TRABAJO DE FIN DE MÁSTER

Caracterización del uso de dispositivos IoT en la población española

Alumna: Cynthia Quintana Tardio

Tutor: Cipriano Quirós Romero

UNIVERSIDAD
COMPLUTENSE
MADRID

Universidad Complutense de Madrid
Máster en Minería de Datos e Inteligencia de Negocio
Curso 2020-2021

Contenido

1. Introducción 1
   1.1. Motivación 1
   1.2. Estado del arte 2
   1.3. Objetivos 4
2. Diseño metodológico 4
   2.1 Tratamiento previo de los datos y selección inicial de variables 5
   2.2 Análisis descriptivo 8
   2.3 Análisis de componentes principales 9
   2.4 Análisis clúster 10
   2.4.1 Análisis clúster jerárquico 11
   2.4.2 Análisis clúster: k-medioids 13
   2.5 Análisis correspondencia simples 14
3. Caracterización de la muestra 16
4. Desarrollo del trabajo: Resultados y discusión. 27
   4.1 Análisis de componentes principales 28
   4.2 Análisis clúster jerárquico 34
      4.2.1 Selección de variables 34
      4.2.2 Análisis clúster jerárquico 35
   4.3 Análisis clúster PAM 39
      4.3.1 Caracterización de los clústeres formados 40
   4.4 Análisis de correspondencias simples 47
      4.4.1 Viabilidad del análisis y número de factores a retener 48
      4.4.2 Caracterización de la sociedad española y el uso de los dispositivos IoT 49
5. Conclusiones 55
   5.1 Limitaciones del trabajo y Futuras líneas de investigación 56
6. Bibliografía 58
7. Anexos 61
   Anexo A: Encuesta INE. 61
   Anexo B: Códigos INE. 65
   Anexo C: Material complementario resultante de la depuración de datos. 65
   Anexo D: Resultados complementarios del análisis clúster. 70

ÍNDICE DE TABLAS

Tabla 1: Ejemplo decodificación valores en el fichero de microdatos. 6
Tabla 2: Variables inicialmente seleccionadas. 6
Tabla 3: Distribución de personas que usan dispositivo IoT según el tipo de dispositivo. 24
Tabla 4: Distribución de selección de las diferentes barreras en el uso de IoT. 27
Tabla 5: Variables utilizadas en el análisis de componentes principales. 28
Tabla 6: Comunalidad final. 31
Tabla 7: Especificidad de las variables. 31
Tabla 8: Matriz de las cargas en las componentes retenidas. 32
Tabla 9: Medidas ajuste del Análisis Componentes Principales. 32
Tabla 10: Resultado de las pruebas de selección de variables. 34
Tabla 11: Variables seleccionadas para realizar análisis clúster. 35
Tabla 12: Características principales clústeres. 36
Tabla 13: Resultados obtenidos para el método de "Enlace medio". 36
Tabla 14: Representación del R-cuadrado versus el R-cuadrado semiparcial. 38
Tabla 15: Características del par de individuos más similar. 39
Tabla 16: Características del par de individuos que más difieren. 40
Tabla 17: Tabla de contingencia variables Clúster y dispHIoT. 48
Tabla 18: Valores esperados estadísticos chi-cuadrado. 48

ÍNDICE DE IMÁGENES

Imagen 1: Distribución de objetos conectados 2017 vs 2019 [8]. 2
Imagen 2: Ejemplo formato archivo microdatos. 5
Imagen 3: Distribución por sexo. 17
Imagen 4: Distribución por edad. 17
Imagen 5: Distribución según la situación laboral. 18
Imagen 6: Distribución según rango de ingresos. 18
Imagen 7: Distribución por nivel de estudios. 19
Imagen 8: Distribución según el uso de internet. 20
Imagen 9: ¿Conoce los dispositivos IoT? 21
Imagen 10: Distribución de personas que conocen dispositivos IoT según nivel de estudios. 21
Imagen 11: Distribución de personas que conocen dispositivos IoT según la edad. 22
Imagen 12: ¿Usa alguno de estos dispositivos IoT? 23
Imagen 13: Porcentaje que usan dispositivo IoT según el tipo de dispositivo. 23
Imagen 14: Distribución de personas según el uso de dispositivos IoT y salario. 24
Imagen 15: Distribución de personas según el uso de dispositivos IoT y estudios terminados. 25
Imagen 16: Distribución de personas según el uso de dispositivos IoT y edad. 26
Imagen 17: Distribución de selección de las diferentes barreras en el uso de IoT. 27
Imagen 18: Representación correlaciones policóricas. 29
Imagen 19: Representación del número de componentes principales. 30
Imagen 20: Proporción de varianza explicada por cada componente principal. 30
Imagen 21: Proporción de la varianza acumulada por cada componente principal. 31
Imagen 22: Representación de las variables y algunos individuos en el espacio de las componentes 1 y 2. 33
Imagen 23: Representación de los valores de los estadísticos Pseudo F y Pseudo T Squared para cada clússter. 37
Imagen 24: Representación del R-cuadrado versus el R-cuadrado semiparcial. 37
Imagen 25: Representación del coeficiente de silueta para cada clúster. 38
Imagen 26: Resumen de características de los medoides de cada clúster. 41
Imagen 27: Distribución de las características en los clústeres. 42
Imagen 28: Representación visual de los clústeres. 42
Imagen 29: Descomposición Inercia y Chi-cuadrado. 49
Imagen 30: Coordenadas, contribuciones parciales y cosenos al cuadrado de las filas. 50
Imagen 31: Gráfico de dispersión para los perfiles fila. 50
Imagen 32: Coordenadas, contribuciones parciales y cosenos al cuadrado de las columnas. 51
Imagen 33: Gráfico de dispersión para los perfiles columna. 51
Imagen 34: Representación de perfiles filas y columnas en los ejes de las dimensiones retenidas. 52

1. Introducción

1.1. Motivación

En los últimos años, los avances de las tecnologías y las comunicaciones (ICTs, por sus siglas en inglés) han impactado positivamente en el desarrollo progresivo de la sociedad. De hecho, con la emergencia del Internet de las Cosas (IoT, por sus siglas en inglés), se están disrumpiendo los comportamientos y las actividades diarias de las personas. Dispositivos anteriormente desconectados, se pueden acceder ahora de forma digital, desde cualquier lugar y en una variedad de dispositivos. Debido a esto, la utilización de dispositivos conectados a internet ha comenzado a despertar el interés de los consumidores por las ventajas y facilidades que conllevan en la mejora de la calidad de vida [1]–[4]. El Internet de las Cosas se puede entender como un ecosistema de una elevada cantidad de dispositivos inteligentes conectados no solo entre sí, sino también al internet, que transmiten y consumen información. Por ejemplo, un concepto incluido dentro del IoT y que ha estado llamando la atención de los consumidores últimamente es el de hogar inteligente, a pesar de no ser nuevo ya que ha sido objeto de debate desde 1980, y que ha evolucionado a partir de la automatización tradicional del hogar [5]. Y es que los productos que traen comodidades al hogar están dejando de ser un lujo para convertirse en una conveniencia como es el caso del control automático de la luz y la calefacción [6].

Es por ello que se puede afirmar que el IoT es sin dudas un mercado con mucho potencial de crecimiento ya que ofrece tecnologías y servicios innovadores que mejoran la experiencia de sus usuarios. Adicionalmente, al proporcionar un acceso fácil y global, los dispositivos inteligentes atraen a más consumidores a participar en dicha tecnología progresivamente. Si se tienen en cuenta datos de informes de analistas, estos aseguran que la cuota del mercado mundial del IoT orientado a consumidores crecerá para 2025 con una tasa de crecimiento anual del 18,95% [7]. También, según *Juniper Research* habrán más de 46 billones de dispositivos IoT conectados en 2021 y, por si fuera poco, *IHS Market* estima que para 2030 habrá una media de 15 objetos conectados por persona.

Adicionalmente, se puede afirmar que el entendimiento y comprensión de los conceptos y tecnologías relacionadas con el IoT, es directamente proporcional a la popularización de los mismos, lo cual resulta consecuentemente, en el aumento de su consumo. En España particularmente, esto ha quedado reflejado en el segundo informe de Telefónica sobre el uso del IoT en el país [8] donde se realizaron 800 encuestas tanto a usuarios de smartphone como a especialistas. El estudio reporta que respecto a 2017, la población que tiene dispositivos conectados ha crecido un 67% en 2019, lo cual es un indicador de que se ha producido un aumento de la penetración del IoT en España. Entre ellos destacan los dispositivos de coches conectados, los dispositivos dentro del hogar como electrodomésticos y bombillas y los relojes inteligentes [8]. En la Imagen 1 se puede apreciar cuáles han sido las diferentes categorías estudiadas, así como el porcentaje de penetración de cada uno para 2017 y 2019.

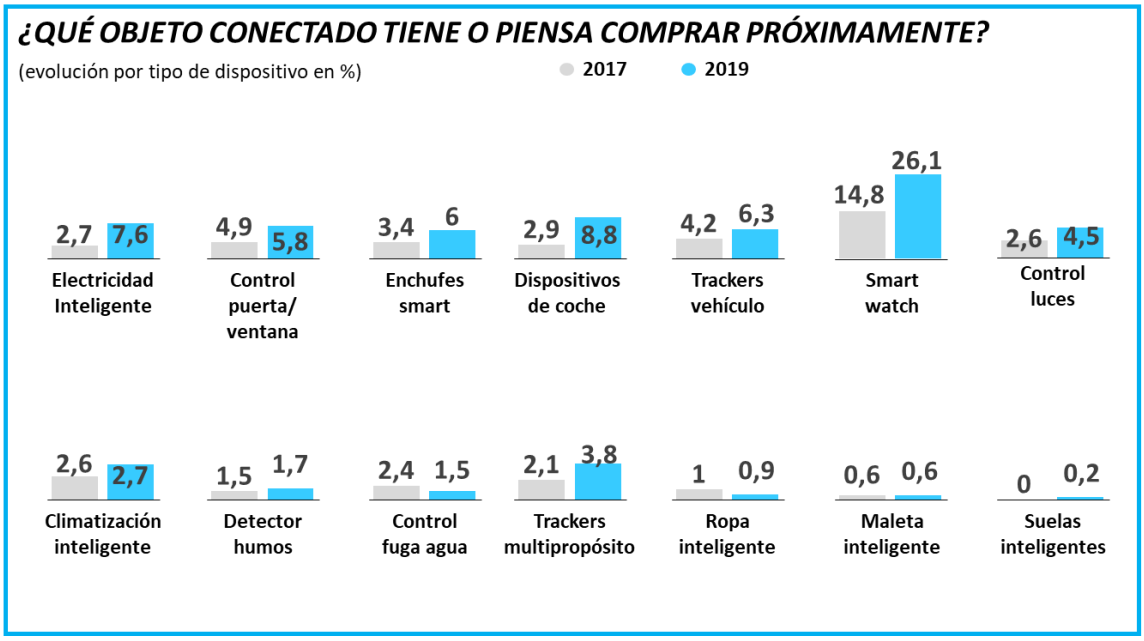


Imagen 1: Distribución de objetos conectados 2017 vs 2019 [8].

Por otro lado, si se hace un estudio exhaustivo de proveedores de tecnologías de IoT para consumidores, existe una gran variedad de empresas locales e internacionales dedicadas a la producción y comercialización de este tipo de dispositivos. Sin embargo, la mayor parte de la atención se centra en los principales actores a nivel mundial: Amazon, Google, Microsoft, Huawei y Apple. Estos lideran tanto en sus propios países como en el extranjero debido a la reputación y popularidad heredadas de otras industrias como son la computación en la nube, el entretenimiento, equipos personales, la inteligencia artificial, entre otros.

1.2. Estado del arte

Como consecuencia del interés que está despertando el IoT en los consumidores y el alto valor monetario que representa para las empresas de este sector, se ha hecho necesario estudiar las características de los usuarios que hacen uso de estas tecnologías, así como las barreras y factores que influyen en la adopción (o no) de estos.

De hecho, el análisis de las percepciones y actitudes de las personas en la adopción de estos tipos de dispositivos ya ha sido tratado por algunas investigaciones anteriores las cuales han sido publicadas desde mediados de los 2000 [9]. Sin embargo, dado que los servicios y productos relacionados con el IoT son muy diversos, la revisión bibliográfica que se ha llevado a cabo se centra en los trabajos relacionados con las barreras y comportamientos de adopción de los consumidores.

Se ha intentado analizar la comprensión de las actitudes en la adopción de dispositivos IoT desde la perspectiva del usuario. Varios de los estudios se han enfocado en grupos de usuarios específicos tales como los ancianos, los discapacitados y los pacientes ya que usualmente los dispositivos “wearable” IoT son utilizados con fines de monitoreo de salud y/o actividad física [5]. Por ejemplo, Vadillo y otros [10] investigaron la adopción del sistema de teleasistencia, el cual es un tipo de servicio domiciliario

inteligente, y descubrió que es de alta importancia la utilidad percibida en la intención de utilizar este sistema. También en [11] se estudió la adopción en Francia de usuarios potenciales de dispositivos de medicina basados en IoT encontrando el coste de los mismos como principal limitante.

Otros estudios han analizado el uso los dispositivos de IoT en el hogar que están más asociados al control remoto, la automatización y la vigilancia del hogar, y a la recreación [1]. En este aspecto [12] analizaron en el contexto de hogar inteligente los principales factores influyentes en su adopción por los consumidores. Entre otros resultados, vale la pena destacar que este estudio concluyó que en la aceptación del hogar inteligente tanto la confianza como el riesgo de seguridad juegan un rol muy importante. En el estudio [6] utilizaron un modelo de aceptación de la tecnología para estudiar simultáneamente la adopción y difusión del hogar inteligente y encontraron que la compatibilidad de la tecnología, facilidad de uso y utilidad percibidas influyen positivamente en la intención de compra [6]. En [13] estudiaron también las posibles barreras en la adopción de productos propios del hogar inteligente, pero desde la perspectiva del riesgo percibido lo cual incluye financiero, de desempeño, de privacidad y psicológico. Haciendo uso de un modelo de ecuación estructural demostraron que estos riesgos influyen en la resistencia a consumir estos tipos de servicios, excepto por el riesgo financiero. En [14] utilizaron también un modelo de ecuaciones estructurales para medir el nivel de aceptación del IoT teniendo en cuenta conceptos como la aplicabilidad de estos dispositivos en tareas diarias, influencia social, facilidad de adaptabilidad al cambio, así como la confianza, utilidad, flexibilidad y facilidad encontrada en el IoT. Por otro lado, en [15] estudiaron varios factores que conformaban el sacrificio y el beneficio percibido, y su influencia en el valor de los hogares inteligentes, lo cual a su vez influía en la actitud y en la intención de usar dichos servicios. Este estudio comprobó que el beneficio percibido tenía una importancia considerablemente mayor que el valor percibido y que también tenía una gran influencia en la actitud y la intención de uso. Finalmente, [5], mediante el análisis factorial confirmatorio examinaron las características de los servicios del hogar inteligente y evaluaron la relación entre los factores críticos y el comportamiento de adopción desde el punto de vista de la automatización, controlabilidad e interconectividad conseguida con los mismos. Sus resultados mostraron que la controlabilidad es el antecedente más importante para la adopción.

No obstante, es importante resaltar que durante la revisión bibliográfica se encontraron más estudios relacionados con los aspectos técnicos del IoT que aquellos sobre los aspectos de comportamiento y adopción. Además, muy pocas publicaciones científicas analizan en detalle la adopción de los servicios y la tecnología relacionada con el IoT como un todo. Uno de ellos, [16] estudió los factores que influyen en la confianza que tienen los consumidores en el IoT y el papel que juegan en la adopción de esta tecnología concluyendo que el rol de la confianza puede ser relevante pero no condición suficiente. Otra de las publicaciones realizada por [17] encontraron que los beneficios percibidos, la actitud y la preocupación por la privacidad de la información afectan de forma significativa y directa a la intención continuada de utilizar los servicios de IoT.

Por todo lo expuesto anteriormente, se puede decir que a pesar de las contribuciones realizadas por estudios anteriores al entendimiento de las barreras y motivaciones de la

adopción del IoT, actualmente todavía existe una falta de datos matizados sobre el uso y otras percepciones de diferentes grupos de edad y a la vez, de dispositivos de IoT de diferente índole. La mayoría de los estudios analizados tienen diferentes enfoques al analizar las causas que influyen en la adopción o rechazo de esta tecnología, no proveyendo una perspectiva generalizable y unificada. Adicionalmente, una gran cantidad se enfocan únicamente en el diseño de experimentos o en la propuesta de la tecnología a emplear, resultando en que los enfoques ingenieriles sobre IoT no logren dar una respuesta correcta ante las verdaderas necesidades de los usuarios.

Actualmente, los diferentes dispositivos comercializables de IoT se están transformado para satisfacer necesidades de uso general y que puedan ser utilizados a conveniencia, ya no son un servicio especializado para un grupo específico de personas. Consecuentemente, se ha notado que existe una falta de modelos de confianza que aclaren los requisitos de los consumidores y no consumidores de IoT y que estudien los principales factores que influyen en la adopción o no de la tecnología y servicios relacionados con el IoT. Adicionalmente, se hace necesario desarrollar un análisis que estudie en profundidad las características de los diversos grupos de usuarios de dispositivos conectados a Internet, así como el nivel de penetración actual existente.

1.3. Objetivos

Es por esto que este trabajo tiene como objetivo general realizar una caracterización del uso o no de las tecnologías y servicios asociados al Internet de las Cosas en los hogares españoles durante el 2020.

Para la realización con éxito del objetivo general, se han trazado una serie de objetivos específicos que ayudaran al cumplimiento del objetivo general:

- Realizar un análisis descriptivo de la población española que conoce y/o hace uso de dispositivos relacionados con el IoT.
- Determinar el perfil entre los diferentes tipos de usuarios (o no) de dispositivos IoT de acuerdo con sus características.
- Evaluar empíricamente la relación entre los factores principales y el comportamiento de adopción o no adopción de los diferentes dispositivos IoT.
- Proporcionar observaciones de utilidad para los profesionales en la industria de IoT que gestionan el marketing o la comercialización de estos dispositivos.

2. Diseño metodológico

Los datos de este trabajo procederán de la encuesta de 2020 sobre Equipamiento y Uso de las Tecnologías de la Información y la Comunicación en los Hogares (TIC-H) realizada por el Instituto Nacional de Estadística de España siguiendo las recomendaciones de la Oficina Estadística de la Unión Europea (EUROSTAT) [18]. Esta encuesta ha sido realizada en los hogares del territorio español para obtener datos comparativos sobre la disponibilidad y el uso que se hace de dichas tecnologías [18]. Además, respecto a otras anteriores se diferencia principalmente por recoger ya el impacto de la pandemia

provocada por la COVID-19, por lo que los resultados de este trabajo mostrarán cómo se ha afectado el uso de dispositivos IoT por este fenómeno mundial.

Como fuente de datos se tomaron los archivos de microdatos [19] asociados a la encuesta en cuestión. Los archivos de microdatos incluyen de forma anonimizada los datos individuales resultantes de una encuesta. Estos archivos tienen formato ASCII donde para cada individuo se tienen los valores que toma cada variable en una estructura de campos.

2.1 Tratamiento previo de los datos y selección inicial de variables

Debido a que, en los ficheros de microdatos los datos se encuentran de manera desagregada tal y como se muestra en la Imagen 2, se hace necesario un tratamiento previo de los mismos.

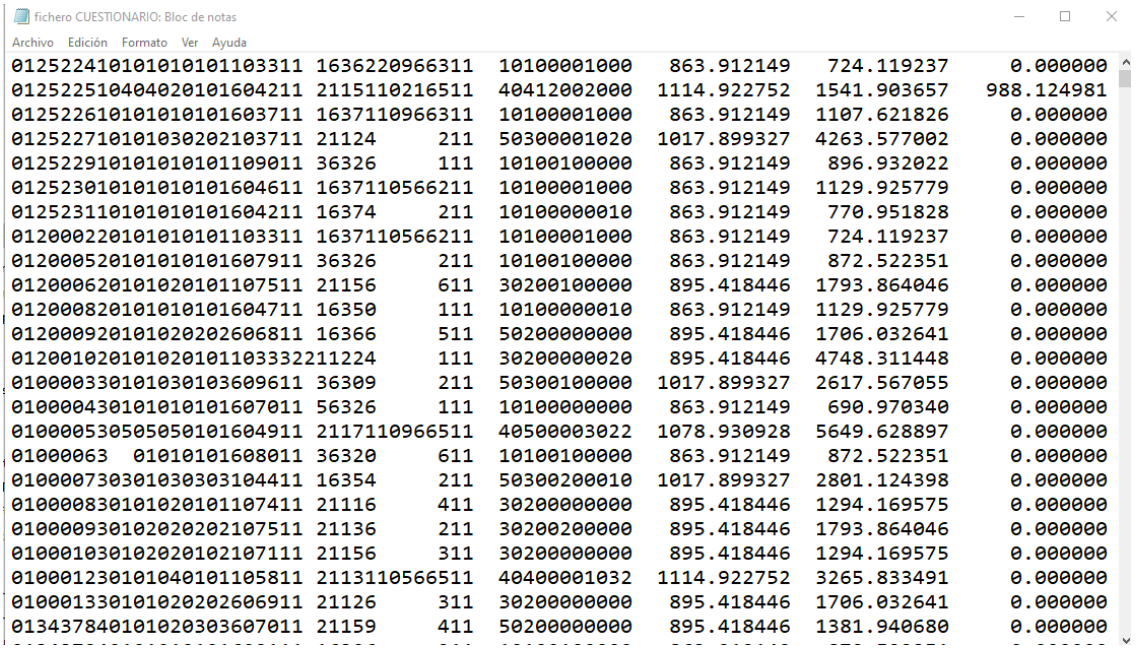


Imagen 2: Ejemplo formato archivo microdatos.

Con este objetivo, se utilizó un código en el lenguaje de programación R y la herramienta R-Studio. Este código para cada variable contemplada en la encuesta recogía, tanto su tamaño como su posición final e inicial. De esta manera se pasaba a convertir en formato tabla el fichero de microdatos. El resultado final fue un archivo con 214 variables o columnas, y 15343 filas. Sin embargo, la codificación de cada variable no estaba del todo interpretable ya que los valores que tomaban cada una necesitaban ser decodificados. Por ejemplo, tomando como referencia la variable que recoge el estado civil legal de la persona encuestada, como se muestra en la siguiente tabla, esta variable tenía como valores originales los números del 1 hasta el 5. Estos valores necesitaban ser reemplazados por aquellos que estaban en la encuesta y que son más entendibles para su posterior procesamiento.

Tabla 1: Ejemplo decodificación valores en el fichero de microdatos.

Valor en el archivo microdatos	Valor reemplazado
1	Soltero/a
2	Casado/a
3	Viudo/a
4	Separado/a
5	Divorciado/a

Para llevar a cabo este reemplazo se utilizó un código en el programa SAS Base. Esta decodificación se llevó a cabo para todas las variables categóricas presentes en la encuesta, es decir, para un total de 191 variables. La salida de este código fue un archivo tipo tabla en formato xlsx con cada variable tomando valores interpretables para un procesado posterior.

