interés de estas en no padecer alguna de las enfermedades lo que los lleva a consumir alimentos saludables? Esta pregunta gana un interés especial cuando se hacer referencia a productos ecológicos, ya que en esta categoría se pueden incluir tanto alimentos considerados saludables (frutas o verduras) como otros poco saludables (bollería de producción ecológica). Por ejemplo, en los antecedentes se podían

ver los resultados de un estudio que mostraba las características de las personas dentro de la población española que más interesadas estaban en consumir productos ecológicos. Estas personas estaban ya preocupadas por cuidar su salud, lo que lleva a pensar que además de comprar productos ecológicos, llevarían una dieta saludable y se cuidarían en otros ámbitos como el deporte, el sueño, etc. En cualquiera de los casos, no queda duda de que, en el camino para no sufrir de alguna de estas enfermedades, llevar una dieta saludable es uno de los factores clave.

Una conclusión más que puede ser extraía de este análisis tiene relación con su escasa profundidad y, es la imposibilidad de hacer recomendaciones a la población a cerca de los beneficios de consumir unos productos en lugar de otros. Se ha podido ver que los alimentos que comúnmente se incluyen en dietas equilibradas con variedad de componentes nutricionales (hidratos como el arroz, legumbres, platos con proteínas…) son aquellos en los que más invertían las comunidades con menor porcentaje casos de alguna de las enfermedades crónicas, pero no se ha podido determinar qué productos concretos o qué tipo de dieta era la seguida por esta comunidad.

Por último, una de las características de la base de datos de alimentación, era la exclusión de alimentos frescos, frutas, verduras, cernes, pescados… Aunque estos alimentos se podían encontrar en productos como las conservas de verduras o los platos preparados de carne, los productos frescos son la base de la alimentación en algunos de los hogares en los que se tiene especial cuidado con la dieta y que se han quedado fuera del análisis.

### Conclusiones del proyecto general

Con la realización de este proyecto, no solo se pretendía demostrar la relación existente entre la alimentación y las enfermedades crónicas, sino que también se buscaba analizar el potencial del uso de datos ya creados, principalmente por grandes empresas, para obtener información de interés general para la población. En este sentido, se han podido extraer diferentes conclusiones que se indicarán a continuación.

Por una parte, el hecho de que cada empresa seleccione y almacene los datos de un modo concreto, dificulta el poder relacionarlos con otras bases de datos. Además, si, como ha ocurrido en este proyecto, se deben mezclar con datos de fuentes oficiales, se observa un gran desacoplo entre las fechas con las que trabajar. En este proyecto, los datos más recientes que recogía el INE eran pertenecientes a 2017, mientras que 2017 era el primer año que guardaba la empresa (se van borrando y actualizando).

Por otra parte, se ha detectado gran dificultad para encontrar datos, especialmente en el caso de las enfermedades crónicas, lo que evidencia la necesidad de que no sean solo las empresas bajo su propio interés las que recolecten y generen bases de datos precisas y con información bien detallada, sino que haya entidades a nivel nacional que trabajen en recolectar los datos que se va generando y los almacene en grandes bases de datos desde las que se pueda obtener información más precisa que la que se puede encontrar en la actualidad. Por ejemplo, en la división temporal se encontraban grandes diferencias entre una y otra base de datos. En el caso de la base de datos la información estaba almacenada por semanas, lo que permitía un análisis de la evolución mucho más amplio que en el caso de la base de datos de las enfermedades crónicas en la que aparecía un único dato de todo 2017.