Después se pasó a realizar un estudio más profundo de los datos con la finalidad de seleccionar aquellos individuos y aquellas variables que fueran de interés para el presente estudio. Es por ello que primeramente se hizo un filtrado de aquellas personas entre 16 y 74 años que habían utilizado internet en los últimos 3 meses. A este filtro se le adicionó que esas personas hubieran respondido el bloque de la encuesta relacionado con uso de dispositivos o sistemas conectados a Internet con fines privados. Esto resultó en una reducción del número de filas de 15343 a 11394. Luego se procedió a seleccionar las variables que en un principio tenían sentido considerar en el estudio lo cual incluyó variables sociodemográficas como son el sexo, género, edad; de capacidades individuales como por ejemplo el uso de internet y conocimientos informáticos; y finalmente económicas como son la situación laboral y los ingresos mensuales. A continuación, se muestra una tabla con la primera versión de selección de variables realizada, así como su significado.

Tabla 2: Variables inicialmente seleccionadas.

Nombre de la variable	**Descripción**
CPRO	Identificador de la Provincia
SEXO	Sexo de la persona seleccionada
EDAD	Edad de la persona seleccionada
HABITAT_EUR	Código de hábitat de Eurostat (DEG_URBA)
TIP_H	Tipo de hogar
TOT_MH	Total de MIEMBROS del hogar
TOT_MEN16	Total de miembros MENORES (ESTRICTO) de 16 años
TOT_MAY74	Total de miembros MAYORES (ESTRICTO) de 74 años
ESTC	Estado civil legal de la persona seleccionada
CONV	Convivencia en pareja
NIVELEST	Estudios terminados
SIT_LAB	Situación laboral en la que se encuentra

TIP_JOR	Tipo de hornada en trabajo principal
ACTIV	Actividad del establecimiento
OCUPACION1	Ocupación principal: Manual / No manual
OCUPACION2	Ocupación principal: Trabajador TIC / No TIC
ING_HOG	Ingresos mensuales netos del hogar
FACTOR_H	Factor de elevación del hogar
FACTOR_P	Factor de elevación del informante
VINTD	Uso de Internet varias veces al día
TMOR1	Tareas relacionadas con móviles y ordenadores: transferir ficheros entre el ordenador y otros dispositivos
TMOR2	Tareas relacionadas con móviles y ordenadores: instalar software o aplicaciones (apps)
TMOR3	Tareas relacionadas con móviles y ordenadores: cambiar la configuración de cualquier software
TAREAINF1	Tareas informáticas realizadas: copiar o mover ficheros o carpetas
TAREAINF2	Tareas informáticas realizadas: usar un procesador de texto
TAREAINF3	Tareas informáticas realizadas: crear presentaciones o documentos que integren diferentes ficheros
TAREAINF4	Tareas informáticas realizadas: usar hojas de cálculo
TAREAINF5	Tareas informáticas realizadas: usar software para editar fotos, video o audio
TAREAINF6	Tareas informáticas realizadas: programar en un lenguaje de programación
PREOPUB	Grado de preocupación respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida
Precook	Cambio configuración navegador para evitar cookies
CONFINT	Grado de confianza en Internet
DISP48_1	Uso de dispositivos conectados a internet: administración de energía en el hogar
DISP48_2	Uso de dispositivos conectados a internet : sistema de alarma
DISP48_3	Uso de dispositivos conectados a internet : electrodomésticos
DISP48_4	Uso de dispositivos conectados a internet : asistente virtual altavoz inteligente
CONOCDIS	Conoce existencia de dispositivos conectados a internet
NODISP1	Razón para no usar dispositivos: no tiene necesidad
NODISP2	Razón para no usar dispositivos: costes elevados
NODISP3	Razón para no usar dispositivos: incompatibilidad con otros dispositivos
NODISP4	Razón para no usar dispositivos: falta de habilidades
NODISP5	Razón para no usar dispositivos: preocupación privacidad
NODISP6	Razón para no usar dispositivos: preocupación seguridad (sea pirateado)
NODISP7	Razón para no usar dispositivos: preocupación seguridad (accidente) o repercusiones sobre salud
NODISP8	Razón para no usar dispositivos: otras razones
DISP51_1	Uso internet dispositivos: televisor conectado
DISP51_2	Uso internet dispositivos: consola conectada
DISP51_3	Uso internet dispositivos: sistema de audio doméstico conectado, altavoces inteligentes
DISP52_1	Uso internet dispositivos: reloj inteligente, auriculares…
DISP52_2	Uso internet dispositivos: Para la salud, controlar presión, azúcar, peso…
DISP52_3	Uso internet dispositivos: juguetes conectados
DISP52_4	Uso internet dispositivos: automóvil conexión inalámbrica

Para poder realizar una exploración más detallada de las variables y sus categorías se procedió a utilizar la herramienta SAS Enterprise Miner Workstation por la facilidad que brinda para estos procesos. Gracias a los nodos *DMDB* y *Explorador de Estadísticos* de

esta herramienta se pudo apreciar la no existencia de datos faltantes en la muestra y también la distribución y el número de niveles de las variables categóricas (Ver Anexo C). Al presentar algunas variables nominales con un número de niveles elevados y algunas con determinadas categorías poco representadas se procedió a realizar una recodificación de esos niveles a través del nodo *Reemplazo* existente en la herramienta. A continuación, se detallan los principales cambios realizados:

- Para el caso de la actividad del establecimiento en el que trabaja la persona encuestada se unió la categoría "Otros servicios" con "No se puede codificar como "Otros". También se unieron las categorías de "Actividades inmobiliarias" y "Construcción" como "Industria inmobiliaria". Adicionalmente se recodificó la categoría de NA como No Aplica porque está asociada a aquellas personas que no se encuentran trabajando activamente.
- Para el nivel de estudio se hicieron algunas modificaciones también como "Analfabetos y estudios primarios incompletos" con "Educación primarios" y "No se puede codificar" que es equivalente a sin título/certificado para formar "Sin estudios/estudios primarios e infantil". Además, se unieron "Segunda etapa de educación secundaria y similar" con "Educación postsecundaria no superior" para formar "Segunda etapa de educación secundaria y educación postsecundaria no superior". Finalmente se unió "Formación Profesional" con "Grados Universitarios" y "Título de Doctorado" para formar la categoría de "Educación superior".
- La situación laboral se modificó para unir las categorías de "Trabajo por cuenta ajena" con "Trabajo por cuenta propia" y "Trabajo Temporal" en la categoría "Trabajando". Adicionalmente se fusionaron las categorías "Otra situación", "Incapacitado" y "Voluntariado" como "Otros".

Gracias a este procesado no solo han disminuido la cantidad de niveles, sino que, además, cada uno de estos se encontraron mejor representados.

- Recodificación de la variable intervalo *edad* como variable categórica con los tramos de edad entre 16-24 años, 25-34 años, 35-44 años, 45-54 años, 55-64 años y más de 65 años.
- Creación de la variable dicotómica *uso_IoT48* que toma el valor "Sí" si la persona encuestada ha hecho uso de alguno de los dispositivos de IoT de la pregunta 48 (Ver Anexo A) y "No" si no ha utilizado ninguno.

2.2 Análisis descriptivo

Como se mencionó anteriormente, de la muestra inicial de la encuesta, el presente estudio trabajó solo con los resultados de unas 10.216.795 personas. Esto se debe a que el objetivo planteado es estudiar aquellos usuarios entre 16 y 74 años que han utilizado internet en los últimos 3 meses y que han respondido el bloque de la encuesta relacionado con uso de dispositivos o sistemas conectados a Internet con fines privados. Además, para dar cumplimiento al objetivo del presente trabajo, con esta muestra, se

realizó un estudio estadístico descriptivo que detalle el comportamiento y distribución del uso de los diferentes dispositivos de IoT en la población española.

2.3 Análisis de componentes principales

- Correlaciones policóricas y tetracóricas

El uso de correlaciones policóricas y tetratóricas en el presente estudio se debió a la naturaleza de los datos. La mayoría de las variables consideradas en este trabajo son binarias o nominales con varias categorías. De ahí que, para poder hallar la matriz de correlaciones entre las mismas se utilizaron este tipo de técnicas pues en la revisión bibliográfica realizada obtenían mejores resultados que el método habitual mediante la correlación de Pearson [20]–[22]. De hecho, normalmente la correlación tetracórica es empleada con datos dicotómicos, mientras que la correlación policórica se emplea para datos que son ordinales y para el análisis de la interacción entre variables dicotómicas y ordinales [23].

El modelo y los supuestos para la correlación policórica y tetracórica son los mismos; solo que en el caso de la policórica existen un mayor número de parámetros de umbral debido al mayor número de niveles de calificación ordenados.

Como no es el objetivo principal de este trabajo ahondar en los procedimientos de cálculos de este tipo de correlaciones, se remite al siguiente trabajo donde entran en el detalle matemático de su obtención [24]. En este estudio, el cálculo de estas correlaciones se llevó a cabo a través del software R-Studio utilizando el paquete *psych* y las funciones *Tetrachoric()* y *polychoric().*

- Análisis de componentes principales

El análisis de componentes principales (PCA por sus siglas en inglés) es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información [25].

Si se supone la existencia de una muestra con n individuos cada uno con p variables (X_1, X_2, …, X_p), siendo el espacio muestral de p dimensiones. A través del uso del análisis de componentes principales se puede encontrar un número de factores subyacentes ($z<p$) que explican aproximadamente lo mismo que las p variables originales. Ahora, con estas z nuevas variables, o componentes principales, bastan para caracterizar a cada individuo cuando antes había p valores [25].

El análisis de componentes principales pertenece a la familia de técnicas conocida como aprendizaje no supervisado. En este tipo de métodos la variable resultante no es relevante porque la intención es extraer información a través de las componentes principales y no predecir. El método de PCA permite por lo tanto “condensar” la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas

estadísticas tales como regresión y/o clustering [26]–[28]. Vale la pena mencionar que esos nuevos vectores o componentes principales se estructuran de forma tal que forman una combinación lineal mediante la cual se pueden expresar las variables originales [25].

- Cálculo de las componentes principales para matriz de correlación

El análisis de componentes principales se empleó para reducir la dimensionalidad de aquellas variables relativas a las habilidades y conocimientos informáticos de los individuos de la encuesta, la frecuencia de uso y nivel de confianza en internet, y el grado de preocupación respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida. El cálculo de dichas componentes se basó en la matriz de correlaciones explicadas anteriormente: tetracóricas y policóricas en dependencia del tipo de variable a tratar. Además, se empleó el lenguaje de programación R mediante el uso de la función *principal()* también del paquete *psych.* Esta función realiza una descomposición del valor propio devolviendo dichos valores propios, las cargas y el grado de ajuste para un número determinado de componentes que se le pasa como parámetro.

Las cargas de las variables pueden interpretarse como el peso que tiene cada variable en cada componente principal. De ahí que se pueda identificar qué información está recogida en cada una de las componentes principales [25].

Una vez calculada la primera componente, se computa el resto de las componentes teniendo en cuenta la condición de que estas no pueden estar correlacionadas. El orden de importancia de las componentes viene dado por la magnitud del autovalor asociado a cada autovector.

- Determinación del número de componentes principales

Existen varios criterios para analizar cuál es el número indicado de componentes principales que se ha de retener.

En este estudio una de las pautas seguidas fue retener la cantidad de componentes que lograra explicar al menos un 70% de la información. Consecuentemente, tanto la proporción de varianza explicada como la proporción de varianza explicada acumulada fueron dos valores de referencia para tomar la decisión en cuanto al número de componentes.

Otro método para determinar el número de componentes fue el "*Método de Catell"* el cual consiste en graficar los autovalores, y para aquellos en los que, al trazar una recta que aglutine los valores más pequeños, queden por encima se corresponderían con las componentes principales a retener [25].

2.4 Análisis clúster

Para dar cumplimiento a los objetivos específicos del presente estudio se hizo necesario utilizar técnicas exploratorias que permitieran agrupar las diferentes personas encuestadas teniendo en cuenta características similares de las mismas. Consecuentemente se emplearon varios tipos de análisis clúster, el jerárquico que tiene un carácter más exploratorio y posteriormente, el no jerárquico mediante el cual se agruparon finalmente los individuos.

- Gower Measure (común a ambos análisis clúster)

La distancia es una métrica numérica utilizada para medir la proximidad o la similitud entre individuos. Existen muchas medidas de distancia, pero la utilizada en este estudio fue la distancia de Gower [29] para ambos análisis clústeres. Esta distancia es la recomendada a utilizar por la bibliografía en el caso de querer hallar la distancia entre variables nominales, ordinales y de intervalo, como es el del presente estudio.

La distancia de Gower se calcula como la media de las disimilitudes parciales entre individuos, tomando valores entre 0 y 1. La estimación de las disimilitudes parciales variará en dependencia del tipo de variable en cuestión. Es decir, básicamente mide cuán diferente es un individuo de otro y aplica una métrica u otra cuando trata con diferentes tipos de variables [29].

2.4.1 Análisis clúster jerárquico

El análisis clúster jerárquico (ACJ) se clasifica como una técnica estadística multivariante de clasificación no supervisada. Se utiliza para agrupar individuos similares en un conjunto de datos [30]. Esta técnica intenta que los individuos dentro de cada clúster sean lo más similares entre sí, y lo más diferentes con los individuos del resto de grupos. Este algoritmo de hecho, se basa em la distancia entre los datos (o similitud) para construir un a jerarquía de clasificación de grupos en varios pasos donde los clústeres de niveles más bajos se engloban en niveles superiores.

El ACJ realiza los siguientes pasos:

1) Se empieza de tantos grupos como observaciones.
2) Una matriz n x n es generada conteniendo la distancia entre todos los pares de observaciones.
3) Se unen los dos clústeres más cercanos, y resultando en una cantidad de grupos menor que en el paso anterior.
4) Se calcula nuevamente la matriz de distancias utilizando los nuevos clústeres formados.
5) Finalmente, los pasos 3 y 4 son repetidos hasta que todas las observaciones estén agrupadas en un solo clúster.

Por ello, se hace necesario definir con anterioridad tanto la distancia a utilizar entre los individuos (distancia de Gower) como la distancia entre los diferentes clústeres.

- Distancia entre clústeres

En el presente estudio se utilizaron tres métodos para el cálculo de distancias entre las agrupaciones de individuos, escogiendo finalmente, el que mejores resultados dio. Vale la pena mencionar que este método se utilizó con el objetivo de detectar el número ideal de conglomerados a realizar en el análisis de clúster no jerárquico que se detallará más adelante. En total se decidió probar los métodos de distancia de Ward, enlace medio y distancia entre centroides para hallar la distancia entre agrupaciones [31]. A continuación, se brinda una breve explicación de esos métodos:

- Distancia de Ward o de la mínima varianza: esta distancia tiene como objetivo minimizar la varianza dentro del propio clúster. De esta forma se puede seleccionar la unión que minimiza la variabilidad interna de cada agrupación resultante.

- Enlace medio: la distancia entre dos clústeres se determina como la distancia media entre individuos de distintos grupos. De esta forma, los clústeres que se forman tienen varianza similar y lo más pequeña posible.

- Distancia entre centroides: distancia entre los centroides de cada clúster.

- Determinar el número óptimo de clústeres

No existe una forma o metodología exacta para determinar el número óptimo de agrupaciones a realizar. Al utilizar varios métodos para el cálculo de la distancia entre clústeres, se tomó una decisión en función de los resultados finales obtenidos. Los indicadores que se explican más adelante se utilizaron con el objetivo de definir la cantidad adecuada de clústeres. Estos indicadores utilizados en el ACJ fueron el R^2 y el R^2-semiparcial (SPRSQ); y además se tuvo en cuenta el Pseudo F y el Pseudo Test de la T [31]. Los conceptos en los que se basan estos indicadores de referencia son la variabilidad dentro de los clústeres (W), la variabilidad entre los clústeres (E), y la variabilidad total (T).

- R^2: Se puede definir como la proporción de la variabilidad explicada por los grupos creados.

$$R^2 = \frac{E}{T}$$

Se buscó explicar una variabilidad de al menos el 70%.

- R^2-semiparcial: es básicamente la diferencia entre la proporción de variabilidad explicada con *l* clúster o con k (*l*>*k*)

$$SPRSQ = R_l^2 - R_k^2$$

- Pseudo Test de la T: afirma que en el caso de que las medias de dos clústeres distintos (l y k) no sean significativamente diferentes, entonces se podrían combinar esos dos clústeres sin la que la variabilidad dentro del clúster

resultante fuese significativamente superior. Por lo tanto, al agrupar esos dos clústeres, la dispersión interna de este último clúster (W_m) debe de ser mayor que la suma de las dispersiones de los clústeres que lo forman.

$$Pseudo - T^2 = \frac{W_m - W_k - W_l}{\frac{W_k + W_l}{n_k + n_l - 2}}$$

En este caso n_k y n_l representan el número de observaciones en los clústeres k y l, respectivamente. Por lo tanto, se buscó un incremento excesivo o un máximo relativo para el Pseudo Test de la T, escogiendo entonces el número de clústeres que venía a continuación.

- Pseudo F: se utiliza para comparar la dispersión entre las agrupaciones con la dispersión dentro del clúster.

$$Pseudo - F = \frac{\frac{E}{g-1}}{\frac{W}{n-g}}$$

En la fórmula anterior n es el número total de individuos, y g es el número de agrupaciones que se está evaluando. Para este caso se buscó aquel número de clústeres donde este valor tuviera máximos relativos o incrementos importantes[31].

El análisis de clúster jerárquico se llevó a cabo utilizando el programa de SAS Base con la sentencia de *proc cluster* y se le indicó el método a utilizar en dependencia de la distancia entre clústeres que se quiso utilizar.

2.4.2 Análisis clúster: k-medioids

El análisis clúster basado en k-medoid es una técnica de clustering de grupos que, al igual que el k-means (algoritmo en el que se basa), intentan minimizar la distancia correspondiente entre los individuos de un clúster formado, y el "centro" o medoide en este caso. Sin embargo, en algo que se difiere el algoritmo k-medoid al de k-means, es que, en vez de escoger la media de las variables, el k-medoid selecciona un individuo en su totalidad o "*datapoint*" como centro de cada agrupación. A partir de ahí, minimiza la suma entre pares de puntos utilizando la distancia indicada, lo que le hace más robusto ante el ruido. Es decir, se puede definir al medoide de cada clúster formado como aquel individuo que, respecto al resto de individuos de ese clúster, tendrá una disimilaridad mínima.

Después de una extensa revisión bibliográfica sobre qué algoritmo de agrupación utilizar en el caso de un set de datos extenso y de variables mixtas, es decir, tanto categóricas como de intervalo, se llegó a la decisión de utilizar un algoritmo basado en k-medoid para realizar el clustering [32]–[37]. Específicamente se seleccionó el método de Partición Alrededor de Medoides (PAM), propuesto por Kaufman y Rousseeuw en 1990 [38]. Este método demostró obtener resultados positivos en cuanto a la reducción del

número de valores atípicos, así como buenos resultados de rendimiento y de ajustes a diferentes conjuntos de datos [39]–[43].

Para realizar este algoritmo se utilizó el programa de R-Studio con la función *pam()* del paquete *cluster.* El algoritmo PAM en R funciona de la siguiente manera:

1. Comienzo: selección de *k* medoides de todos los individuos existentes, siendo k la cantidad de clústeres indicados.
2. Calcular la matriz de disimilitud, si no se le proporciona.
3. Unir cada individuo con el medoid menos disimilar.
4. Para cada clúster formado se busca si alguno de los individuos en ese clúster hace que el coeficiente de disimilitud promedio disminuya. Si eso pasa, se selecciona como medoid aquel objeto que hace que más disminuya ese coeficiente. Si al menos algún medoid ha cambiado se regresa al paso 3 y si no se termina el algoritmo.

- Selección del número adecuado de clústeres: coeficiente de "Silhouette"

Para ayudar a la selección del número adecuado de clústeres para realizar la agrupación se tuvo en cuenta el coeficiente de "Silhoutte". La silueta, por su nombre en español, es una ayuda gráfica para interpretar y validar en el análisis de clúster [44]. El valor de la silueta mide la similitud de un objeto con su propio clúster en comparación con otros. Y su valor se encuentra entre -1 y +1. Un valor alto indica que el objeto está fuertemente emparejado con su propio clúster e inadecuadamente emparejado con los clústeres vecinos. La fórmula de la silueta se muestra a continuación:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

donde $a(i)$ es la distancia media entre *i* y todos los demás puntos de datos del mismo clúster, $b(i)$ es la menor distancia media de *i* a todos los puntos de cualquier otro clúster.

Es decir, la media $S(i)$ de todos los miembros de un grupo da una idea de cuán estrechamente agrupados están todos los miembros ese clúster. De hecho, ese valor al que se le denomina Coeficiente de Silueta (SC).

2.5 Análisis correspondencia simples

El Análisis de Correspondencias Simples (ACS) es una de las técnicas multivariantes utilizadas en el caso de variables categóricas. Esta técnica facilita la detección de relaciones existentes entre variables mediante la visualización de tablas de contingencia. Dicha tabla de contingencia es obtenida cuando dos variables nominales son cruzadas y repartidas en la muestra en correspondencia con el número de individuos que presentan cada categoría de cada una de las variables. El ACS es utilizado con el

objetivo principal de encontrar relaciones entre las categorías de dos variables a la vez que se logra reducir la dimensionalidad [45].

En el presente trabajo se hizo uso de esta técnica para identificar y caracterizar entre los diferentes tipos de individuos, cuáles eran los que hacían uso o no de los dispositivos IoT. Para su ejecución se utilizó tanto el programa SAS Base, con el procedimiento *proc corresp*, y el programa R-Studio. La idea de utilizar ambos programas fue debido a la facilidad que proveen ambos para este análisis y para poder hacer un uso intercalado de los gráficos y visualizaciones que proveen.

- Estudio de la tabla de contingencia

En la tabla de contingencia ya mencionada anteriormente se representa, en las columnas las categorías de una de las variables analizadas y en las filas, las modalidades de la otra variable. Con las frecuencias, tanto absolutas como relativas, que forman esta tabla se procedió a realizar diferentes tipos de análisis para dar cumplimiento a los objetivos del presente estudio.

- Viabilidad del ACS

Para que el análisis de correspondencia simple tenga sentido, existen una serie de condiciones que se deben cumplir. Una de ellas, es que las variables en cuestión no sean independientes para poder así ver qué tipo de asociaciones existen entre las mismas. Se consideró que dos variables fueran independientes si el valor de una variable no contribuye a la distribución de la otra variable. Consecuentemente, un contraste de hipótesis fue realizado con el objetivo de poner a prueba la independencia de las variables. Para esto se utilizó el estadístico de Chi-Cuadrado de Pearson. La hipótesis nula viene dada por la asunción que las variables serán independientes, evaluando así las desviaciones que presenta la muestra respecto al valor teórico previsto [45].

El estadístico Chi-cuadrado se define como:

$$\chi^2_{(r-1)(c-1)} = \sum_{i}^{r}\sum_{j}^{c}\frac{(f_{ij} - e_{ij})^2}{e_{ij}},$$

donde f_{ij} se entiende como la frecuencia al cruzar las categorías i de la fila (r) y j de la columna (c). Por otro lado, la variable e_{ij} se entiende como la frecuencia esperada bajo la hipótesis cero del cruce de categorías i y j. Este estadístico se distribuye con $(r-1)(c-1)$ grados de libertad.