Sin embargo, a pesar de que los datos de Nielsen estaban más detallados y abrían permitido un análisis más profundo, también hay que destacar que, entre las premisas de las que se partía a la hora de utilizar los datos era que no se podían llegar a analizar productos de marcas concretas o de poblaciones concretas. Con esta información, ya se puede ver el problema que tienen las empresas a la hora de ceder la información y, es que, no pueden decantarse o perjudicar a alguna de las empresas que estén incluidas entre sus datos puesto que podrían retirarles el permiso para seguir recolectándolos. Este conflicto de interese limita la obtención de resultados útiles y veraces para la población. Además, si obtener los datos es una inversión para las empresas, es de esperar que muchas de ellas no quieran ceder sus datos, o al menos aquellos que más beneficiosos les sean

A pesar de las limitaciones que se han ido mostrando a lo largo de este documento, se ha podido apreciar también que el uso de técnicas que trabajen con cantidades masivas de datos tiene un gran potencial para la realización de futuros proyectos. Y, es que, si con solo 18 datos se ha podido corroborar la relación entre los diferentes tipos de alimentos y las enfermedades crónicas, con unos datos mejor seccionados, se podrían comprender mejor muchas de las preguntas que se planteaban en el apartado de antecedentes.

Por último, como conclusión general de los resultados obtenidos en el proyecto, se observa la amplitud y complejidad que conlleva tratar de discernir las causas de las enfermedades crónicas y, es que es muy limitado analizar únicamente el gasto en alimentación sin observar otros factores como el nivel social, la edad, la actividad física, la educación, etc. De esta complejidad se corrobora la necesidad de contar con bases de datos amplias sobre los hábitos de la población para poder establecer relaciones de causa y efecto y poder comprender mejor este fenómeno.

### Recomendaciones

Entre los objetivos del trabajo se encontraba el de ser un primer acercamiento a proyectos en este ámbito, por lo que, tras haberlo realizado y obtenido las conclusiones arriba descritas, se procederá dar una serie de recomendaciones, tanto para futuros proyectos que continúen investigando ese tema (especialmente por parte de So Good Data), como de cara a la sociedad para que otros proyectos similares puedan ser llevados a cabo con resultados más precisos.

En primer lugar, es necesario obtener la mayor cantidad de información acerca del tema sobre el que se vaya a realizar el proyecto. Para ello, no solo es importante realizar una amplia búsqueda en diferentes páginas científicas de internet, sino contactar con expertos que estén trabajando en este tema. Esto, permitirá obtener una visión completa y actual del problema y enfocarlo de tal forma que los resultados puedan ser interesantes también para la comunidad científica. Esto da lugar a que se recomiende, no solo contactar en las primeras fases del desarrollo del proyecto, sino a llevar una colaboración más estrecha entre la ONG y los expertos en el tema, en este caso en las enfermedades crónicas y su relación con la alimentación. Esto ayudará a tener una opinión experta mientras que se van obteniendo los resultados y se podrá ir adaptando según las posibilidades del análisis de la manera que más útil e interesante les parezca a estos expertos.

Por otro lado, es muy importante realizar una amplia búsqueda de datos, y puesto que seguramente, en las fuentes oficiales no se encontrarán los datos buscados, sería de gran interés poder trazar colaboraciones con diferentes empresas o entidades que tengan capacidad para recolectar grandes cantidades de datos. Esto podría limitar el estudio, por ejemplo, delimitando un territorio concreto

donde esos datos pueden ser recolectados, pero permitirá profundizar en los resultados obtenidos, aunque sea solamente para esa población. Posteriormente, sería posible determinar diferentes formas de extrapolarlos al resto de España, si se creyese conveniente.

Otra de las recomendaciones que se pueden ofrecer tras el proyecto, es la necesidad de tratar un mismo problema desde diferentes perspectivas. En este caso concreto, ya se han citado diferentes factores de riesgo que no se han incluido en el estudio y que, por lo tanto, limitan los resultados de este. Por ello, es importante ampliar los aspectos que se estudian de un mismo problema lo máximo posible.

Finalmente, una recomendación general para los proyectos de análisis datos que estudien comportamientos sociales, sería la de no ser cortoplacista. El análisis de datos es un campo que ha llegado para quedarse y que se irá ampliando con el tiempo, por lo que no se debe de dudar a la hora de invertir en él, no solo dinero, sino tiempo. Por ejemplo, establecer relaciones con diferentes sectores de la sociedad con los que se puedan crear vínculos a largo plazo y que colaboren con su conocimiento en los resultados que va obteniendo la ONG. Aunque un proyecto concreto debe de tener una duración acotada, este será parte de una investigación más amplia que sí que tendrá más duración en el tiempo.

## REFLEXIÓN CRÍTICA

El análisis de datos es un paso imprescindible que se debe dar a la hora de tomar decisiones y poder mejorar en un aspecto determinado. Si en nuestro día a día tiene gran importancia el hecho de que, ante un problema incluso personal, nos paremos y analicemos el origen y las posibles soluciones antes de tomar una decisión, cuando se trata de decisiones que influyen en la vida de un gran grupo de personas, parece totalmente necesario.

Todo esto no es algo nuevo, sino que se trata de un campo con un largo recorrido. Sin embargo, el gran volumen de datos que ahora somos capaces de almacenar y analizar, sí que se trata de una novedad. En el ámbito social, los datos que se analizan se obtienen principalmente a través de encuestas a la población. Comparando la cantidad de datos que pueden ser obtenidos a través de encuestas con la cantidad de datos que recogen las empresas de cualquier movimiento de la población que les resulte de interés, se aprecia una enorme diferencia.

Analizar grandes cantidades de datos, presenta una ventaja y es la precisión de los modelos que se pueden generar y el nivel de conocimiento que es posible obtener. Vivir en una sociedad en la que rige el factor económico hace que todo aquello que ayude a aumentar las ganancias de una empresa, sea bienvenido. Sin embargo, en el sector social, recoger estas grandes cantidades de datos y analizarlas posteriormente, no representa un beneficio económico para un actor de la sociedad concreto, lo que hace difícil que se invierta en esto.

La diferencia de intereses y capital disponible que se aprecia entre las empresas y las entidades sociales hace que se plantee la cuestión de si contribuirá a aumentar la desigualdad entre ambos campos (empresa y sector social). Por ello, se ve necesario aumentar la concienciación y la implicación de diversos sectores que reclamen más inversión en estos proyectos que ya demuestran su eficacia en otros campos.

Por otro lado, a pesar de las ventajas que presenta el Big Data, hay una desventaja derivada de estos grandes volúmenes de datos. Esta es la pérdida de control sobre los resultados. El hecho de utilizar algoritmos muy complejos y cantidades de datos que no podemos comprender completamente hace que tengamos que demostrar una confianza “ciega” en los resultados. Además, cuando se trata de temas sociales, tenemos que recordar que las causas de un problema no son simples, sino que pueden presentarse multitud de causas a la vez.

En el caso concreto del proyecto, se veía una limitación a la hora de aceptar los resultados y es, precisamente, el hecho de que solo se ha tenido en cuanta la alimentación de las personas. Existen numerosos y diversos aspectos que influyen en la salud de una persona, genética, educación, entorno, actividad física, deporte… Además, en el caso de la alimentación, también habría que preguntarse qué es lo que ace que una persona o familia invierta un porcentaje determinado de su gasto en unos productos y no otros. Podría haber factores económicos que limiten a la población a la hora de elegir los productos de la compra en perjuicio de la salud. O podrían ser factores sociales y educativos los que hagan que unas personas sean más sensibles en cuanto a la calidad del alimento a la hora de comprarlo.

Además, en proyectos como el que se ha realizado, es muy importante contar con la visión de expertos en el ámbito concreto que se esté analizando. Quizás, si se tomaran varios de los factores de riesgo de sufrir alguna de esas enfermedades, los expertos con los que se ha contactado también serían insuficientes. Contar con la visión de expertos que puedan dar información sobre factores sociales que estén relacionados con el aumento de casos de enfermedades crónicas, habría sido una buena forma de obtener un análisis más completo. No solo la escasez de datos con los que finalmente se ha podido trabajar, sino también la poca amplitud desde la que se ha enfocado hace que los resultados resulten poco interesantes.