Para poder rechazar la hipótesis nula se necesita que las diferencias entre los valores observados y los esperados sea lo suficientemente grande. Por ello, si el estadístico Chi-cuadrado presenta valores altos, se puede decir que la relación entre las variables estudiadas es bastante fuerte. Adicionalmente, en el caso de que el valor del p-valor fuera lo suficientemente pequeño, también se rechazaría la hipótesis nula.

- Número adecuado de dimensiones a retener

Para elegir el número correcto de dimensiones a retener se tuvieron en cuenta dos criterios basados en la inercia. La inercia *(I)* se puede entender como una medida de la dispersión semejante a la varianza de datos numéricos [45]. De hecho, es una medida

que coincide con el estadístico Chi-cuadrado, representando la distancia de los perfiles al perfil medio ponderados por la masa de los perfiles. Adicionalmente, este parámetro toma valores altos cuando las variables están relacionadas entre sí. Se consideró como criterio principal explicar al menos el 70% de la inercia, escogiendo tantas dimensiones como fueran necesarias para cumplir este criterio.

3. Caracterización de la muestra

Como se comentó en el capítulo anterior, los datos utilizados en este trabajo están formados por las respuestas de personas del territorio español (entre 16 y 74 años) sobre la disponibilidad y el uso que hacen de diferentes dispositivos IoT [11]. Para darle cumplimiento a los objetivos del presente trabajo y poder conocer las características de la muestra en cuestión, en este capítulo se realizará un análisis descriptivo univariante y bivariante de las variables más significativas y que aportan más información.

Para poder llevar a cabo el análisis descriptivo de los datos, fue necesario tener en cuenta el factor de elevación/ponderación. Este es muy relevante a la hora de que los datos de la encuesta se asemejen a un nivel representativo de la población española. Es por ello, que para poder crear los gráficos de distribución de frecuencias para cada variable era necesario utilizar el factor de expansión. En el caso específico de esta encuesta, se trabajó con la variable *factor_p* que es el factor de elevación del informante.

- Análisis general

La muestra de este estudio está formada aproximadamente por unas 32.844.074 personas debido a que el foco es, como se comentó en capítulos anteriores, estudiar aquellos usuarios entre 16 y 74 años que han utilizado internet en los últimos 3 meses y que han contestado el bloque de la encuesta relacionado con el uso de dispositivos o sistemas conectados a Internet. A continuación, se presenta un análisis descriptivo de la muestra general del presente trabajo.

Por ejemplo, si se estudia la distribución por sexo (Imagen 3) se aprecia que tanto los hombres como las mujeres están casi igualmente representados teniendo un 49,6% y 50,4% respectivamente.

Sin embargo, al mirar los diferentes tramos de edad considerados (Imagen 4), se nota que la mayoría se encuentra entre 35-54 años con un 43,91%. De hecho, es el rango de 45 a 54 años el que mayor porcentaje tiene de todos los intervalos con un 22,33%, mientras que los mayores de 65 años son los que menor representados están, siendo un 9,84% del total. También se aprecia cómo las personas de menos de 35 años son en total un 29% de toda la muestra estudiada.

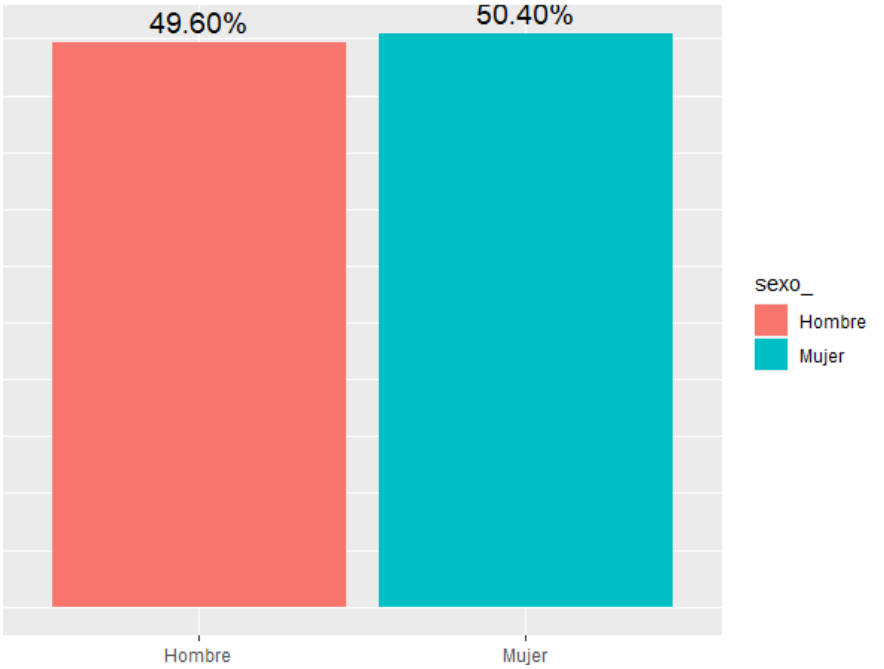


Imagen 3: Distribución por sexo.

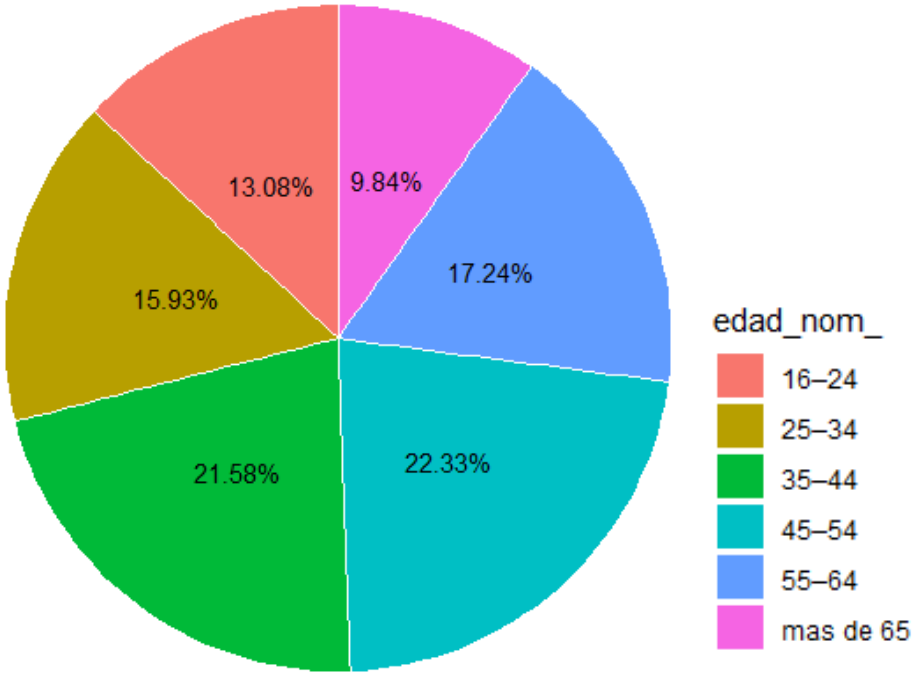


Imagen 4: Distribución por edad.

En correspondencia con la distribución por edad, en la Imagen 5 se puede apreciar cómo más de la mitad de las personas en la muestra se encuentran trabajando activamente representando un 51,65% de la muestra, frente al 16% se encuentra parado. Adicionalmente, un 4,18% se encuentra realizando tareas del hogar, siendo el porcentaje más bajo. Por otro lado, un 28,15% de la muestra se encuentra inactivo en términos laborales repartido en 10,7% de estudiantes, 10,17% de jubilados o prejubilados, y un 7,28% de otros (voluntariado, incapacitado, otros).

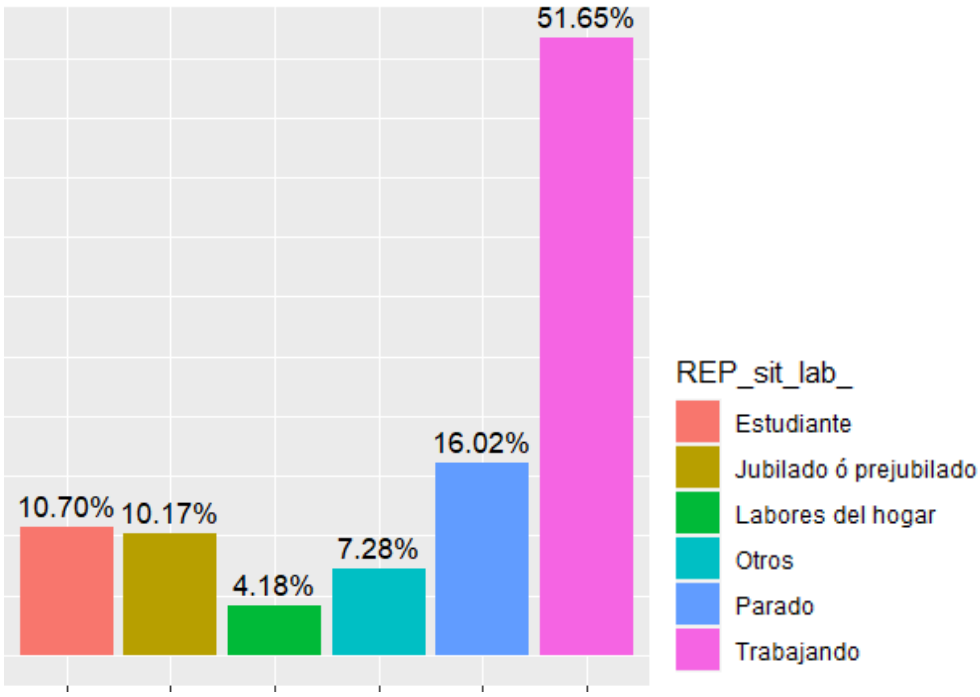


Imagen 5: Distribución según la situación laboral.

En cuanto a la distribución según el nivel de ingresos, la Imagen 6 muestra que la mayoría de los intervalos están relativamente igual representados excepto por los que cobran menos de 900€ que son la minoría siendo un 14.9% del total. También se ha de notar que la cantidad de personas que cobran entre 900€ y 1600€ al mes son el intervalo más representado siendo un 34,9% del total.

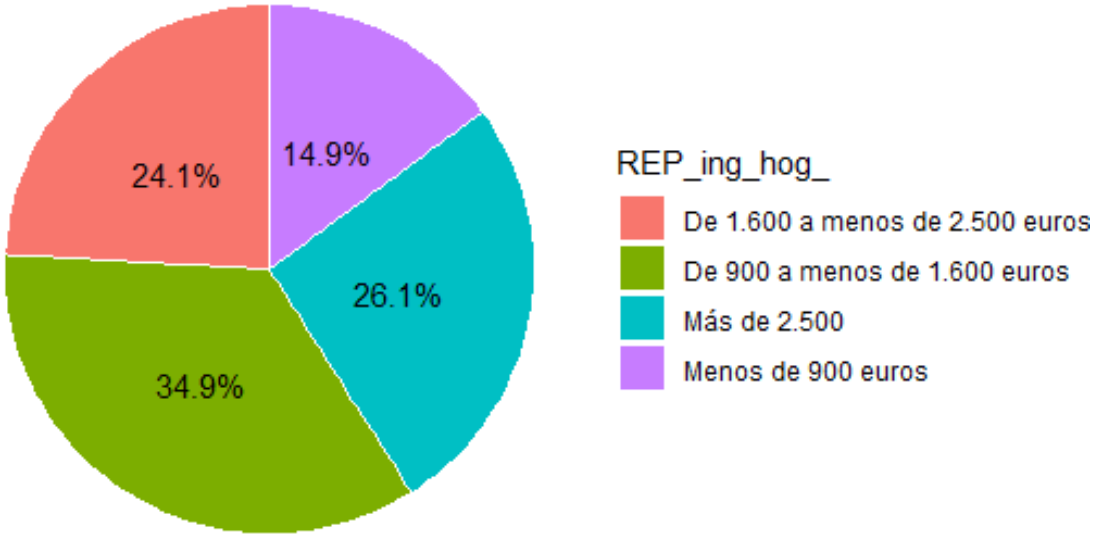


Imagen 6: Distribución según rango de ingresos.

A continuación, en la Imagen 7, se muestra la distribución de la muestra según el nivel de estudio que presentan los individuos y se puede apreciar que casi el 39% cuenta con una Educación Superior mientras que solo el 11,4% no tiene estudios o cuenta con estudios primarios. Es decir, la muestra utilizada en este estudio está formada por un

88,6% de personas que han completado al menos la primera etapa de la educación secundaria. Para más información respecto a qué estudios se encuentran en cada categoría ver el Anexo B.

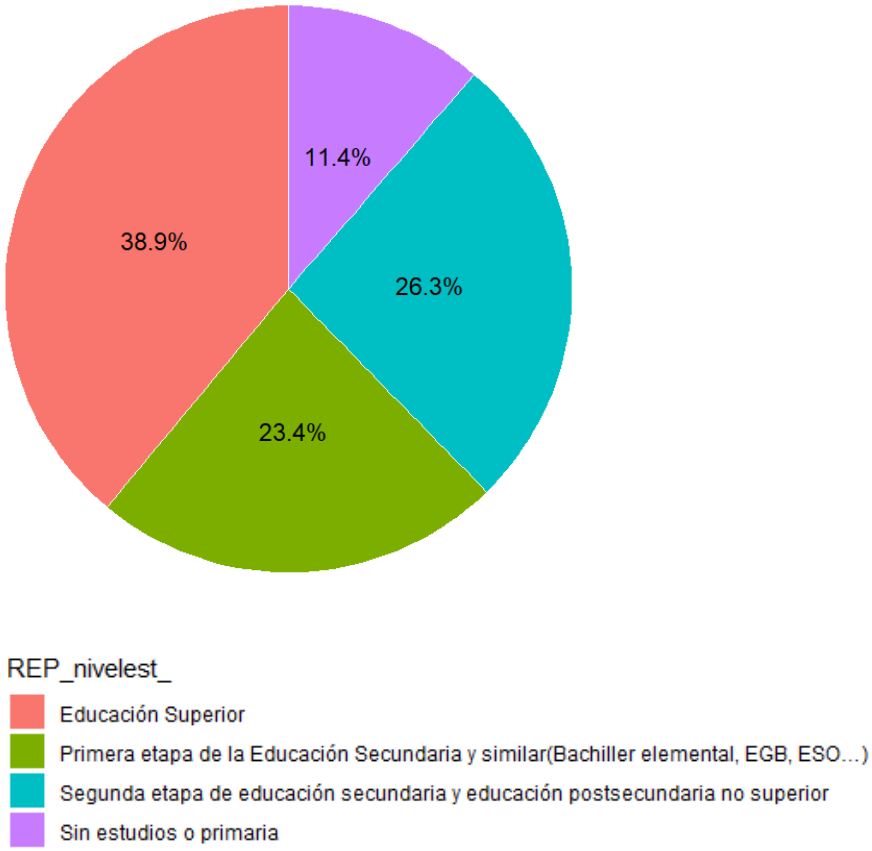


Imagen 7: Distribución por nivel de estudios.

Finalmente, y por tener una idea de cuánto es el uso de internet en el siguiente gráfico se puede apreciar cómo las personas encuestadas usan internet diariamente representando un 89,1% del total.

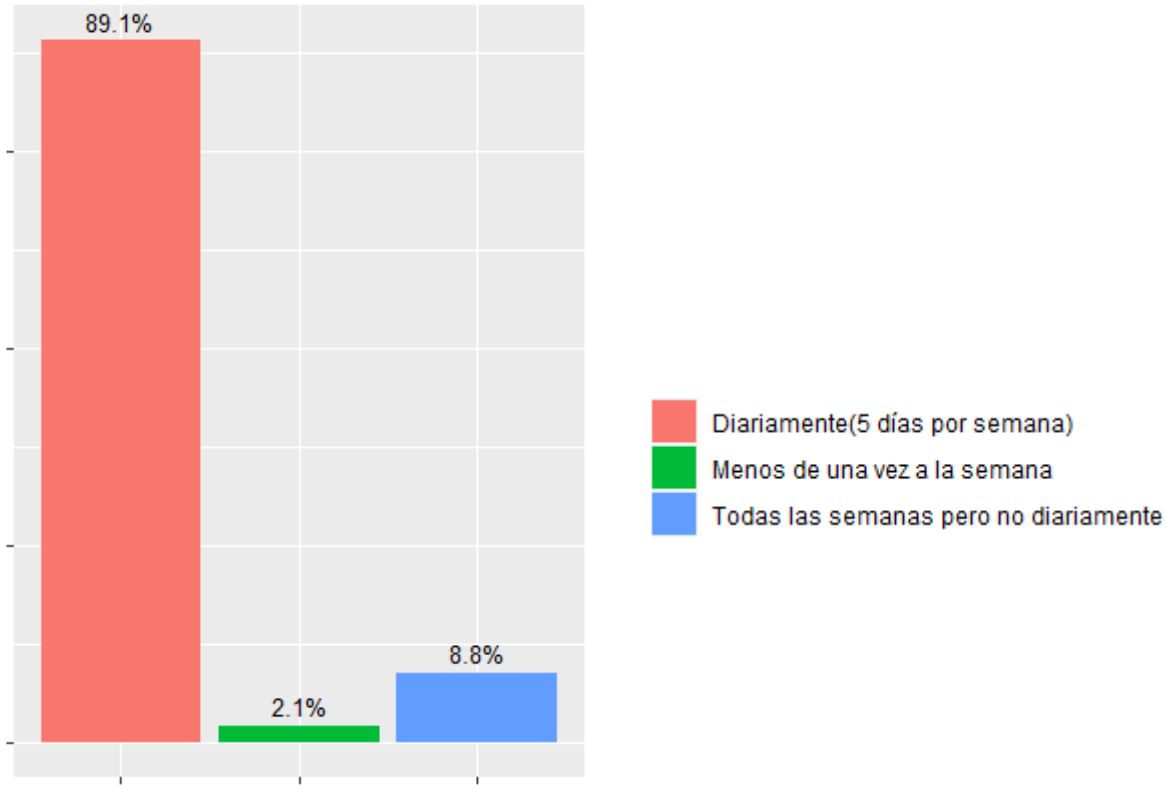


Imagen 8: Distribución según el uso de internet.

- Análisis descriptivo IoT

A continuación, se describirá la muestra según su conocimiento y uso de los dispositivos IoT considerados por la encuesta sobre el Equipamiento y Uso de las Tecnologías de la Información y la Comunicación en los Hogares (TIC-H).

En la Imagen 9 se puede apreciar cómo un 85% de la muestra conoce este tipo de dispositivos mientras que solo un 15% no sabe lo que son. Si se representa la distribución del 85% de las personas que los conoce según los estudios finalizados, en la Imagen 10 se aprecia que el 42,1% tienen una educación superior. Además, en total, un 91,3% han completado al menos la primera etapa de la educación secundaria y similar. Esto se puede deber a que al aumentar el nivel de educación aumenta el nivel de conocimientos y cultural de cada individuo y por eso pueden estar más familiarizados con esta tecnología.

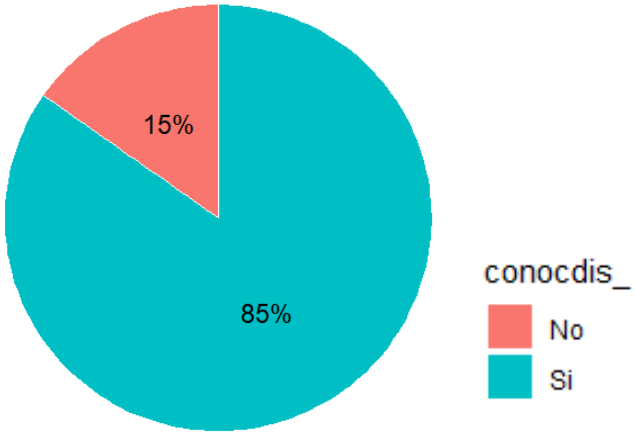


Imagen 9: ¿Conoce los dispositivos IoT?

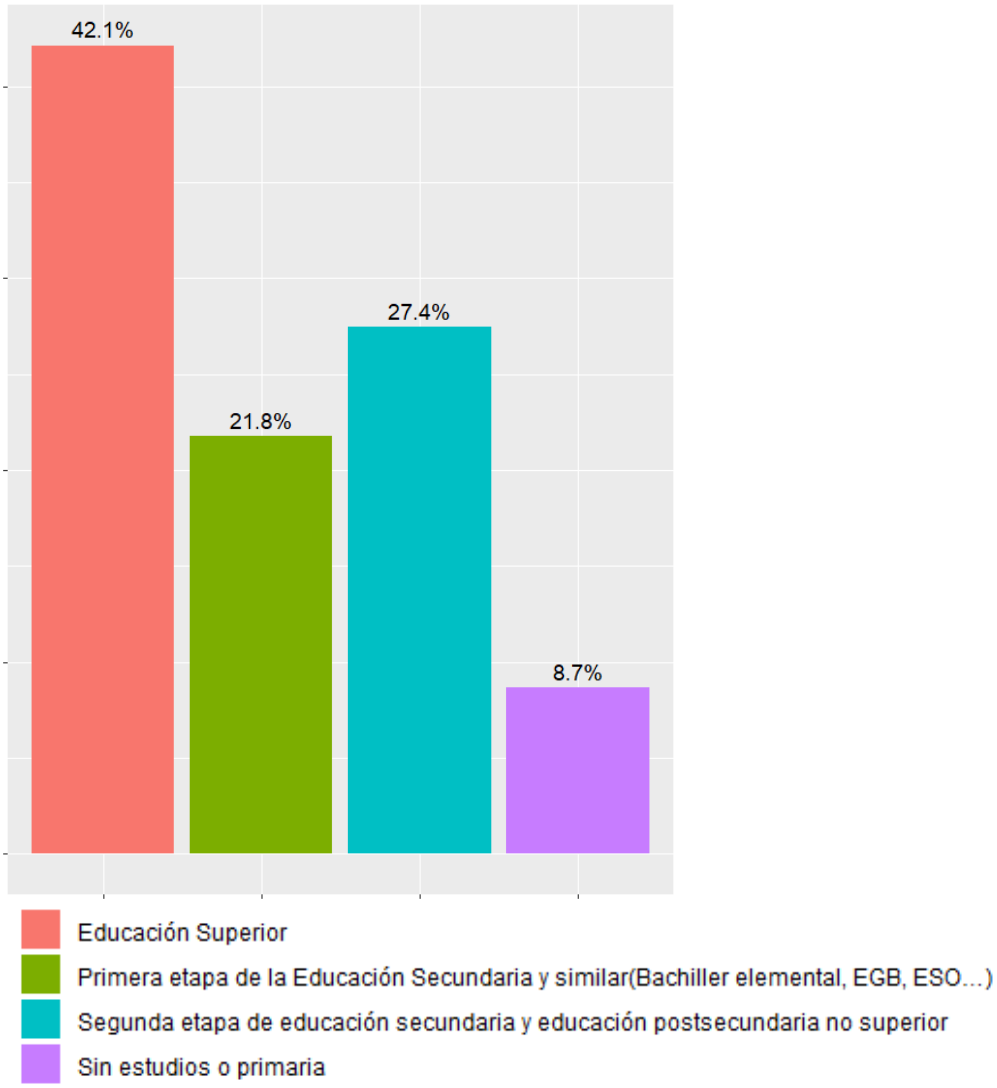


Imagen 10: Distribución de personas que conocen dispositivos IoT según nivel de estudios.