Por todas estas razones, se cree que el proyecto es un buen ejemplo sobre el potencial que tiene el empleo de técnicas de data science en problemáticas sociales, pero se encuentra muy limitado en cuanto a los resultados obtenidos. Los resultados se pueden ver como una comprobación de que los datos empleados se pueden utilizar y los resultados se obtienen conforme a lo ya sabido, es decir, no inducen a errores, pero no se puede obtener nuevo conocimiento de los resultados del proyecto.

Finalmente, se cree que no debe de ser un proyecto aislado, sino que debe iniciar una línea de trabajo e la cual se creen lazos fuertes de colaboración entre expertos en los diferentes factores que puedan estar aumentado los casos de enfermedades crónicas de modo que puedan orientar a los analistas de datos a buscar resultados novedosos. Ir profundizando con diferentes análisis y desde diferentes enfoques será lo que permita comprender esta situación y aportar diferentes posibles soluciones.

## BIBLIOGRAFÍA


Amat, J. (2016) *Correlación lineal y regresión lineal simple.* Recuperado de: https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal#Inferencia_mediante_regresi%C3%B3n_lineal_Significancia_e_intervalo_de_confianza_para_(beta_0)_y_(beta_1)

Ambit (2019) *Qué es la huella digital y cuál es su importancia.* Recuperado de https://www.ambit-bst.com/blog/huella-digital-importancia

Ayuso, M. (2015). "Así comemos los españoles: ocho gráficas que explican por qué engordamos". En Alma Corazón y Vida. Recuperado de https://www.elconfidencial.com/alma-corazon-vida/2015-11-03/asi-comemos-los-espanoles-seis-graficas-que-explican-por-que-engordamos_1081006/

Bastis Consultores (2021) *Análisis univariante.* Recuperado de: https://online-tesis.com/analisis-univariante/

Echanove, M y Puértolas, S. (2019) "Healthy foods & brands" Google https://drive.google.com/file/d/101WQIRWMJhaSSxCcG5cH4zzN0RgRgu7r/view

Fundación Mapfre (s.f.) *¿Cuánta información se genera y almacena en el mundo?* Recuperado de https://www.fundacionmapfre.org/blog/cuanta-informacion-se-genera-y-almacena-en-el-mundo/

García, M. (2021) *¿Qué es el análisis multivariado? Aprende a dominar datos y variables.* Recuperado de: https://www.crehana.com/es/blog/desarrollo-web/analisis-multivariado/

Hyland, C., Bradman, A. et al. (2018) "Organic diet intervention significantly reduces urinary pesticide levels in U.S. children and adults" en Environmental Research nº 171, pp. 568-575. Recuperado de: https://www.sciencedirect.com/science/article/pii/S0013935119300246?via%3Dihub

Instituto Nacional del Cáncer (s.f.) *Definición de enfermedad crónica.* Recuperado de: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/enfermedad-cronica

Instituto Nacional de Estadística (2019) Población que usa Internet (en los últimos tres meses). *Tipo de actividades realizadas por Internet*. Recuperado de: https://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259925528782&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout

Kersting, K., Bauckhage, C. et al. (2016) "Feeding the World with Big Data: Uncovering Spectral Characteristics and Dynamics of Stressed Plants" en Computational Sustainability pp. 99-120. Recuperado de https://link.springer.com/chapter/10.1007/978-3-319-31858-5_6

Lytras, M. y Visvizi, A. (2019) "Big Data Research for Social Scienece and Social Impact" en Sustainability. Recuperado de: https://www.mdpi.com/2071-1050/12/1/180/htm

Marti, A., Calvo, C. y Martínez, A. (2021) "Consumo de alimentos ultraprocesados y obesidad: una revisión sistemática". En Nutrición hospitalaria, vol. 38 nº 1. Recuperado de: https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112021000100177

Maté, C. (2014). “Big data. Un nuevo paradigma de análisis de datos”. En Anales de mecánica y electricidad, Universidad de Comillas. Recuperado de https://www.iit.comillas.edu/docs/IIT-14-153A.pdf