Por otro lado, si se representa la distribución por edad de las personas que conocen los dispositivos IoT, en la Imagen 11 se puede apreciar cómo solo un 7,81% son personas de

más de 65 años y que las distribuciones de edades que más conocen estos dispositivos son de 35-44 años y de 45-54 años con un 22,37% y 22,19% respectivamente. Además, en total, más de un 53% de la muestra tiene una edad de menos de 44 años. Esto puede explicarse por el motivo de que las personas más avanzadas en edad encuentran más complicados a hacerse al uso de nuevas tecnologías como regla general, mientras que las generaciones más jóvenes son nativos a la evolución de estas.

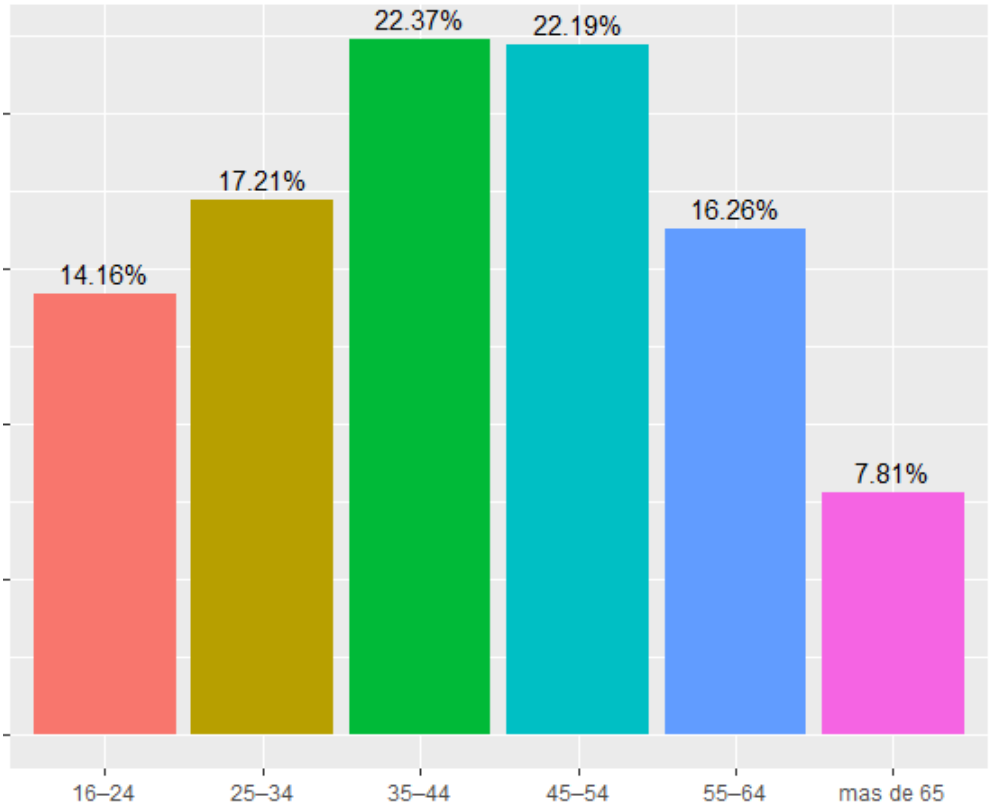


Imagen 11: Distribución de personas que conocen dispositivos IoT según la edad.

En cuanto al **uso de dispositivos o sistemas conectados a Internet** con fines privados, en los que están incluidos:

- Sistemas para la administración de energía en su hogar, termostatos, luces, enchufes u otras soluciones
- Sistemas de seguridad para el hogar,
- Electrodomésticos conectados
- Un asistente virtual en forma altavoz inteligente o de app.

En la Imagen 12 se puede apreciar que, aunque la mayoría de las personas sí conocen estos dispositivos, solo un 31% hace uso de los mismos mientras un 69% no. Esto demuestra que la adopción de esta tecnología en territorio español, aunque en crecimiento, aún está en sus inicios.

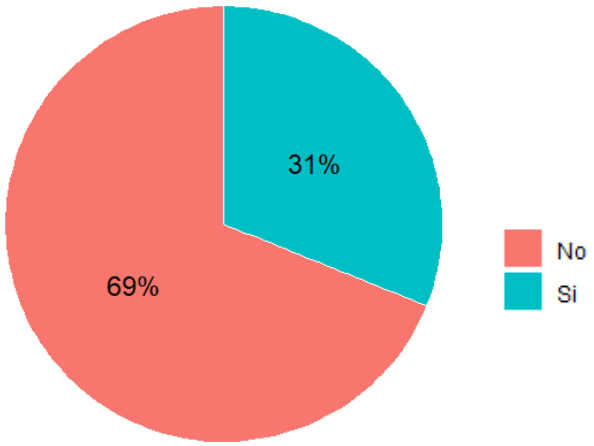


Imagen 12: ¿Usa alguno de estos dispositivos IoT?

Por tener un conocimiento más detallado de cuáles son los dispositivos que se han utilizado alguna vez y su distribución según la cantidad de personas que lo han hecho, en la Imagen 13 y la Tabla 3 se encuentra el detalle de estos datos. En ambas se puede apreciar como la mayoría de las personas lo que más usan son los asistentes virtuales en forma altavoz inteligente o de app, como Alexa, Siri, Google Home, etc con unas cantidades que son casi el doble que el resto. El segundo lugar lo ocupan los electrodomésticos conectados (aspiradoras de robot, frigoríficos, hornos) con un 10,8% y 3,5 millones de personas. Por otro lado, el menos usado corresponde a los sistemas para la administración de energía en el hogar como termostatos, luces, enchufes u otras soluciones con un total de un 8,4% y 2,7 millones de personas.

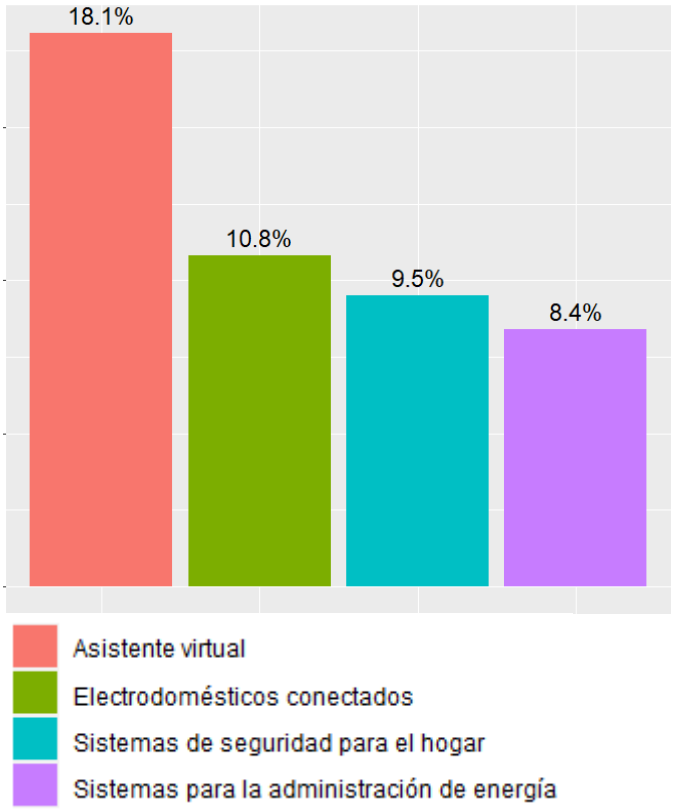


Imagen 13: Porcentaje que usan dispositivo IoT según el tipo de dispositivo.

Tabla 3: Distribución de personas que usan dispositivo IoT según el tipo de dispositivo.

Nombre Dispositivo	**Cantidad**	**Porcentaje**
Sistemas para la administración de energía	2.765.308	0,08419505
Sistemas de seguridad para el hogar	3.128.035	0,09523893
Electrodomésticos conectados	3.553.987	0,10820785
Asistente Virtual	5.941.055	0,18088667

Continuando el estudio de las personas que usan o no este tipo de dispositivos según la distribución de ingresos, en la Imagen 14 se puede apreciar como del 31% que sí los usan, el mayor porcentaje (aproximadamente 34%) pertenecen a los que cobran más de 2500€ y solo un 8,6% de los que cobran 900€ los usan, lo cual es evidencia de la influencia del salario en el uso de estos dispositivos ya que actualmente el precio no es particularmente bajo. Además, más de la mitad de la concentración del porcentaje de uso (61,34%) está en los hogares que perciben más de 1600€ al mes.

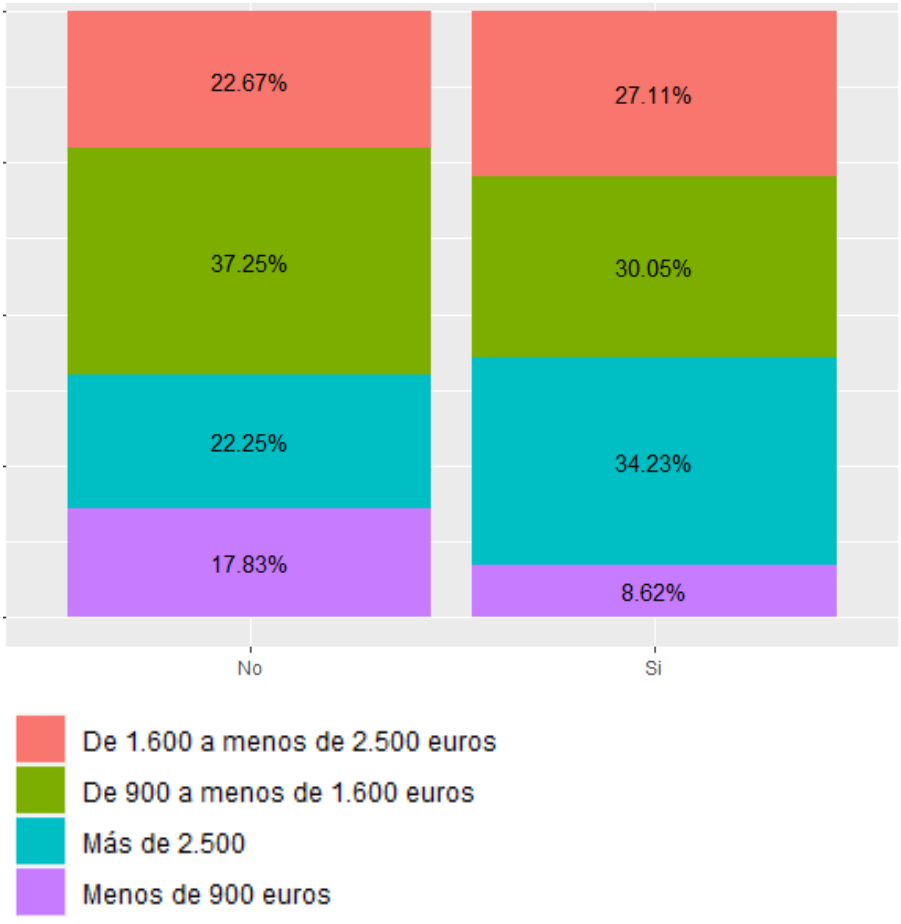


Imagen 14: Distribución de personas según el uso de dispositivos IoT y salario.

Por otro lado, en la Imagen 15 se puede apreciar que de los que usan dispositivos, casi el 48% tienen una formación perteneciente al rango de educación superior, lo cual evidencia la correlación entre el conocimiento de estos dispositivos y el uso de los mismos. Sin embargo, llama la atención que de los que no los usan, el mayor porcentaje

(35%) también esté mayormente representado por personas que tienen una educación superior. Esto probablemente se deba a que las personas con un nivel superior de estudios tengan una menor confianza y seguridad del uso que se le pueda dar a sus datos personales a través de estos dispositivos. Adicionalmente, se puede ver como las personas son estudios o con último nivel terminado como los estudios primarios con los que, en ambos casos, presentan un menos porcentaje.

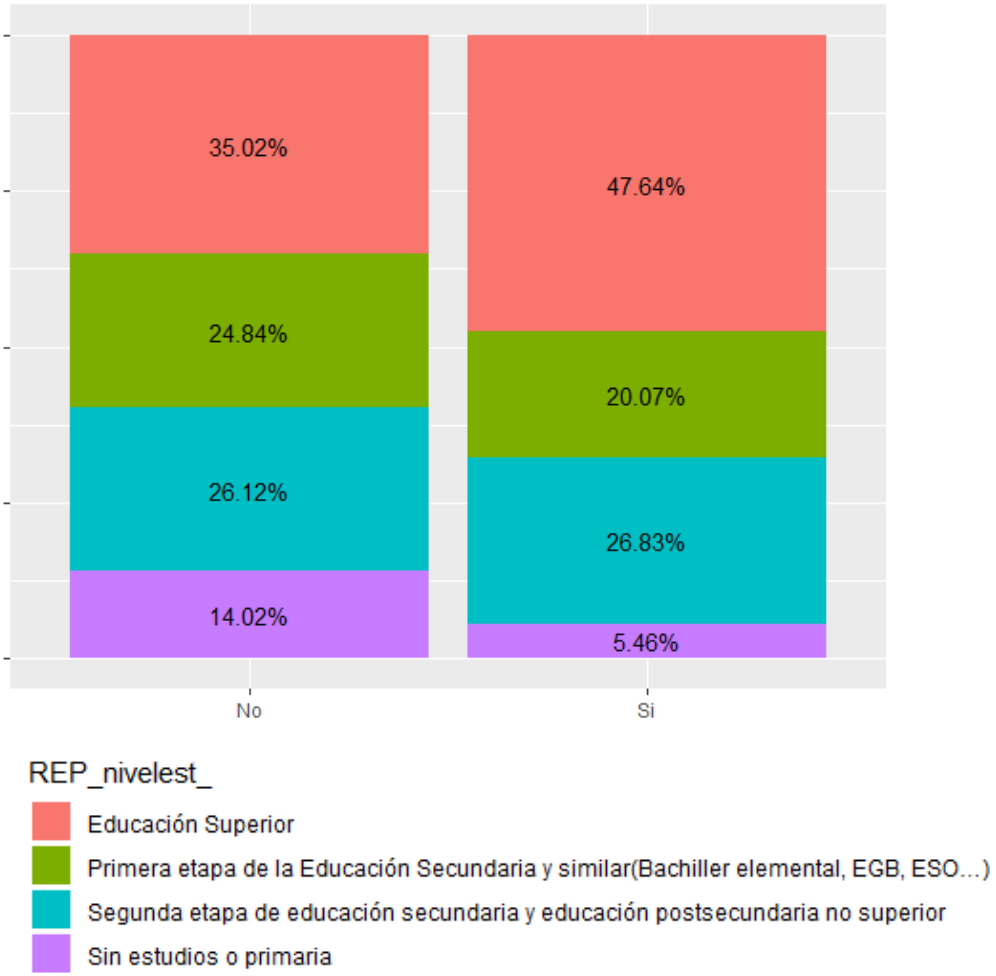


Imagen 15: Distribución de personas según el uso de dispositivos IoT y estudios terminados.

En temas de edad y como se muestra en la Imagen 16 se puede decir que:

- De los que sí usan dispositivos IoT, el mayor porcentaje se encuentra entre las edades de 25-54 años representando casi un 66% respecto al total en esa categoría.
- En cuanto a los que no lo usan, el rango se mueve un poco hacia edades más avanzadas, ya que más del 53% son de personas con más de 45 años.

Esto se corresponde con lo expresado anteriormente respecto al conocimiento de este tipo de dispositivos y su relación con la edad.

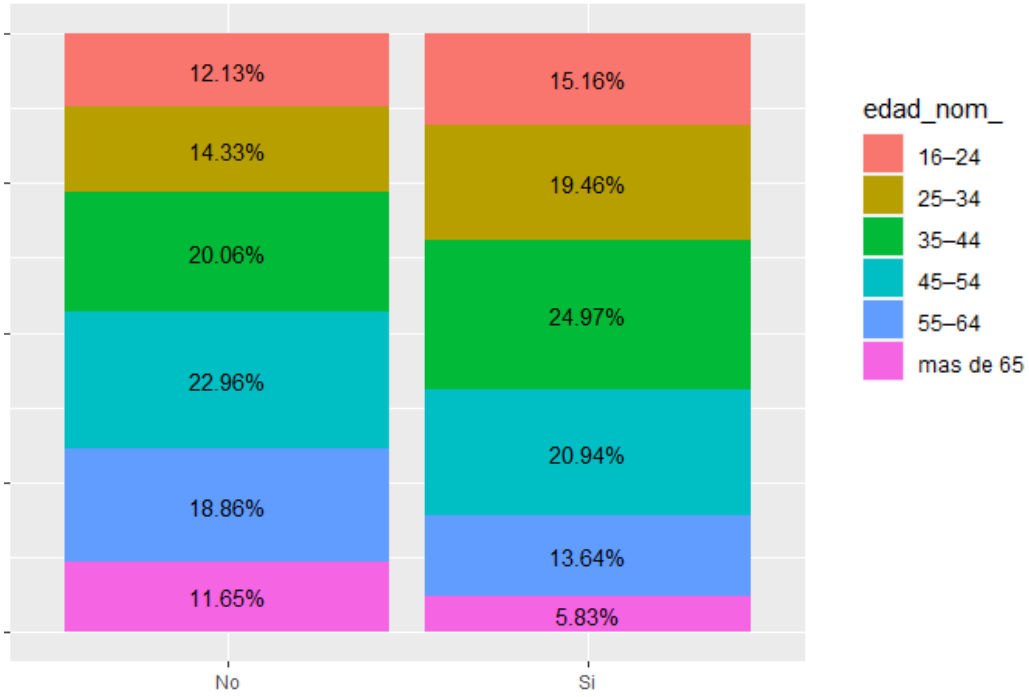


Imagen 16: Distribución de personas según el uso de dispositivos IoT y edad.

Debido al alto porcentaje de personas que no hacen uso de los dispositivos se hace muy interesante para el presente trabajo hacer un estudio más detallado del por qué tantas personas no los utilizan actualmente. A continuación, se presentan las barreras en el uso de dispositivos IoT. Los porcentajes en la Imagen 17 y en la Tabla 4 indican la cantidad de personas que respondieron sobre qué barrera en cuestión era la razón del no uso de estos dispositivos. El porcentaje representa la cantidad de personas que respondieron que "Sí" a la barrera en cuestión respecto al total que respondieron esa misma pregunta. Se puede apreciar que la mayoría de las personas no usan dispositivos conectados a Internet porque no tienen necesidad de hacerlo, representando un total del 74,8% y una cantidad de 13 millones de personas. La segunda barrera más seleccionada fue la de costes demasiado altos, con un 38,7% de las personas seleccionándolo como motivo lo que representan un total de 6,8 millones de individuos. Adicionalmente, existen dos barreras con un peso similar que en este caso fueron las de preocupación sobre la privacidad y protección de datos personales generados por esos dispositivos o sistemas y la de preocupación sobre la seguridad (por ej. que sea pirateado), ambas con un 35% aproximadamente. El resto de las barreras como son falta de compatibilidad con otros dispositivos o sistemas, falta de habilidades para utilizar esos dispositivos o sistemas, preocupaciones sobre la seguridad o la salud (p. ej. pueda provocar un accidente o un problema de salud) y otras tienen distribuciones similares. Sin dudas, esta distribución contribuye a entender los principales impedimentos que actualmente presenta la sociedad española en el uso de dispositivos IoT. Adicionalmente, establece las principales líneas de acción a seguir, teniendo en cuenta que, a priori, aún existe la necesidad de informar sobre los beneficios de estas tecnologías, así como de su uso.

Tabla 4: Distribución de selección de las diferentes barreras en el uso de IoT.

Nombre Barrera	**Cantidad**	**Porcentaje**
NoNecesidad	13.183.278	0,7481839
Costes	6.820.468	0,3870785
Compatibilidad	3.820.990	0,2168507
Habilidades	4.518.296	0,2564245
PreocupPriv	6.280.193	0,3564166
PreocupSeg	6.270.348	0,3558579
PreocupSegSal	3.714.007	0,2107791
Otras	3.725.254	0,2114174

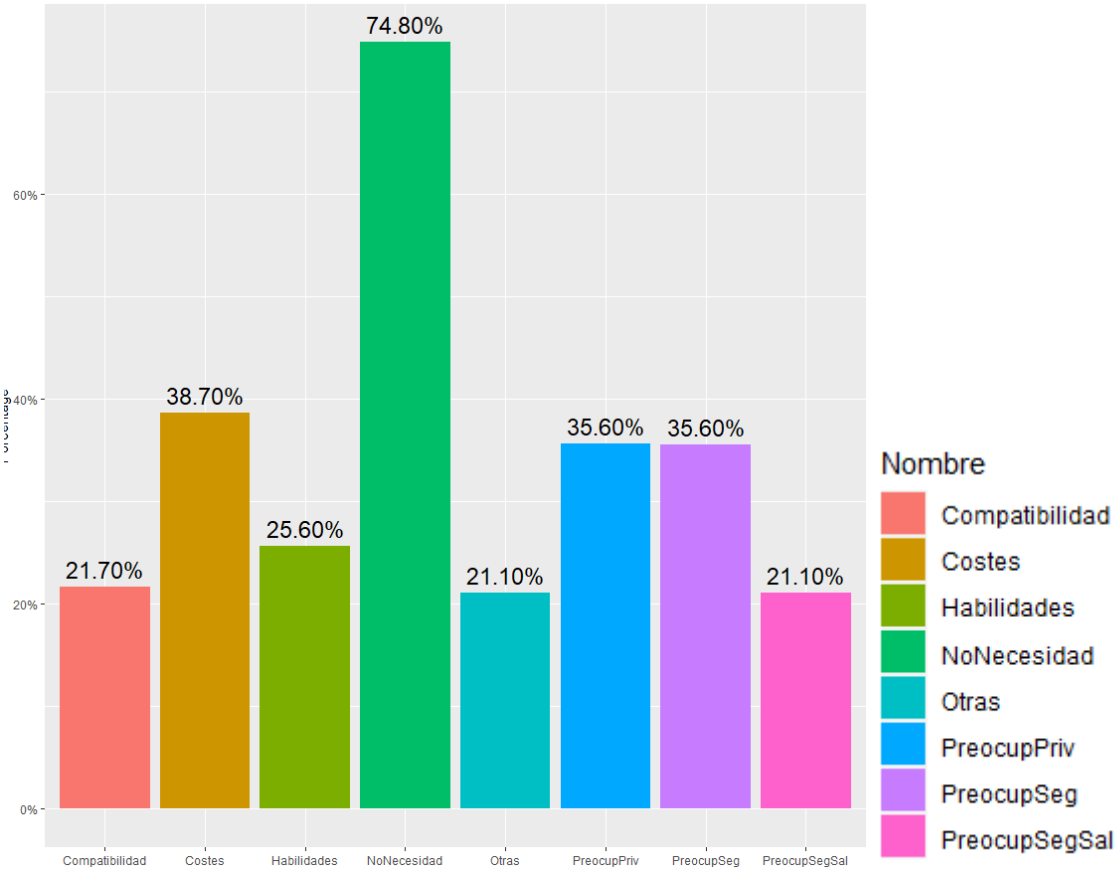


Imagen 17: Distribución de selección de las diferentes barreras en el uso de IoT.

4. Desarrollo del trabajo: Resultados y discusión.

En este capítulo se detallan los resultados de la metodología y procedimientos explicados en el Capítulo 2. Primeramente, se explicará la reducción de la dimensionalidad de la muestra para alguna de las variables utilizando el análisis de componentes principales. Después se procederá a explicar la realización y resultados de los diferentes tipos de análisis clúster realizados para poder agrupar a los individuos de la muestra en grupos similares. Por último, se presentará un análisis de

correspondencias simples para relacionar a esos individuos con el uso de los distintos dispositivos de IoT recogidos por la encuesta del INE.

4.1 Análisis de componentes principales

Al contar con un número elevado de variables, se hizo uso de la técnica de componentes principales para poder reducir la dimensionalidad del conjunto de datos, a la vez se pudiera contar con la información de esas variables para los estudios posteriores. De ahí que se decidiera realizar el análisis de componentes principales para las siguientes variables:

Tabla 5: Variables utilizadas en el análisis de componentes principales.

Nombre variable	**Descripción**	**Valores**
VINTD	Uso de Internet varias veces al día	SI NO
TMOR1	Tareas relacionadas con móviles y ordenadores: transferir ficheros entre el ordenador y otros dispositivos	SI NO
TMOR2	Tareas relacionadas con móviles y ordenadores: instalar software o aplicaciones (apps)	SI NO
TMOR3	Tareas relacionadas con móviles y ordenadores: cambiar la configuración de cualquier software	SI NO
TAREAINF1	Tareas informáticas realizadas: copiar o mover ficheros o carpetas	SI NO
TAREAINF2	Tareas informáticas realizadas: usar un procesador de texto	SI NO
TAREAINF3	Tareas informáticas realizadas: crear presentaciones o documentos que integren diferentes ficheros	SI NO
TAREAINF4	Tareas informáticas realizadas: usar hojas de cálculo	SI NO
TAREAINF5	Tareas informáticas realizadas: usar software para editar fotos, video o audio	SI NO
TAREAINF6	Tareas informáticas realizadas: programar en un lenguaje de programación	SI NO
PREOPUB	grado de preocupación respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida	Muy preocupado Algo preocupado Nada preocupado
PRECOOK	Cambio configuración navegador para evitar cookies	SI NO
CONFINT	Grado de confianza en Internet	Bastante Mucho Poco o nada

Como el principal supuesto de este tipo de análisis es que las variables en cuestión estén relacionadas, primeramente, se computó la tabla de correlaciones, en este caso policórica, debido a la naturaleza de los datos. En la Imagen 18 se puede apreciar el nivel de correlación entre las variables elegidas. Los colores rojos indican una correlación positiva, mientras que los colores azules indican una correlación negativa. Además, a mayor intensidad del color, mayor nivel de correlación. También en esta imagen se puede ver cómo todas las variables están de alguna forma correlacionadas entre sí. Una relación interesante para destacar es la de las variables de *preopub* y *confint.* Se puede apreciar cómo a medida que aumenta el grado de preocupación por el uso de información para ofrecer publicidad a medida, la confianza en internet disminuye. Esta relación también parece darse con las variables que recogen los conocimientos

informáticos y la preocupación, mientras más conocimientos informáticos, la preocupación respecto al uso de los datos personales también es mayor.

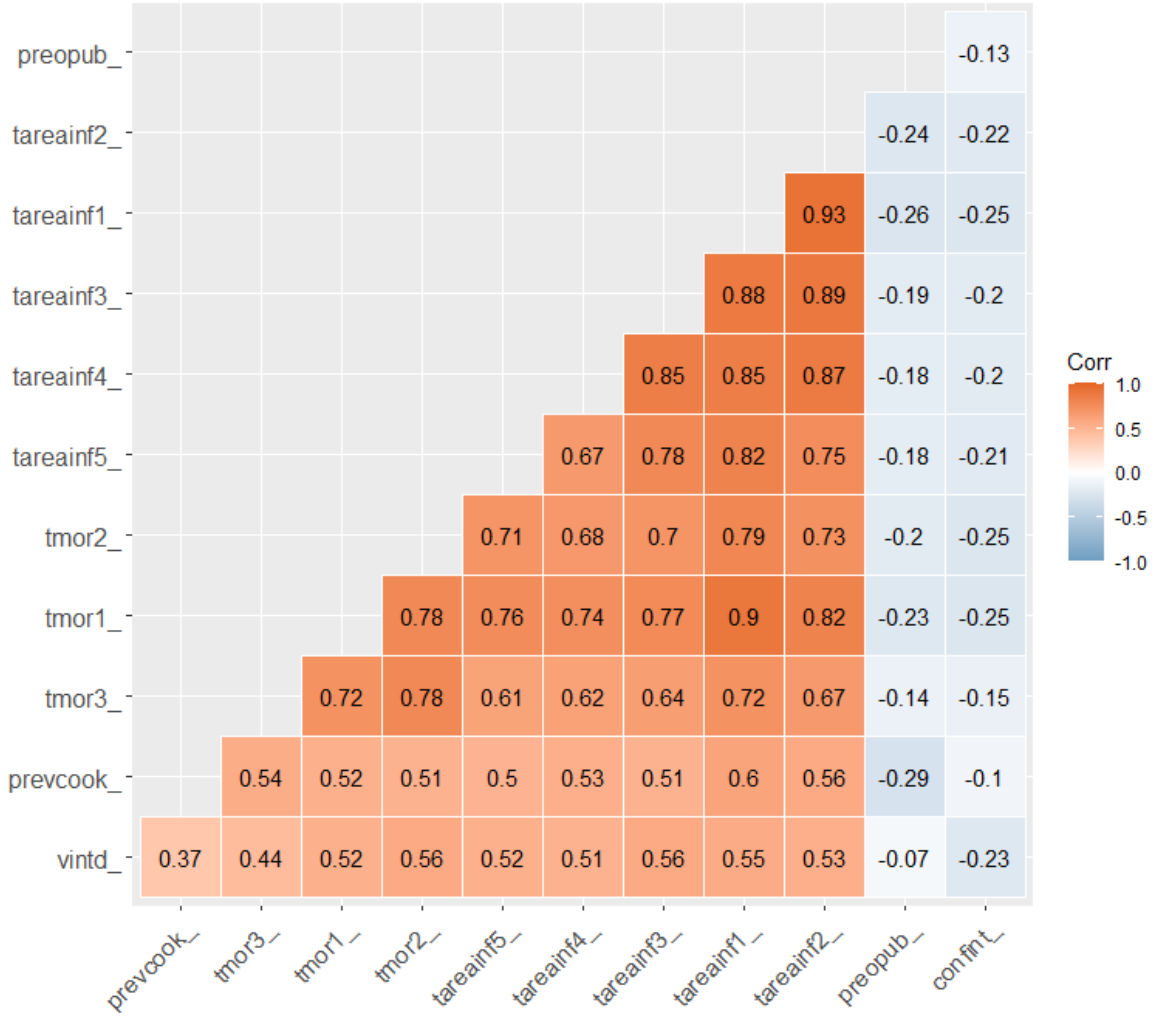


Imagen 18: Representación correlaciones policóricas.

Para elegir el número de componentes principales a retener se tuvieron diferentes criterios en cuenta, todos basados en la matriz de correlaciones policórica. En la Imagen 19, se muestra el resultado del "Método de Catell". Según este método parece conveniente retener las dos primeras Componentes Principales ya que los dos primeros autovalores se encuentran claramente por encima de la recta que se puede trazar conteniendo en su entorno al resto de autovalores. Además, se puede apreciar que estos autovalores obtienen valores por encima de uno, que también se corresponde con la teoría.

Además, para que se considere suficiente, la variabilidad explicada debe de estar por encima del 70%. Según el resultado, como se aprecia en la Imagen 20 y en la Imagen 21, con dos componentes principales se logra explicar un 71% de la variabilidad de la información, por lo que finalmente se retuvieron solo dos. Con estas dos componentes se pueden resumir los datos con la mínima pérdida de información.

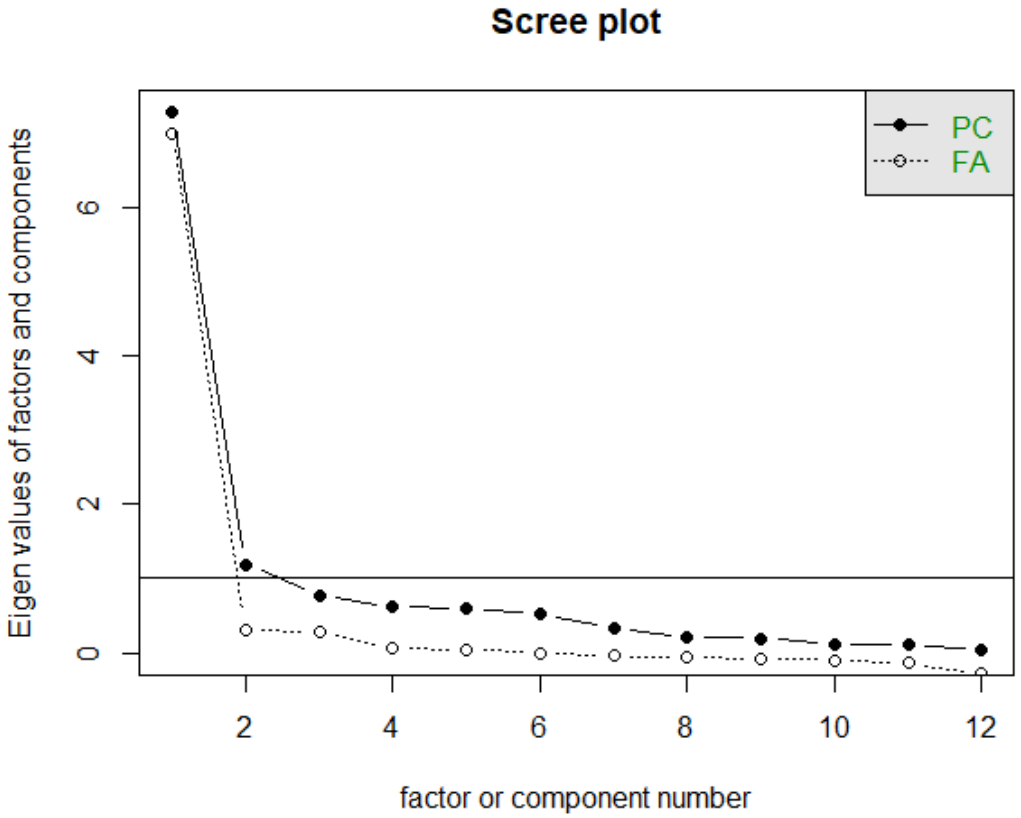


Imagen 19: Representación del número de componentes principales.

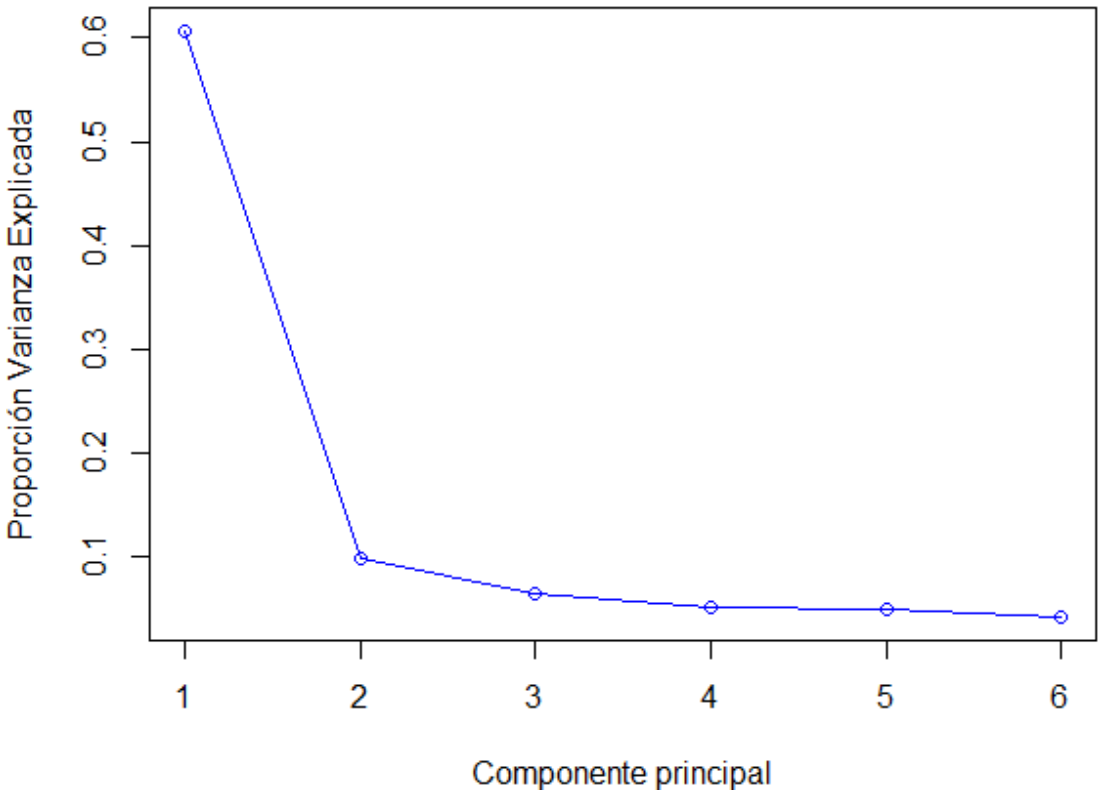


Imagen 20: Proporción de varianza explicada por cada componente principal.

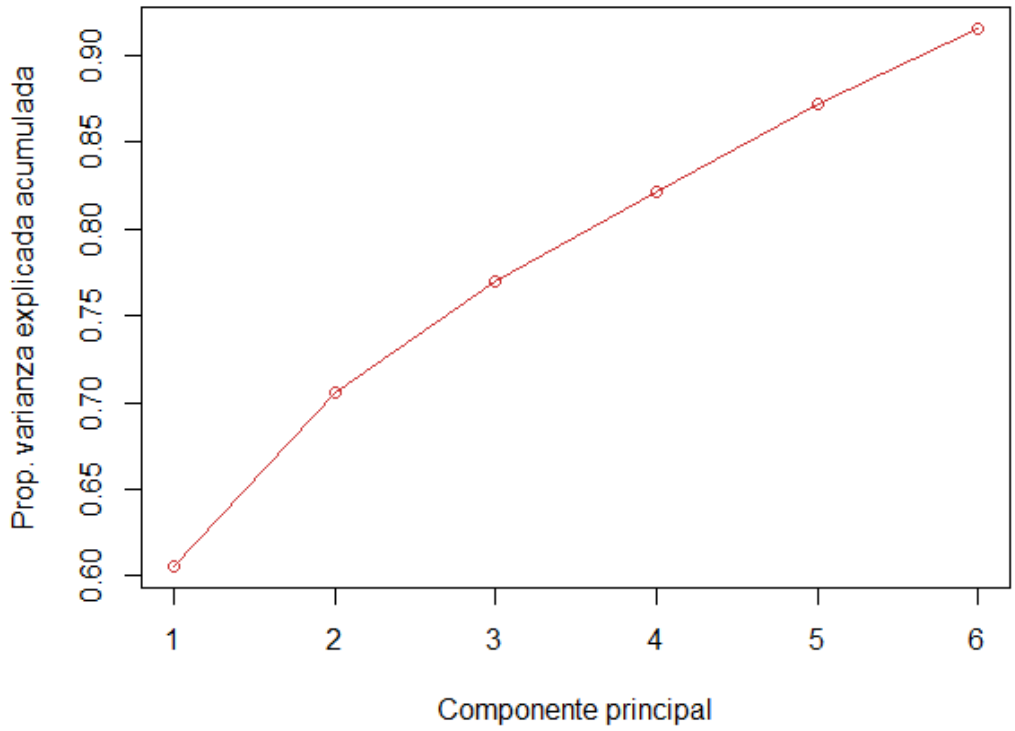


Imagen 21: Proporción de la varianza acumulada por cada componente principal.

Los resultados de la realización de este análisis se muestran a continuación. En la Tabla 6 se muestra la comunalidad final para cada una de las variables indicando la varianza común explicada. Luego, en la Tabla 7 se representan la especificidad, o la varianza residual, es decir, la varianza que no ha sido explicada.

Tabla 6: Comunalidad final.

vintd_	**tmor1_**	**tmor2_**	**tmor3_**	**tareainf1_**	**tareainf2_**
0,4621026	0,81444429	0,74259025	0,63245522	0,93058137	0,86502946
tareainf3_	**tareainf4_**	**tareainf5_**	**prevcook_**	**preopub_**	**confint_**
0,82089051	0,76402283	0,72145398	0,50185713	0,64781643	0,56374616

Tabla 7: Especificidad de las variables.

vintd_	**tmor1_**	**tmor2_**	**tmor3_**	**tareainf1_**	**tareainf2_**
0,5378974	0,18555571	0,25740975	0,36754478	0,06941863	0,13497054
tareainf3_	**tareainf4_**	**tareainf5_**	**prevcook_**	**preopub_**	**confint_**
0,17910949	0,23597717	0,27854602	0,49814287	0,35218357	0,43625384

Finalmente, en la Tabla 8 se muestra la matriz de las cargas en las componentes retenidas. Aquí se puede observar cuáles variables están recogidas por cada una de las componentes principales. Por ejemplo, se puede ver que las variables relacionadas con conocimientos y tareas informáticos realizados, y el uso de internet, están recogidas principalmente por la primera componente principal. Por otro lado, aquellas variables que recogen el grado de preocupación respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida y el grado de confianza en

Internet, están recogidas principalmente por la segunda componente principal. También en la Tabla 8 se puede ver que los autovectores de estas variables con respecto a la PC1 y todos son mayores que 0,6. Esto pasa de igual manera con los autovectores de las variables para la segunda componente. Por otro lado, vale la pena destacar que la segunda componente principal, la cual recoge la información de las variables *preopub* y *confint,* tiene para esta última un vector con dirección opuesta a PC2.

Tabla 8: Matriz de las cargas en las componentes retenidas.

	PC1	PC2
*vintd_*	**0,64438716**	0,21648969
*tmor1_*	**0,90243995**	0,00681413
*tmor2_*	**0,86020056**	0,05143194
*tmor3_*	**0,79502032**	-0,01994777
*tareainf1_*	**0,96448602**	-0,0186573
*tareainf2_*	**0,92973955**	-0,0247757
*tareainf3_*	**0,90582659**	0,01920138
*tareainf4_*	**0,87404008**	0,00876197
*tareainf5_*	**0,84899387**	0,02575629
*prevcook_*	**0,65746024**	-0,26382412
*preopub_*	-0,26001984	**0,76171262**
*confint_*	-0,27408549	**-0,69901595**

Tabla 9: Medidas ajuste del Análisis Componentes Principales.

	PC1	PC2
Eigenvalue	7,28	1,19
Proportion Var	0,61	0,10
Cumulative Var	0,61	0,71
Mean item complexity = 1,1,		
Test of the hypothesis that 2 components are sufficient,		
The root mean square of the residuals (RMSR) is 0,07 with the empirical chi square 6414,34 with prob < 0		
Fit based upon off diagonal values = 0,99		

En la Tabla 9 se puede apreciar que los resultados confirman la validez del modelo ya que el RMSR = 0,07 (teóricamente el modelo representa una solución adecuada cuando RMSR es menos que 0,08). Los dos autovalores asociados a las componentes seleccionadas valen 7,28 y 1,19, y en términos de porcentajes explican el 61% y el 10% de la variabilidad respectivamente. Teniendo un total de variabilidad explicada del 71%.

En la Imagen 22 se encuentra una representación tanto de algunos individuos (al tener una muestra tan elevada se decidió representar solo 100 individuos para mejorar la interpretación del gráfico) como de cada una de las variables en los ejes de las dos componentes principales retenidas. También, según se puede observar en la representación gráfica, para la segunda componente principal hay dos variables, *preopub* y *confint* como ya se mencionó anteriormente. Estas variables tienen autovectores próximos a 1 y -1 respectivamente, lo cual indica que hay una correlación negativa entre ambos grupos de variables. Es decir que, si un grupo de individuos aumenta su valor sobre una de esas variables, ese mismo grupo de individuos lo disminuye sobre la otra, como quedó en evidencia en la tabla de correlaciones. Por ejemplo, la variable *confint,* al tener un valor de coordenada cercano a -1 sobre el segundo eje, establece una clasificación de los individuos según valores decrecientes con relación a los valores sobre esa variable.

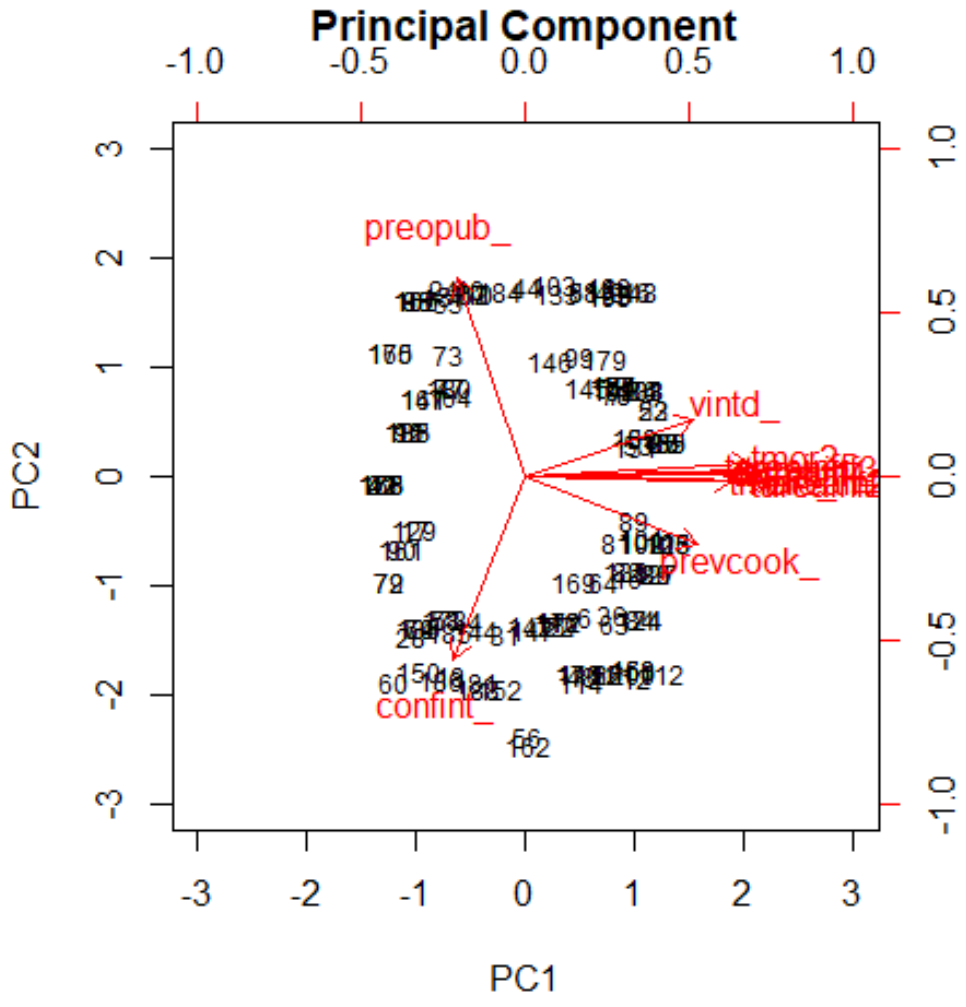


Imagen 22: Representación de las variables y algunos individuos en el espacio de las componentes 1 y 2.