Mayoral Cortes, J. M., Aragonés Sanz, N. et al. (2016) “Las enfermedades crónicas como prioridad de la vigilancia de la salud pública en España” en Gaceta Sanitaria. Número 30, 2016, pp.154-157. Recuperado de: https://scielo.isciii.es/pdf/gs/v30n2/especial.pdf

Miralles Navarro, M. (2020) “Tendencias en alimentación 2020-2021 ¿Qué comeremos en el presente?” Sabuma. Recuperado de https://sabuma.es/consumidores-en-el-centro-responsabilidad-salud-hedonismo-y-plant-based/

López-Cano, L., Restrepo-Mesa, S. et al. (2014). “Etiquetado nutricional, una mirada desde los consumidores de alimentos”. En Perspectivas en Nutrición Humana, volumen 16, nº 2. Recuperado de: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0124-41082014000200003

Organización Mundial de la Salud (2021) *Enfermedades no transmisibles.* Recuperado de: https://www.who.int/es/news-room/fact-sheets/detail/noncommunicable-diseases

Organización Mundial de la Salud (2002) “Informe sobre la salud en el mundo reducir los riesgos y promover una vida sana”. Recuperado de: https://apps.who.int/iris/bitstream/handle/10665/42557/WHR_2002_spa.pdf?sequence=1&isAllowed=y

Organización Mundial de la Salud (2003) “Dieta, nutrición y prevención de enfermedades crónicas” en OMS, Serie de Informes Técnicos 916. Recuperado de: https://www.who.int/nutrition/publications/obesity/WHO_TRS_916_spa.pdf

Redacción Médica (2018) “El 80% del gasto sanitario autonómico se destina a cuatro enfermedades” en Redacción médica. Recuperado de: https://www.redaccionmedica.com/secciones/gestion/el-80-del-gasto-sanitario-autonomico-se-destina-a-cuatro-enfermedades-9971

Swinburn, B., Kraak, V et al. (2019) “The Global Syndemic of Obesity, Undernutrition, and Climate Change: The Lancet Commission report” en The lancet commissions. Volume 393, pp. 791-846. Recuperado de: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)32822-8/fulltext

Tovar, J. (2020). “La esperanza de vida en España, amenazada por enfermedades crónicas y covid”. En EFE salud. Recuperado de: https://www.efesalud.com/esperanza-vida-espana-amenazada-enfermedades-cronicas-covid/

Universidad de Alcalá (s.f.) *5 ejemplos de uso real de data analytics.* Recuperado de: https://www.master-data-scientist.com/ejemplos-master-en-data-analytics/

Varela Moreiras, G., Serrano Iglesias, M. et al. (2015). “Alimentación y sociedad en la España del siglo XXI” en Fundación Mapfre. Recuperado de: https://sennutricion.org/media/Estudio_Alimentaci__n_y_Sociedad_en_la_Espa__a_del_s_XXI.pdf

Yubero, B. (2021) “Si España invirtiera en digitalización se podría incrementar el PIB en un 4.38 por ciento”. En El plural. Recuperado de: https://www.elplural.com/economia/espana-invirtiera-digitalizacion-incrementar-pib-438-ciento_262498102

Zaforas, M. (2016) “¿Qué puede aportar el Big Data al mundo de la medicina?” En Paradigma. Recuperado de: https://www.paradigmadigital.com/dev/puede-aportar-big-data-al-mundo-la-medicina/

## ANEXOS

**Tabla equivalencias unidades de medida de almacenamiento de datos.**

| Decimal | |
|---|---|
| 1 byte (B) | 8 bits |
| 1 kilobyte (KB) | 1000 B |
| 1 megabyte (MB) | 1000 KB |
| 1 gigabyte (GB) | 1000 MB |
| 1 terabyte (TB) | 1000 GB |
| 1 petabyte (PB) | 1000 TB |
| 1 exabyte (EB) | 1000 PB |
| 1 zettabyte (ZB) | 1000 EB |
| 1 yottabyte (YB) | 1000 ZB |

**Fuente: Fundación Mapfre (s.f.)**