Se puede decir que, dentro de cada grupo de variables, la correlación existente entre las variables que los constituyen está próximo a uno, lo que hace que la distancia entre esas variables esté cercana a cero. Es por ello que, para la componente principal uno se

puede afirmar que los grupos de individuos que poseen varios conocimientos informáticos también hacen un uso frecuente de internet. Por otro lado, las variables *preopub* y *confint* presentan un valor de coordenada respecto al primer eje (CP1) cercano al cero (menos que 0,3 en absoluto). De ahí que la correlación tanto con la CP1 como con los grupos de variables que esta recoge no es alta y, por lo tanto, no relacionable. Sin embargo, vale la pena mencionar que *confint* está negativamente relacionada con la PC1.

Por todo lo expresado anteriormente se decidió renombrar a la PC1 como *CPConInf_UsoInt* y a la PC2 como *CPConfSegInt* para utilizarlas en los análisis posteriores.

4.2 Análisis clúster jerárquico

En esta sección se procede a seleccionar las variables más relevantes para el análisis clúster y la selección del número de clúster a realizar.

4.2.1 Selección de variables

Entre el conjunto de variables disponibles en la muestra en cuestión, se decidió llevar a cabo una selección de aquellas variables que aportarían más información y que a la vez fueran más relevantes para los próximos análisis a realizar.

La siguiente tabla muestra un resumen de los mejores resultados obtenidos para el análisis clúster con las distintas combinaciones de variables de algunas de las pruebas realizadas. En la primera columna se indica el número de variables. En la segunda, la cantidad de clústeres necesarios para explicar la variabilidad recogida, la cual es medida a partir del estadístico R^2 que aparece, a su vez, en la tercera columna. Las pruebas consistieron en ir probando con todo el set de variables disponibles e ir quitando aquellas que disminuían el valor del R^2 y dejando aquellas variables que aumentaban su valor.

Tabla 10: Resultado de las pruebas de selección de variables.

Cantidad de variables	Número de clústeres	Variabilidad explicada (%)
20	50	70%
14	26	70,4%
10	10	71,1%

Después de todas las pruebas realizadas, finalmente se decidió continuar con las 10 variables que con 10 clústeres lograban explicar más de un 70% de la variabilidad. Esas variables, su descripción y valores se muestran en la tabla a continuación.

Tabla 11: Variables seleccionadas para realizar análisis clúster.

Nombre de la variable	**Descripción**	**Valores**
NIVELEST	Estudios terminados	Sin estudios o primaria, Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…), Segunda etapa de educación secundaria y educación postsecundaria no superior, Educación Superior
SEXO	Sexo de la persona seleccionada	Hombre, Mujer
EDAD	Edad de la persona seleccionada	16 - 74 años
SIT_LAB	Situación laboral en la que se encuentra	Estudiante, Jubilado ó prejubilado, Parado, Otros, Labores del hogar, Trabajando
TOT_MH	Total de MIEMBROS del hogar	1 - 13
TIP_JOR	Tipo de hornada en trabajo principal	A tiempo completo, A tiempo parcial, No aplica
OCUPACION2	Ocupación principal: Trabajador TIC / No TIC	No Aplica, No TIC, TIC
ING_HOG	Ingresos mensuales netos del hogar	Más de 2.500, De 1.600 a menos de 2.500 euros, De 900 a menos de 1.600 euros, Menos de 900 euros, NS/NR
CPConInf_UsoInt	Componente principal que recoge conocimientos informáticos y uso de internet	-1,3 – 1,33
CPConfSegInt	Componente principal que recoge confianza en internet y preocupación por el uso de datos para publicidad	-2,47 – 1,79

4.2.2 Análisis clúster jerárquico

Una vez se definió el conjunto de variables a utilizar, se procedió a determinar el número óptimo de clústeres a escoger. Debido a la existencia de variables mixtas en el set de datos, es decir, tanto variables categóricas como de intervalo, el método para determinar la distancia entre los individuos que se utilizó fue el *DGOWER* (distancia de Gower), como se explicó en el capítulo de metodología. Esto se le indica en la sentencia del *proc distance* en el programa SAS que fue el utilizado para este paso del estudio.

```
proc distance data=HIoT.FichClus4 method=DGOWER
out=HIoT.distgower;
var nominal(sexo_     REP_ing_hog_     REP_nivelest_
     REP_sit_lab_    REP_ocupacion2_    REP_tip_jor_    );
var INTERVAL( tot_mh_ edad_  CPConInf_UsoInt     CPConfSegInt
);
id IDEN;
run;
```

- Determinar número de clúster

Para las variables seleccionadas y la distancia calculada se procedió a determinar el número de clústeres óptimos a utilizar a través de la sentencia *proc cluster*. La idea es formar grupos que sean lo más homogéneos dentro de ellos, pero lo más diferente entre ellos. Debido a esto, para la elección adecuada del número de agrupaciones, se buscó

máximos relativos en el estadístico Pseudo F, y valores pequeños del Pseudo T-Squared seguidos de un máximo relativo o un incremento excesivo. Adicionalmente, el número de agrupaciones mínimo que empezaremos a valorar será a partir de un R^2 superior al 50%.

La siguiente tabla recoge la elección de clústeres con cada método empleado, así como el porcentaje de variabilidad explicada con cada uno según los criterios explicados con anterioridad.

Tabla 12: Características principales clústeres.

Método empleado	**Clústeres**	**Variabilidad**
Distancia de Ward	10	70,8%
Enlace medio	10	71,1%
Distancia entre centroides	16	70,2%

Con la información de la tabla anterior, se concluye que fue el método de "Enlace Medio" el que mejores resultados ofreció, ya que un con número menor de clústeres, respecto al de "Distancia entre centroides" y el mismo número respecto al método de "Ward", consiguió explicar mayor variabilidad de la información.

A continuación, se detallan más resultados de este método y los motivos de la selección de 10 clústeres. Por reducir los resultados obtenidos, la Tabla 13 y los siguientes gráficos solo muestran los resultados hasta de 5 a 11 clústeres ya que, además, con esos son suficientes para explicar el porqué de la selección final. Por ejemplo, la Imagen 23 recoge el valor de los estadísticos en cada clúster, vemos que en el caso del *Pseudo F* las variaciones no son tan acentuadas como en el caso del *Pseudo T Squared*. Siguiendo las explicaciones de la metodología, en el clúster 7 y 9 se observan un máximo relativo en el *Pseudo T Squared*, y a continuación un fuerte descenso del mismo para los clústeres 8 y 10 respectivamente. Esto mismo lo podemos comprobar analizando la Tabla 13, donde se encuentran los valores de las medidas en las cuales se basó el criterio de selección de clústeres. En la misma se pueden ver los valores que toman algunos indicadores como son: el R-cuadrado semiparcial, el R-cuadrado, el Estadístico pseudo F, el T-cuadrado pseudo, y la distancia de los centroides.

Tabla 13: Resultados obtenidos para el método de "Enlace medio".

Número de clusters	**Frec**	**R-cuadrado semiparcial**	**R-cuadrado**	**Estadístico pseudo F**	**T-cuadrado pseudo**	**Norm Centroid Distance**
11	1378	0,0095	,716	2868	376	0,6756
10	753	0,0053	,711	3105	202	0,6814
9	4531	0,0404	,670	2891	1901	0,6847
8	443	0,0004	,670	3299	17,0	0,7209
7	5513	0,0304	,639	3364	1042	0,7263
6	5956	0,0192	,620	3717	565	0,7785
5	2462	0,0258	,594	4171	924	0,7823
	Distancia del cuadrado de la raíz media entre observaciones: 0.506586					

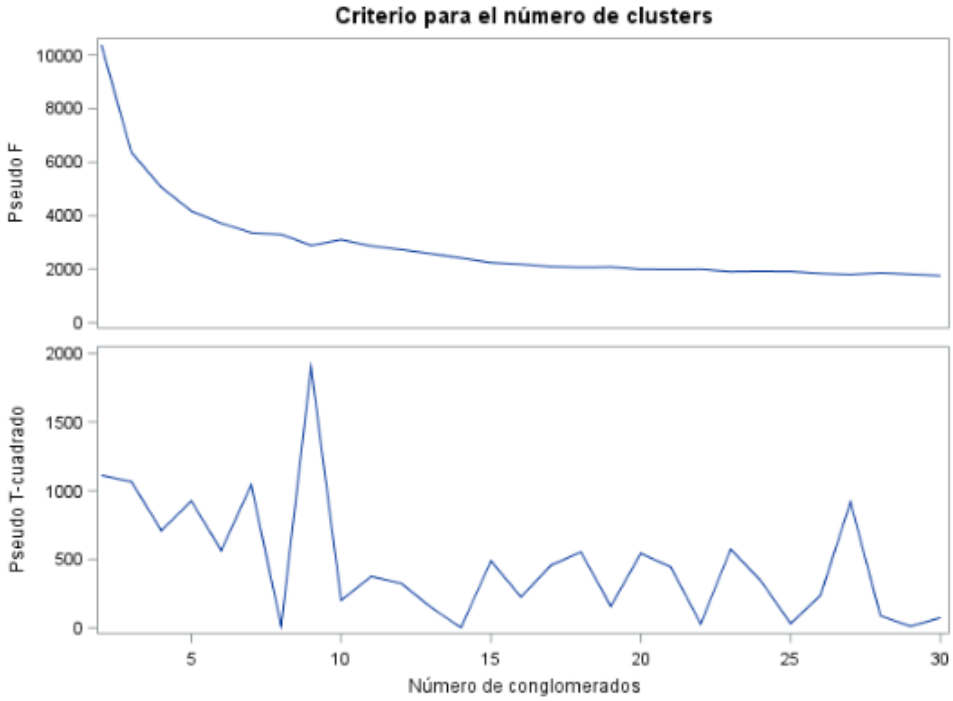


Imagen 23: Representación de los valores de los estadísticos Pseudo F y Pseudo T Squared para cada clússter.

Además, si vemos la representación del R-cuadrado versus el R-cuadrado semiparcial en la Imagen 24 se observan también un máximo y un mínimo respectivamente para los clústeres 8 y 10 también, siendo visiblemente mayor la diferencia para 10 clústeres. Para comprobar esto último se presenta la diferencia en la Tabla 14.

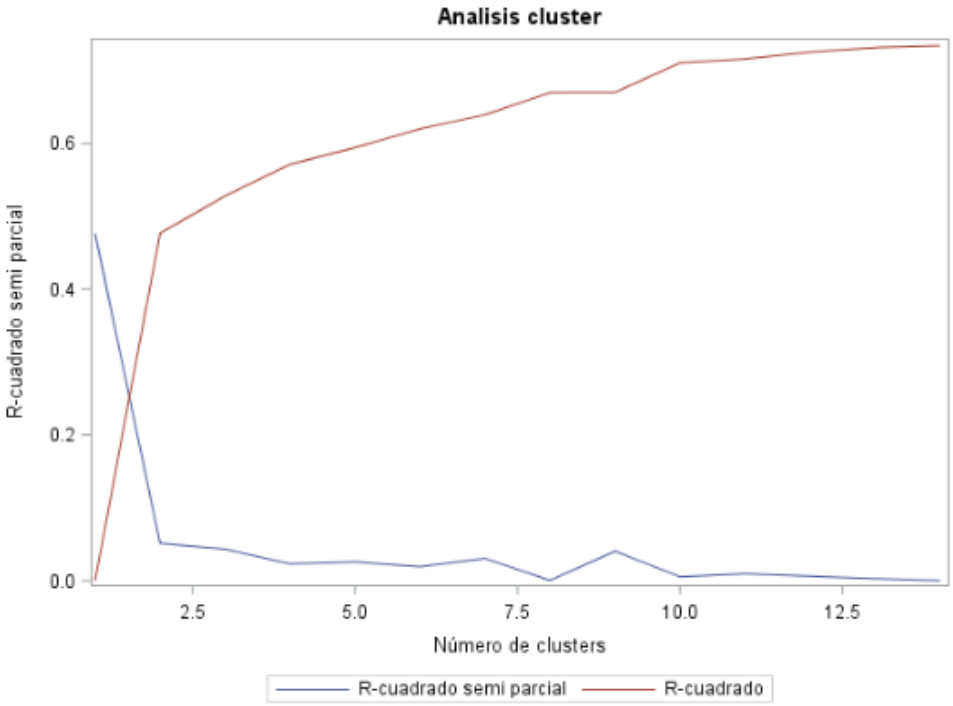


Imagen 24: Representación del R-cuadrado versus el R-cuadrado semiparcial.

Tabla 14: Representación del R-cuadrado versus el R-cuadrado semiparcial.

Clúster	**_RSQ_**	**_SPRSQ_**	Diferencia
8	0,66974	0,00041	0,66933
10	0,71057	0,00528	0,70529

Como método decisor final en la elección de si utilizar 8 o 10 clústeres, se tuvo en cuenta el coeficiente de la silueta de cada clúster. El coeficiente de silueta contrasta la distancia media a los elementos del mismo clúster con la distancia media a los elementos de otros clústeres. Debido a esto, los clústeres con un valor de silueta alto se consideran bien agrupados. El cálculo de este coeficiente se realizó en RStudio mediante la siguiente sentencia:

```
sil_width <- c(NA)

for(i in 6:12){

  pam_fit <- pam(gower_mat3,

          diss = TRUE,

          k = i)

  sil_width[i] <- pam_fit$silinfo$avg.width  }
```

En la Imagen 25 se aprecia cómo el coeficiente de "silhouette" obtiene mucho mejores resultados para 10 clústeres que para 8. Consecuentemente se decidió agrupar a todos los individuos de la muestra en 10 clústeres.

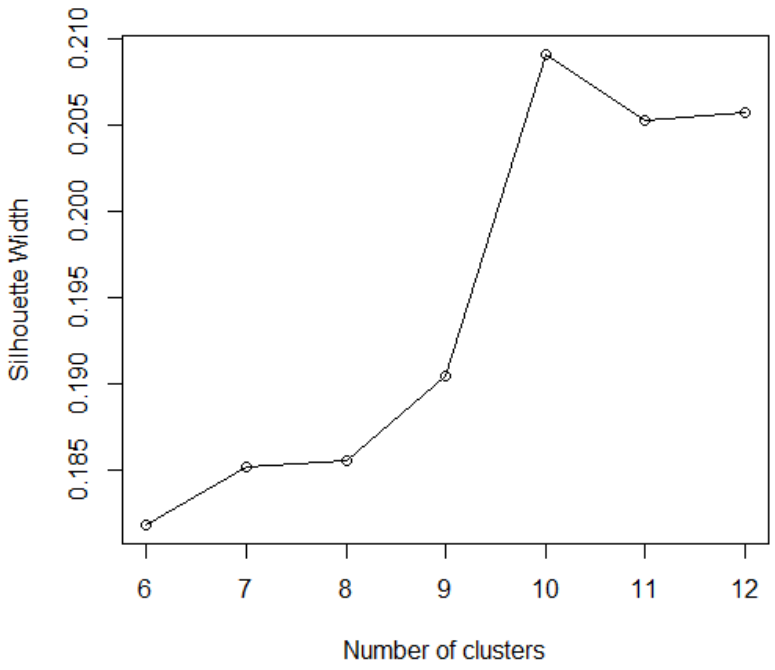


Imagen 25: Representación del coeficiente de silueta para cada clúster.

4.3 Análisis clúster PAM

Al tener decido el número de clústeres a agrupar, se procedió a realizar el análisis clúster no jerárquico con el algoritmo k-medoids en R-Studio utilizando el lenguaje de programación R. A través del paquete *cluster*, la sentencia utilizada fue la que se muestra a continuación donde a la función *pam* se le pasa como argumento la matriz de distancia calculada con el método de "GOWER".

pam_clus10 <- pam(gower_mat3, diss = TRUE, k = 10)

Para calcular esta distancia se utilizó la función *daisy()* del mismo paquete con las variables escogidas en la sección anterior y a través de la siguiente sentencia:

gower_df3 <- daisy(fich_clusPam[,-c(1)],metric = "gower")

Gracias al lenguaje de programación R, con la salida de la función *daisy()* también podemos ver cuáles son los individuos que más se parecen y cuáles son los que más difieren entre sí. Por ejemplo, en la Tabla 15 se exponen los individuos que son muy parecidos entre sí, siento ambos hombres cuyo hogar tiene ingresos de más de 2500€ mensuales, que son estudiantes de 20 años y que han realizados diversas actividades informáticas.

Tabla 15: Características del par de individuos más similar.

serial_	**Individuo 1**	**Individuo 2**
*sexo_*	Hombre	Hombre
*tot_mh_*	4	4
*REP_ing_hog_*	Más de 2.500	Más de 2.500
*REP_nivelest_*	Segunda etapa de educación secundaria y educación postsecundaria no superior	Segunda etapa de educación secundaria y educación postsecundaria no superior
*REP_ocupacion2_*	No Aplica	No Aplica
*REP_sit_lab_*	Estudiante	Estudiante
*REP_tip_jor_*	No Aplica	No Aplica
CPConInf_UsoInt	1.02392441	1.01883519
CPConfSegInt	0.34905132	0.35227511
*edad_*	20	20

Por otro lado, la Tabla 16 muestra las personas con características que los hacen muy diferentes entre ellos. Por ejemplo, se ve que son de diferente sexo y que viven en hogares con una totalidad de personas muy diferentes, mientras que el hombre vive en un hogar unipersonal, la mujer vive en un hogar conteniendo 6 personas. También se aprecian diferencias entre el nivel de estudios terminados, así como en la situación

laboral actual. Adicionalmente se ve cómo la diferencia de edad es bastante relevante lo cual influye en los conocimientos informáticos, ya que la joven presenta mayores conocimientos que la persona con 61 años.

Tabla 16: Características del par de individuos que más difieren.

serial_	**Individuo 1**	**Individuo 2**
*sexo_*	Mujer	Hombre
*tot_mh_*	6	1
*REP_ing_hog_*	NS/NR	De 900 a menos de 1.600 euros
*REP_nivelest_*	Educación Superior	Sin estudios o primaria
*REP_ocupacion2_*	No Aplica	No TIC
*REP_sit_lab_*	Estudiante	Trabajando
*REP_tip_jor_*	No Aplica	A tiempo completo
CPConInf_UsoInt	1.24879242	-0.74139158
CPConfSegInt	-1.79700757	1.71469239
*edad_*	21	63

4.3.1 Caracterización de los clústeres formados

Una vez ejecutadas las sentencias anteriores se procedió a realizar un estudio del resultado del análisis de k-medoids.

En el lenguaje de programación R existen dos maneras de investigar y detallar los resultados de este ejercicio de agrupación, con el fin de derivar alguna interpretación relevante para el negocio. Una es realizar un resumen de cada clúster, utilizando la función summary() en R que muestra las características de los medoides de cada clúster. La otra es mediante la visualización en un espacio de menor dimensión, con t-SNE, utilizando la función Rtsne() en R. t-Distributed Stochastic Neighbor Embedding (t-SNE) es una técnica para la reducción de la dimensionalidad que es particularmente adecuada para la visualización de conjuntos de datos de alta dimensión.

Primeramente, se utilizó la función *summary()* para sacar un resumen de cada clúster, así como la cantidad de elementos dentro de cada categoría. Esto es sumamente útil en el caso de las variables categóricas con más de dos categorías pues permite hacer una caracterización más adecuada de cada clúster. Otra ventaja para destacar del algoritmo PAM en R con respecto a la interpretación es que los medoides sirven como ejemplares de cada clúster como se muestra en la Imagen 26.

La obtención de este resumen se realizó con la siguiente sentencia:

```
pam_results10 <- fich_clusPam[,-c(1)] %>%
  mutate(cluster = pam_clus10$clustering) %>%
```

```
  group_by(cluster) %>%
  do(the_summary = summary(.))
pam_results10$the_summary
```

A continuación, se muestran diferentes gráficos con los resultados del análisis clúster realizado. Como se mencionaba anteriormente, en la Imagen 26 se pueden apreciar cuáles son los mediods de cada clúster, así como los valores que toman cada una de las variables utilizadas para los mismos.

Cluster	1	2	3	4	5	6	7	8	9	10
sexo_	Hombre	Mujer	Mujer	Hombre	Mujer	Mujer	Mujer	Hombre	Hombre	Hombre
tot_mh	2	3	3	2	2	2	2	3	3	3
REP_ing_hog	De 1.600 a menos de 2.500 euros	Más de 2.500	De 1.600 a menos de 2.500 euros	De 900 a menos de 1.600 euros	De 900 a menos de 1.600 euros	De 900 a menos de 1.600 euros	Menos de 900 euros	Más de 2.500	NS/NR	De 900 a menos de 1.600 euros
REP_nivelest_	Educación Superior	Educación Superior	Educación Superior	Segunda etapa de educación secundaria y educación postsecundaria no superior	Segunda etapa de educación secundaria y educación postsecundaria no superior	Segunda etapa de educación secundaria y educación postsecundaria no superior	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	Educación Superior	Segunda etapa de educación secundaria y educación postsecundaria no superior	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)
REP_ocupacion2_	No TIC	No TIC	No TIC	No Aplica	No TIC	No Aplica	No Aplica	No TIC	No Aplica	No TIC
REP_sit_lab_	Trabajando	Trabajando	Trabajando	Jubilado ó prejubilado	Trabajando	Parado	Parado	Trabajando	Estudiante	Trabajando
REP_tip_jor_	A tiempo completo	A tiempo completo	A tiempo completo	No Aplica	A tiempo completo	No Aplica	No Aplica	A tiempo completo	No Aplica	A tiempo completo
CPConInf_UsoInt	0.53344492	0.77121942	0.52388077	-0.73084473	-0.226589	-0.27814064	-1.00838725	0.896234	0.78614992	-0.60831451
CPConfSegInt	0.27221761	0.31630893	0.32571624	0.23187935	-0.52185356	0.31049077	-0.48057598	0.04893549	0.30496	-0.21266645
edad_	42	41	43	61	48	55	62	45	22	49

Imagen 26: Resumen de características de los medoides de cada clúster.

Por otro lado, la Imagen 27 ayuda a ver la formación de cada uno de los clústeres de forma más visual. En la misma se refleja la distribución de las características de cada medoid formado por el algoritmo utilizado. Es básicamente un mapa de calor para representar el número de observaciones que entran en cada nivel de cada variable categórica. El verde más intenso corresponde a un mayor número relativo de observaciones dentro de un clúster mientras que la ausencia de color, en este caso gris, significa la ausencia de individuos pertenecientes a esa categoría. En este caso se puede ver cómo las categorías de mujer y hombre son las que más cantidades tienen en la mayoría de los clústeres. Si se analiza, por ejemplo, el clúster 6 se puede apreciar visualmente está mayormente formado por mujeres cuyos ingresos del hogar están principalmente en el rango de 900 a 1600€ mensuales debido a que la tonalidad de esa categoría es mayor que el resto de las categorías del rango de ingresos. Otra conclusión que se puede sacar de esta visualización es que la mayoría de las personas que se encuentran trabajando lo hacen en una ocupación No TIC ya que su tonalidad es parecida para aquellos clústeres donde tienen presencia. También se puede ver que, en el caso de las diferentes categorías de nivel de estudios terminados, la cantidad de personas que tienen una Educación Superior es relevante para casi todas las agrupaciones realizadas.

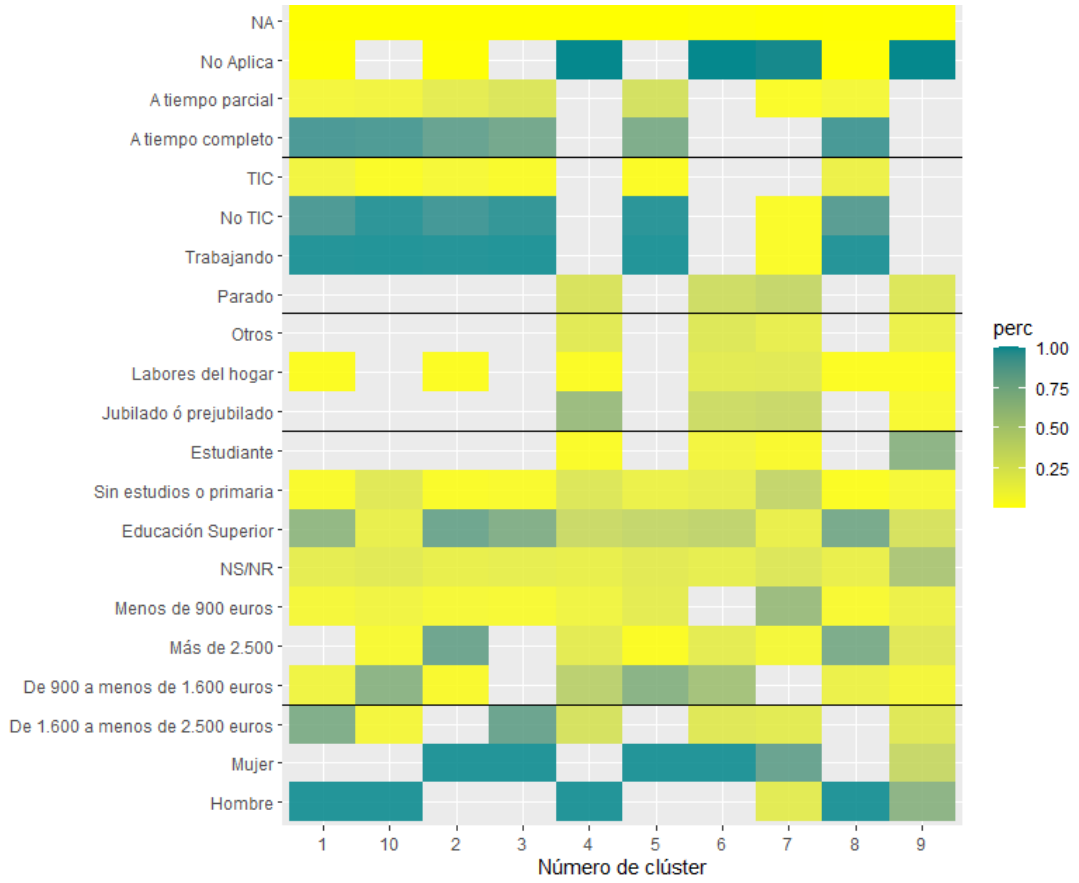


Imagen 27: Distribución de las características en los clústeres.

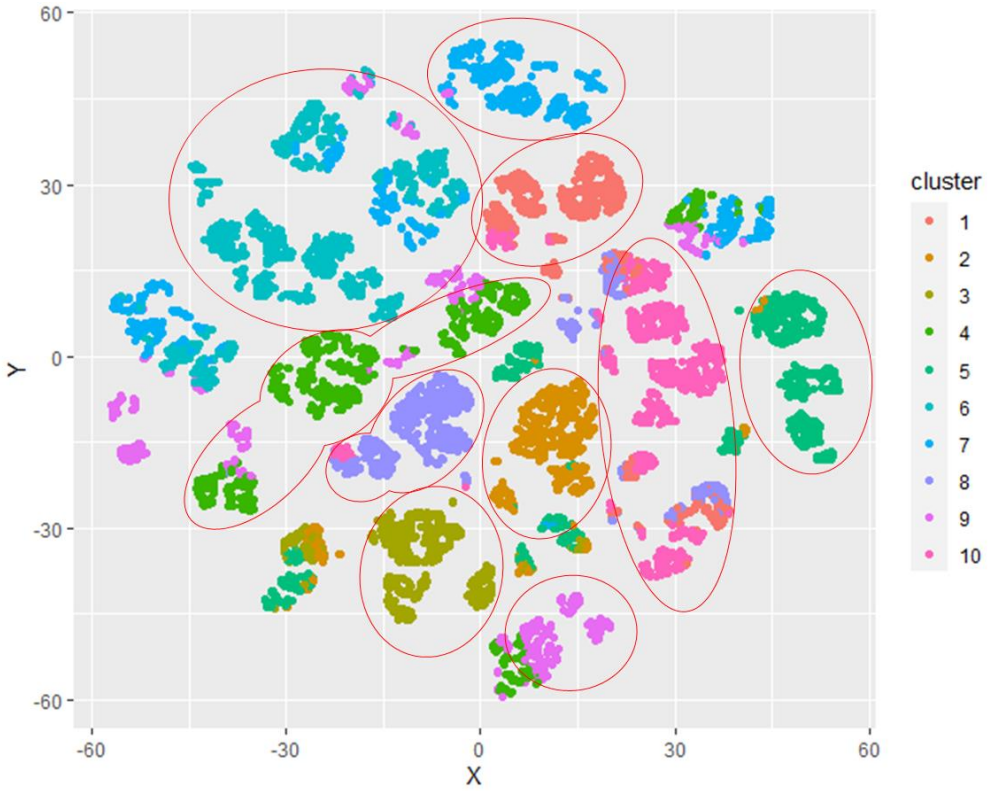


Imagen 28: Representación visual de los clústeres.

También se pasó a realizar una representación visual de los clústeres para poder ver cómo estaban agrupados y ver cuáles de ellos estaban peor representados. En la Imagen 28 se puede ver cómo, de manera general todos los clústeres se encuentran bien definidos con pocos individuos fuera del alcance del conglomerado al que pertenece. Sin embargo, hay algunos clústeres para lo cual esto no se cumple y es el caso del clúster 9 ya que se puede apreciar cómo los individuos que pertenecen al mismo están bastante dispersos y mezclados con el resto, por lo que se puede inferir que es donde mayor variabilidad de características habrá. Vale la pena mencionar que esta representación para el caso de 8 clústeres era mucho menos homogénea ya que las agrupaciones se encontraban mucho menos apreciables y bien delimitados.

Gracias a este resumen y junto con el detalle de cada uno de los clústeres que se encuentra en el Anexo D se pasó a realizar la caracterización de cada uno.

- Clúster 1: Formado solamente por hombres cuyo hogar está formado principalmente de 1 a 3 personas en su hogar (media 2 y como máximo 7; 75% de los individuos conviven con un máximo de 3 personas). En términos educativos, la mayoría de las personas de este grupo tienen una educación superior principalmente y de Segunda etapa de ES y EPNS. Profesionalmente hablado, todos los miembros de este clúster se encuentran trabajando en profesiones "No TIC" (Ver Anexos) y a tiempo completo principalmente. La edad media de este grupo ronda los 42 años, con edad del primer cuartil de 18 y del tercero 53 años (principalmente edad entre 40-50). Los ingresos de esos hogares se encuentran de media entre los 1600 a 2500€ mensuales. Además, la variable *CPConInf_UsoInt* es mayormente positiva en este clúster por lo que se infiere que las personas presentan conocimientos informáticos y realizan un uso frecuente de internet. Lo mismo pasa con la variable *CPConfSegInt,* la mayoría de los individuos de este grupo presentan valores positivos de *CPConfSegInt*, por lo que se infieren que su nivel de confianza en internet es más bien elevado.

- Clúster 2: Compuesto en su totalidad por mujeres que conviven en un hogar donde de media existen 3 miembros (como máx. 14) con ingresos en el rango de más de 2500€ al mes. En temas de educación, un elevado número de individuos presentan alguna educación superior y se encuentran trabajando en ocupaciones No *TIC*, y a tiempo completo. La edad media de este grupo ronda los 41 años (principalmente edad entre 40-50). En términos de la variable *CPConInf_UsoInt,* esta es sobre todo positiva con valores más altos que Clúster 1, por lo que se puede entender que las personas pertenecientes a este grupo han realizados la mayoría de las tareas informáticas en la encuesta y realiza un uso bastante frecuente de internet. En este caso, la variable *CPConfSegInt* también presenta valores en su mayoría positivos y elevados, por lo que se puede decir que el grado de preocupación respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida es bajo o nulo.

- Clúster 3: Pertenecen solo mujeres cuyo hogar está formado en media por 3 personas, pero como mucho conviven 6 y el 75% de los individuos conviven con

un máximo de 4 personas. Los ingresos totales de los hogares de estas personas se encuentran en el rango de 1600 a 2500€ principalmente. Respecto al nivel de formación, estas personas presentan una educación superior principalmente (algo de segunda). Además, en este grupo sus miembros se encuentran trabajando a tiempo completo y presentan una edad media de 45 años (principalmente edad entre 40-50). Por otro lado, *CPConInf_UsoInt* es mayormente positiva en este clúster por lo que se infiere que las personas presentan conocimientos informáticos y realizan un uso frecuente de internet. Lo mismo pasa con la variable *CPConfSegInt,* la mayoría de los individuos de este grupo presentan valores positivos de *CPConfSegInt*, por lo que se infieren que su nivel de confianza en internet es más bien elevado.

- Clúster 4: Conformado en su totalidad por hombres con miembros en su hogar entre 2 y 3 mayoritariamente (máx. 10 y tercer cuartil es igual a 3 miembros). Los ingresos de estos hogares se encuentran repartidos entre los diferentes rangos salariales establecidos, pero sobre todo se concentran entre los rangos de más de 900 y menos de 2500 ("cantidad de más de 2500" > "cantidad de menos de 900"). Al igual que el rango salarial, el nivel de estudios terminados de este grupo varía entre algunas de las categorías siendo la educación superior la que más se repite y la mayoría de los miembros habiendo terminado niveles superiores a los de la educación primaria por lo que tienen algo de formación (81%). En términos de situación laboral, las personas que conforman el Clúster 4 son en su mayoría jubilados o prejubilados, y a continuación también parados. Esto explica que la media de edad se encuentre en los 60 años con solo un 25% por debajo de los 55. Respecto a la variable *CPConInf_UsoInt* en este caso su valor medio es negativo y elevado y un 75% de los individuos también presentas valores menores que cero por lo que se pudiera decir que estas personas no realizan un uso muy frecuente de internet y tampoco realizan muchas actividades informáticas, no tienen claro su confianza en internet y su preocupación es más bien nula.

- Clúster 5: En su totalidad formado por mujeres que conviven en su hogar con otras dos o tres personas y como máximo son 7 los miembros de su hogar. Los ingresos de los hogares de estas personas se encuentran mayoritariamente en el rango de los 900€ y 1600€ mensuales, aunque existen algunos con menos de 900€ también. En términos de estudios, a modo general los individuos de este clúster tienen diferentes niveles de estudios. La categoría que toma el medoide de este clúster como estudios terminados son los de Segunda etapa de educación secundaria y educación postsecundaria no superior, por lo que no clasifican como los estudios más altos. Es interesante resaltar que la cantidad de individuos en este grupo que no tienes estudios superiores es mayor que la cantidad que los tiene. Adicionalmente, estas personas se encuentran trabajando en ocupaciones No Tic y en general a tiempo completo, aunque existe un porcentaje que es a tiempo parcial. Respecto a la edad, se puede decir que como media tienen alrededor de 47 años, pero el rango mayoritario se encuentra entre los 40 y los 55 años. Por otro lado, la variable *CPConInf_UsoInt* toma valores casi siempre negativos por lo que se puede concluir que las mujeres de

este grupo no realizan muchas de las actividades informáticas recogidas en la encuesta y puede ser que su uso de internet tampoco sea muy frecuente. De igual manera, la variable *CPConfSegInt* también toma en su mayoría valores negativos o positivos muy bajos. Esto puede ser un indicador de que la confianza que tienen estos individuos en internet sea poco o nada y en consecuencia, estén algo preocupados por el uso que se les da a sus datos personales en internet.

- Clúster 6: A este grupo pertenecen solo mujeres, que al igual que el caso anterior conviven en su hogar con otras dos o tres personas y como máximo son 7 los miembros de su hogar. Sin embargo, algo en lo que se diferencia con el clúster 5 es que el rango de ingresos en los hogares en los que conviven estas mujeres es algo superior ya que todos se encuentran por encima de 900€ mensuales. Además, existen cantidades significativas de individuos conviviendo en hogares cuyo ingreso se encuentra por encima de los 2500€ mensuales. En términos de estudios, a modo general los individuos de este clúster tienen diferentes niveles de estudios. La categoría que toma el medoide de este clúster como estudios terminados son los de Segunda etapa de educación secundaria y educación postsecundaria no superior, por lo que no clasifican como los estudios más altos. Al igual que en el caso anterior, la cantidad de individuos en este grupo que no tienes estudios superiores es mayor que la cantidad que los tiene. Otra característica diferente entre el clúster 5 y el clúster 6 es que este último los individuos se encuentran sobre todo jubilados o prejubilados, aunque el valor que toma el medoid es de parados contando con una cantidad significativa también. Además, existen varios individuos que se encuentran de voluntariado o realizando labores del hogar. Como era de esperar la edad de este clúster es más bien alta respecto al resto de grupos con una edad media de 52 años y el 25% estando por encima de 41 años. En términos de actividades informáticas realizadas, la variable *CPConInf_UsoInt* toma valores casi siempre negativos por lo que se puede concluir que las mujeres de este grupo no tienen mucho de los conocimientos informáticos preguntados. Respecto a la confianza en internet, como *CPConfSegInt* es más bien positiva se podría decir que estas personas tienen una confianza en internet entre baja y media porque también hay un 25% de valores que son negativos.

- Clúster 7: Este es uno de los clústeres que está compuesto tanto por hombres como por mujeres, aunque en su mayoría son mujeres. Los miembros de los hogares de los individuos de este grupo están entre 1 y 2 miembros en su mayor proporción. Adicionalmente, los ingresos de estos hogares están principalmente en los rangos más bajos que se han considerado en este estudio ya que en su mayoría se encuentran por debajo de 900€ mensuales. Sin embargo, encontramos algunos individuos que se encuentran con hogares con 1600€ y 2500€ al mes. Si se entra en el detalle del nivel de estudio más alto terminado se ve que en este caso es el de Primera etapa de la educación secundaria y similar seguidos de personas que no tienen estudios o solo han terminado la primaria. Consecuentemente se puede afirmar que es uno de los grupos con los niveles más bajos de todos los existentes. Respecto a la situación laboral actual, en este grupo se encuentra concentrado sobre todo en los niveles de Parado, seguido de

jubilados y prejubilados. La variable *CPConInf_UsoInt* en este caso toma valores negativos casi en su totalidad por lo que se puede decir que los individuos de este grupo no realizan casi ninguna de las actividades informáticas encuestadas y puede ser que su uso de internet tampoco sea muy frecuente. El rango de edad de estas personas se encuentra entre los 49 y los 67 años de edad, tomando el medoid el valor de 65 años, haciendo de este clúster los que concentran a las personas con mayor edad del resto. Similarmente, la variable *CPConfSegInt* también toma en su mayoría valores negativos o positivos muy bajos lo cual puede indicar que las personas en este grupo tienen una confianza en internet sea poco o nada.

- Clúster 8: En este grupo se encuentran individuos que son todos hombres viviendo en hogares cutos miembros se encuentran entre tres y cuatro integrantes y como máximo 8. El rango de ingresos percibidos por estos hogares se encuentra bastante dispares ya que, o son más de 2500€ mensuales (lo cual es el valor del medoid) o menos de 1600€ al mes. En este grupo los hombres han terminado en su mayoría algunas de los estudios incluidos en la categoría de Educación superior y además se encuentran también trabajando en ocupaciones No Tic, y a tiempo completo. *CPConInf_UsoInt* sobre todo positiva, conocimientos, uso de internet. Algo preocupado por el uso de su trato en internet y con confianza media-baja. Edad entre 39 y 53 principalmente con media de 45. En este caso la variable *CPConInf_UsoInt* toma valores positivos en la mayoría de los individuos de este clúster por lo que se infiere que las personas presentan conocimientos informáticos, habiendo realizado varias de las tareas informáticas preguntadas y realizan un uso frecuente de internet. Sin embargo, en el caso de la variable *CPConfSegInt,* toma valores tanto positivos como negativos. Consecuentemente se puede decir que estos individuos estarían algo preocupados respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida es bajo o nulo y su confianza en internet es media.

- Clúster 9: En este grupo existen tanto hombres como mujeres, pero en su mayoría son hombres. Estas personas conviven en hogares que tienen como media entre tres y cuatro individuos. En términos de ingresos se puede decir que este grupo está constituido por personas cuyos hogares reciben ingresos mensuales bastante variados considerando los rangos contemplados, pero en su mayoría son de más de 1600€. En este grupo es muy relevante la edad, ya que es el clúster que presenta los individuos más jóvenes, teniendo una edad entre los 18 y 28 años. Es por ello que es de esperar que los estudios en su mayoría no sean todos superiores, sino que se encuentren más bien repartidos entre las diferentes etapas de los rangos estudiantiles considerados. Adicionalmente, si se tiene en cuenta la situación laboral actual, se reafirma lo planteado anteriormente ya que en su mayoría se clasifican como estudiantes. En términos de la variable *CPConInf_UsoInt,* esta es sobre todo positiva con valores uno de los valores más altos de todos los clústeres, por lo que se puede entender que las personas pertenecientes a este grupo han realizados la mayoría de las tareas informáticas en la encuesta y realizan un uso bastante frecuente de internet. En

este caso, la variable *CPConfSegInt* también presenta valores en su mayoría positivos y elevados. Es por ello que se puede decir que el grado de preocupación respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida en este caso es prácticamente irrelevantes.

- Clúster 10: Compuesto en su totalidad por hombres que conviven en un hogar donde de media existen entre uno y tres 3 miembros (como máximo 9). Los ingresos en este caso se encuentran en su mayoría por debajo de los 1600€ al mes. Adicionalmente estas personas se encuentran trabajando en su totalidad en ocupaciones No Tic y a tiempo completo. En términos de educación se puede decir que este clúster está conformado por individuos cuyos últimos estudios terminados son principalmente los de la primera etapa de la educación secundaria y similar. También del tema de estudios se puede decir que de manera general el nivel educativo de este grupo es más bien bajo. La variable *CPConInf_UsoInt* en este caso toma valores negativos casi en su totalidad por lo que se puede decir que los individuos de este grupo no tienen muchos de los conocimientos informáticos considerados en la encuesta del INE y puede ser que su uso de internet tampoco sea muy frecuente. Similarmente, la variable *CPConfSegInt* también toma en su mayoría valores negativos o positivos muy bajos lo cual puede indicar que las personas en este grupo tienen una confianza en internet sea poco o nada. El rango de edad de estas personas se encuentra entre los 40 y los 55 años, tomando el medoid el valor de 49 años

4.4 Análisis de correspondencias simples

Una vez obtenidos los diferentes grupos de individuos respecto a los datos presentes en este estudio se procedió a realizar su caracterización respecto al uso de dispositivos IoT de la encuesta. Para ello, se creó una variable dispHIoT que tomaba tres niveles:

- Con_usa: Si el individuo conoce los dispositivos de IoT y además los usa
- Con_Nousa: Si el individuo conoce los dispositivos de IoT pero no lo usa
- Ncon_Nousa: Si el individuo no conoce los dispositivos de IoT y por lo tanto, no los usa.

El desarrollo del análisis de correspondencias simples se llevará a cabo tanto en RStudio como en SAS ya que ambos programas se complementan en cuanto a las visualizaciones que muestran y a la facilidad que brindan para esta técnica.

Como se explicó en la metodología, la técnica del análisis de correspondencias simple se utilizó para visualizar las tablas de frecuencias de las variables cualitativas *Clúster* y *dispHIoT* y poder detectar relaciones entre ellas. Es por ello que a continuación se muestra la tabla de contingencia que recoge la repartición de la muestra según el número de individuos que presentan una categoría de cada una de las variables. En este caso las categorías de la variable *dispHIoT* se encuentran en las columnas de la Tabla 17,

mientras que las filas son los diferentes clústeres que se formaron en el apartado anterior.

Tabla 17: Tabla de contingencia variables Clúster y dispHIoT.

Tabla de contigencia

Clúster	con_Nousa	conusa	Nocon_Nousa	Suma
1	503	324	74	901
2	445	342	64	851
3	464	262	89	815
4	755	276	323	1354
5	693	271	212	1176
6	902	405	326	1633
7	733	178	506	1417
8	487	467	45	999
9	557	402	86	1045
10	721	250	232	1203
Suma	6260	3177	1957	11394

4.4.1 Viabilidad del análisis y número de factores a retener

Para evaluar si existe algún tipo de asociación entre las variables en cuestión se utilizó la hipótesis de independencia a través del estadístico Chi-cuadrado de Pearson. Como los valores esperados estadísticos Chi-cuadrado (representados en la Tabla 18) versus los valores reales de los datos (valores representados en la Tabla 17) difieren en gran medida, se puede afirmar que se puede realizar el análisis por correspondencia.

Tabla 18: Valores esperados estadísticos chi-cuadrado.

Clúster	con_Nousa	conusa	Nocon_Nousa
1	495.020	251.227	154.753
2	467.550	237.285	146.165
3	447.771	227.247	139.982
4	743.904	377.537	232.559
5	646.108	327.905	201.986
6	897.190	455.331	280.479
7	778.517	395.103	243.380
8	548.863	278.552	171.585
9	574.136	291.378	179.486
10	660.943	335.434	206.624

Por otro lado, se sabe que si el estadístico Chi-cuadrado calculado es mayor que el valor crítico, se puede concluir que las variables de fila y columna no son independientes entre sí. Esto implica que están significativamente asociadas y los valores que toma una de las variables influye en la distribución de la otra. En el presente estudio, las variables de fila y columna están asociadas de forma estadísticamente significativa (p valor = 0) tal y como se muestra en la Imagen 29.

Los valores propios corresponden a la cantidad de información retenida por cada eje. En el análisis de correspondencia las dimensiones se ordenan de forma decreciente y se enumeran según la cantidad de varianza explicada en la solución. La dimensión 1 explica la mayor parte de la varianza en la solución, seguida de la dimensión 2 y así sucesivamente. Los valores propios pueden utilizarse también para determinar el número de ejes que hay que conservar. El porcentaje acumulado explicado se obtiene sumando las proporciones sucesivas de variación explicada para obtener el total corrido. No existe una "regla general" para elegir el número de dimensiones que hay que mantener para la interpretación de los datos. Sin embargo, para este caso, el 100% de la variación se explica por las dos primeras dimensiones.

Pearson's Chi-squared test

X-squared = 1025.4, df = 18, p-value < 2.2e-16

Descomposición de inercia y chi-cuadrado

Valor singular	Inercia principal	Chi-cuadrado	Porcentaje	Porcentaje acumulado
0.29352	0.08615	981.63	95.73	95.73
0.06200	0.00384	43.79	4.27	100.00
	0.09000	1025.43	100.00	

Pr > Chisq < .0001 DF = 18

Imagen 29: Descomposición Inercia y Chi-cuadrado.

4.4.2 Caracterización de la sociedad española y el uso de los dispositivos IoT

Para poder obtener más infromación respecto a la relación entre este conjunto de variables se procedió a realizar un análisis más detallado considerando tanto los perfiles filas, como los perfiles columna. El estudio de estos perfiles consistió en el análisis de la información sumplementaria aportada por los coeficientes generados por las contribuciones parciales y los cosenos al cuadrado. Como se explicó en la metodología las contribuciones parciales representan la proporcion de la inercia de un factor asociadas a la modalidad correspondiente de la variable en cuestión. Por otro lado, el coseno al cuadrado mide el grado de asociación entre filas/columnas y un eje concreto. Por lo que si un elemento de la fila o columna está bien representado por dos dimensiones, la suma de los cosenos se acerca a uno.

- Perfiles fila

En la Imagen 30 se encuentran los resultados para las coordenadas, contribuciones parciales y cosenos al cuadrado de los perfiles fila. Con esta información se puede decir, por ejemplo, que la primera dimensión se construye para los clústeres 7 y 8 ya que presentan las contribuciones más altas para este eje. Por otro lado, a través de las aportaciones a los cosenos al cuadrado se puede ver que la mayoría de los clústeres se encuentran prácticamente situados en el eje ya que casi todos sus valores son

cercanos a 1. Esto era de esperar teniendo en cuenta que la dimensión uno recogía prácticamente toda la información de este análisis.

Coordenadas de la fila

	Dim1	Dim2
1	-0.2632	-0.0323
2	-0.3296	0.0353
3	-0.1661	-0.0492
4	0.2150	-0.0047
5	0.0770	-0.0760
6	0.0891	-0.0010
7	0.5266	0.0936
8	-0.4672	0.0990
9	-0.2949	0.0168
10	0.1284	-0.0934

Contribuciones parciales a la inercia para los puntos de la fila

	Dim1	Dim2
1	0.0636	0.0214
2	0.0942	0.0241
3	0.0229	0.0450
4	0.0638	0.0007
5	0.0071	0.1552
6	0.0132	0.0000
7	0.4003	0.2832
8	0.2221	0.2236
9	0.0926	0.0068
10	0.0202	0.2399

Cosenos cuadrados para los puntos de la fila

	Dim1	Dim2
1	0.9852	0.0148
2	0.9887	0.0113
3	0.9194	0.0806
4	0.9995	0.0005
5	0.5064	0.4936
6	0.9999	0.0001
7	0.9694	0.0306
8	0.9570	0.0430
9	0.9968	0.0032
10	0.6537	0.3463

Imagen 30: Coordenadas, contribuciones parciales y cosenos al cuadrado de las filas.

Por otro lado, a través del gráfico de dispersión de la Imagen 31 se obtuvo una idea de a qué polo de las dimensiones contribuyen realmente las categorías de las filas. En este caso, colores rojos y naranjas significan que la contribución a la inercia es mayor, mientras que colores verdes significa que la contribución no es mucha. Por ejemplo, se puede decir que el clúster 7 contribuye de forma importante al polo positivo de la primera dimensión, mientras que el 8 contribuyen de forma importante al polo negativo de la primera dimensión.

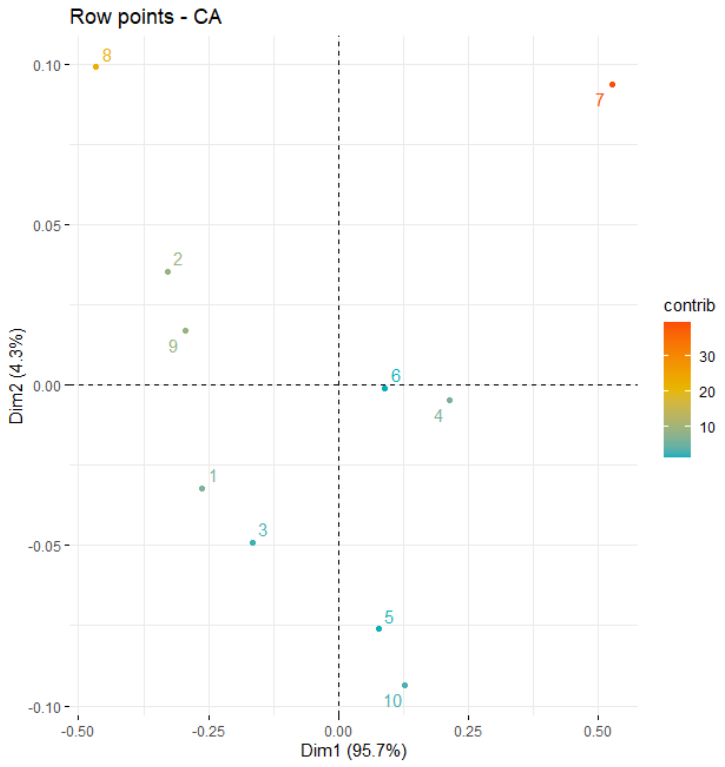


Imagen 31: Gráfico de dispersión para los perfiles fila.

- Perfiles columna

Al igual que para los perfiles filas, en la Imagen 32 se representan los resultados para las coordenadas, contribuciones parciales y cosenos al cuadrado, pero para los perfiles

columnas. En este caso se puede ver, según las contribuciones a la inercia que para el primer eje son las categorías de conusa y Nocon_Nousa las que tienen mayor aportación. En cuanto a la segunda dimensión, es la categoría con_Nousa la que contribuye mayormente a este eje. Y lo mismo pasa para el caso de las contribuciones de los cosenos al cuadrado. De la misma manera que para el caso anterior, en la Imagen 33 se representó el gráfico de dispersión para saber a qué polo de las dimensiones contribuyen realmente las categorías de las columnas. En este caso, se observa cómo la categoría Nocon_Nousa contribuye de forma importante al polo positivo de la primera dimensión, mientras que conusa contribuyen de forma importante al polo negativo de la primera dimensión.

Coordenadas de las columnas

	Dim1	Dim2
con_Nousa	0.0146	-0.0561
conusa	-0.3604	0.0644
Nocon_Nousa	0.5385	0.0748

Cosenos cuadrados para puntos de columnas

	Dim1	Dim2
con_Nousa	0.0632	0.9368
conusa	0.9691	0.0309
Nocon_Nousa	0.9811	0.0189

Contribuciones parciales a la inercia para los puntos de la columna

	Dim1	Dim2
con_Nousa	0.0014	0.4492
conusa	0.4204	0.3007
Nocon_Nousa	0.5782	0.2500

Imagen 32: Coordenadas, contribuciones parciales y cosenos al cuadrado de las columnas.

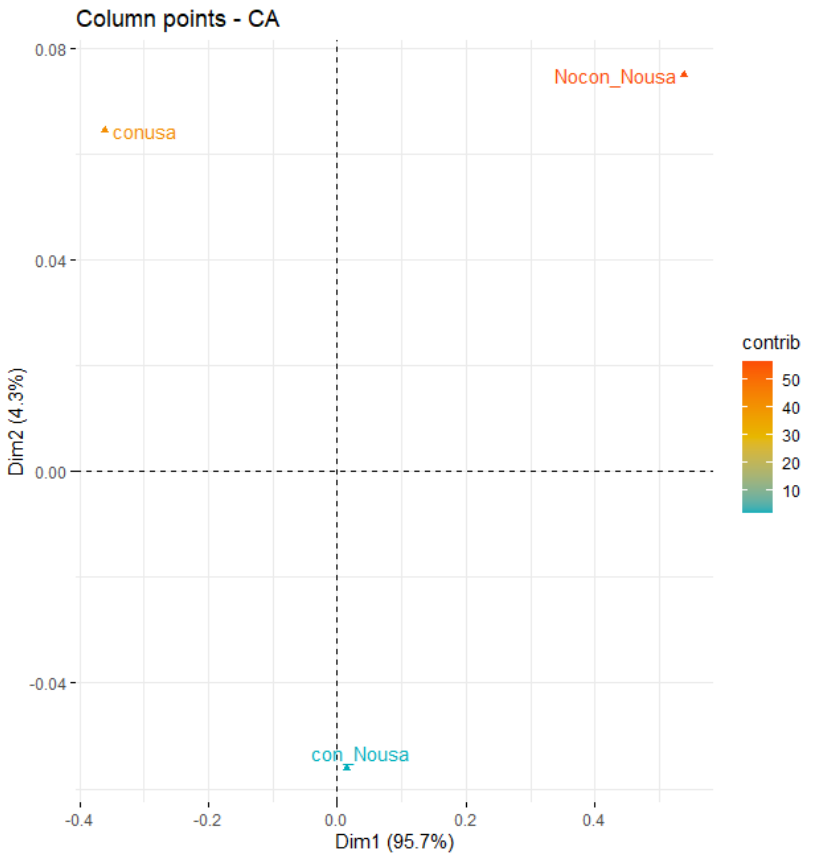


Imagen 33: Gráfico de dispersión para los perfiles columna.

Finalmente, el gráfico que se muestra a continuación en la Imagen 34 se denomina gráfico simétrico y muestra un patrón global dentro de los datos. Las filas están representadas por puntos azules y las columnas por puntos rojos. La distancia entre cualquier punto de fila o columna da una medida de su similitud (o disimilitud). Los puntos de fila con perfil similar se cierran en el mapa de factores. Lo mismo ocurre con los puntos de las columnas. Además, gracias a esa representación se puede apreciar aquellos clústeres que hacen un uso de los dispositivos IoT y por tanto, conocer los perfiles de las personas que lo integran. Por ejemplo, a partir de este gráfico se pueden sacar las conclusiones que se exponen a continuación.

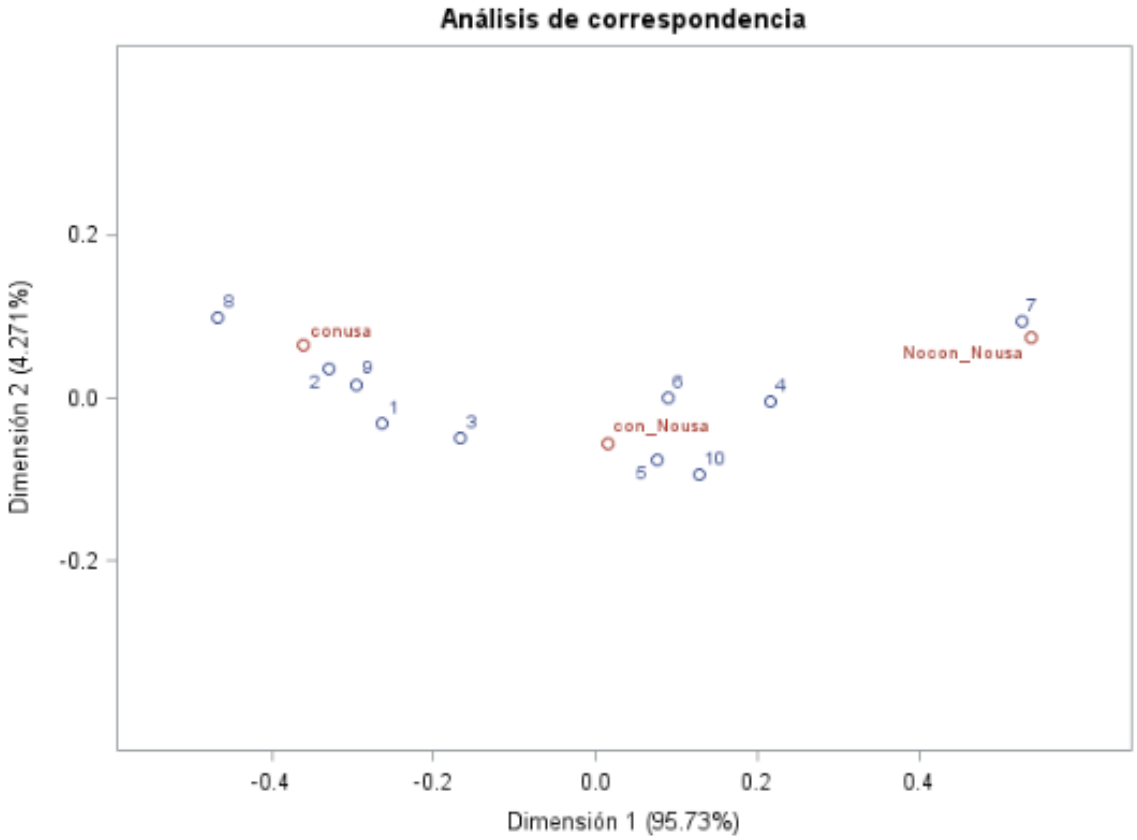


Imagen 34: Representación de perfiles filas y columnas en los ejes de las dimensiones retenidas.

- Perfil de columna con_usa:

Se puede apreciar cómo los clústeres 2 y 9 son los que mayormente se asocian con el conocimiento y uso de dispositivos IoT. De aquí se puede inferir que aquellas características similares entre los individuos que conforman estas agrupaciones serán las más relacionadas tanto con el conocimiento como con el uso de dispositivos o sistemas conectados a internet. En este caso, el nivel de educación es una de las variables que comparten ya que los individuos de ambos clústeres están mayormente formados por personas que han terminado alguna de las carreras o formaciones pertenecientes a este nivel. Adicionalmente, algo que resalta entre características

similares son, por un lado, la cantidad de actividades informáticas que saben realizar las personas de estos dos clústeres, que en cierta forma pueda estar ligada también al nivel educacional. Sin embargo, lo que no cabe duda es que el tener conocimientos informáticos está relacionado con la adopción de IoT. Lo mismo pasa con el uso frecuente de internet, el cual, en este caso, es elevado también por los valores elevados y positivos que toma la variable *CPConInf_UsoInt.* Algo a destacar también es que estas personas presentan, a modo general, un grado de preocupación bajo o nulo respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida. De ahí que su confianza en internet tenga tendencia a tomar valores repetidos de mucha o bastante. Por otro lado, se puede ver que el clúster 9 está mucho más cercano a este perfil de columna que el clúster 2. Y es que una de las diferencias relevantes entre estas dos agrupaciones es la edad que los caracteriza. El clúster 9, como se mencionó la sección anterior, es el clúster que presenta los individuos más jóvenes, conteniendo a personas con una edad que en su mayoría se encuentra entre los 18 y 28 años. Además, las edades que toman los medoid de estos clústeres se diferencian en casi el doble uno de otro.

Otros dos clústeres que se encuentran relativamente cercanos a esta categoría son los números 8 y 1. En ellos destacan, al igual que en el caso anterior, que el nivel educativo que los caracteriza sea el de educación superior, con edades alrededor de los 40 años e ingresos en el hogar de más de 1600€ mensuales por lo general. En el caso del clúster 8, este presenta el medoid cuya variable *CPConInf_UsoInt* presenta el nivel más alto positivamente, y esto se puede deber a que algunas de las personas que trabajan en este grupo, lo hacen en ocupaciones TIC. También vale la pena destacar que estos conglomerados están formados por individuos cuya confianza en internet es media-alta.

Puede llamar la atención la diferencia entre la cercanía que pueden tener los clústeres 1 y 3 con este perfil de columna cuando estos son tan similares entre sí. La única diferencia notable entre el clúster 1 y el clúster 3 es el sexo. De hecho, varios investigadores han concluido que en la adopción de nuevas tecnologías, las mujeres suelen tener una predisposición negativa en su uso prematuro [46], [47]. Consecuentemente, dado que el clúster 3, que se encuentra formado en su totalidad por mujeres, es el clúster de estos dos que menos relación tiene con el conocimiento y el uso de dispositivos IoT.

- Perfil de columna con_Nousa:

De la misma manera que para el caso anterior, se puede inferir que aquellas características similares entre los individuos que conforman las agrupaciones más cerca a esta categoría serán las más relacionadas con el conocimiento de dispositivos IoT pero no con su uso. En la Imagen 34 se puede apreciar que, para este caso, son los clústeres 5, 6 y 10. Para estas agrupaciones destaca, por ejemplo, la edad que los caracteriza. A modo general, el rango de edad de las personas que conforman estos clústeres no es muy bajo, encontrándose entre los 45 y los 66 años. A pesar de ello, y algo que destaca también es que, a pesar de que no estén formados por personas jóvenes y que por lo general tienden a encontrarse estudiando, el nivel educativo no es muy elevado. De

hecho, el clúster 10 es el que presenta el nivel educativo más bajo de todas las agrupaciones realizadas, lo que explica que se encuentre un poco más alejado. Algo a destacar es que la variable *CPConInf_UsoInt* toma valores casi siempre negativos por lo que se puede concluir estos individuos no realizan muchas de las actividades informáticas recogidas en la encuesta y puede ser que su uso de internet tampoco sea muy frecuente. Por otro lado, en estos casos, la variable *CPConfSegInt* toma valores negativos o muy bajos. Esto resulta en que la confianza que tienen estas personas que integran los clústeres 5, 6 y 10 tengan una confianza en internet nula o muy baja.

- Perfil de columna Nocon_Nousa:

Finalmente, para este perfil, es evidente en la Imagen 37 que el clúster que más relacionado está con el no tener conocimientos sobre tecnologías relacionadas con el IoT y por tanto, no usan estos dispositivos, es el clúster 7. Esto se puede explicar si entra en el detalle de las características de las personas que conforman este grupo. Para empezar, se puede tomar como referencia la edad, siendo este clúster los que concentran a las personas con mayor edad de todos los demás. De hecho, solo el 25% de las personas de este grupo tiene una edad menos que 49 años, la cual ya es elevada y además, el medoide toma el valor de 65 años. Por otra parte, al ser tan mayores estas personas, casi todas se encuentran en paro o estan jubilados/prejubilados, lo que hace que los ingresos sean bajos, ya que en su mayoría se encuentran por debajo de 900€ mensuales. Sumándole a todo esto está el hecho de que el medoide para la variable *CPConInf_UsoInt* tome el valor más negativo de todos, indicando que, los conocimientos informáticos de este grupo son más bien escazos. Esto lo avala también el hecho de que si se entra en el detalle del nivel de estudio más alto terminado se ve que en este caso es el de primera etapa de la educación secundaria y similar seguidos de personas que no tienen estudios o solo han terminado la primaria.

Teniendo en cuenta que el clúster 4 se encuentra un poco más alejado del perfil con_Nousa que otros mencionados anteriormente (clústeres 5,6 y 10), vale la pena volver a mencionar las características de este y encontrar una explicación a este caso. Dado que, por un lado, la edad de este grupo también es elevada, con solo un 25% por debajo de los 55 años y por otro, que la situación laboral sea de parados o jubilados/prejubilados se explica la cierta independencia que se mencionaba anteriormente. Sin embargo, algo que diferencia al clúster 4 del 7 es que el nivel de estudios terminados es superior, por lo que se infiere que los conocimientos de las personas serán mayores también.

5. Conclusiones

El presente trabajo se ha centrado en desarrollar un análisis detallado del estado del uso y conocimientos de los dispositivos IoT en la sociedad española. Además, en este estudio se han analizado las características de los diversos grupos de usuarios de estos dispositivos, así como el nivel de penetración actual existente.

Haciendo uso de los datos de la encuesta sobre "Equipamiento y Uso de las Tecnologías de la Información y la Comunicación en los Hogares (TIC-H)" del INE, y empleando diferentes técnicas, desde descriptivas hasta multivariantes, se ha dado cumplimiento al objetivo principal de este estudio. Es decir, se ha llevado a cabo una caracterización de los perfiles de los hogares españoles en el uso dispositivos y servicios asociados al Internet de las Cosas en durante el 2020. A diferencia de otros estudios anteriores, los datos en los cuales se basó este trabajo, además de ser más cuantiosos, estudian a la vez las dos tipologías de individuos en cuanto a adopción. Es decir, los datos tratados en esta investigación han recogido y procesado la información tanto de las personas que hacen uso de los dispositivos IoT como de los que no. Adicionalmente, al utilizar características asociadas a los propios usuarios como situación laboral, nivel de estudio, ingresos, etc., se ha proporcionado una perspectiva que va más ligada a finalidades comerciales o de segmentación de mercado.

Gracias principalmente al análisis clúster y al de correspondencias simple, se logró identificar aquellos perfiles de la población española que ya están adoptando y haciendo uso del IoT en sus hogares, y por lo tanto se pueden considerar como "early adopters". Estos usuarios se caracterizan sobre todo por tener una formación de grados superiores y universitaria ya que su nivel de estudio es elevado respecto al resto de individuos. Adicionalmente, este grupo de individuos están acostumbrados a realizar una serie de tareas informáticas que les hace estar mucho más familiarizados con la tecnología de IoT, y comprender más rápidamente el funcionamiento de la misma. Como era de esperar también, son los jóvenes los que más hacen uso de los dispositivos IoT, principalmente porque se pueden considerar nativos digitales y más propensos a hacer uso de forma intuitiva de los mismos. Los niveles de ingresos altos en los hogares para la adopción del IoT se estableció como categoría necesaria pero no decisiva ni excluyente. Vale la pena mencionar que estos individuos hacen un uso muy frecuente de internet y que, además, su grado de confianza en este es muy elevado.

También se han caracterizado a las personas que, aunque conocen los dispositivos asociados al Internet de las Cosas, no han hecho uso de estos con finas privados, considerándose como rezagados en la adopción de esta tecnología. Estas personas, aunque sí tenían algún nivel de formación, por lo general no era muy alto lo que, adicionado a rangos de edades más bien elevados para la muestra, resultan en que los conocimientos informáticos no sean diversos. De ahí que no vean una necesidad real en el uso de dispositivos IoT. Algo que destacó en estos grupos de individuos fue que el grado de preocupación respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida es mucho. Esto puede deberse a que, al

no tener muchos conocimientos informáticos, pues no conocen o entienden el detalle de las políticas de privacidad del tratamiento de los datos. Además, por esa misma razón, pueden que tampoco sepan saben utilizar mecanismos antirrastreo o bloqueo en internet.

Por último, se han detectado aquellas personas que no conocen las tecnologías IoT y por lo tanto no hacen uso de los mismos. Estas personas se caracterizaron por tener una edad muy elevada y casi ninguna formación. Adicionalmente, muchos estaban en paro o se encontraban jubilados/prejubilados por lo que su situación económica no les permite ampliamente hacer un destino de presupuestos para estas tecnologías.

A modo general, en este estudio se ha verificado que la penetración actual del IoT en España no es en realidad elevada aún pues actualmente existen más de 22.627.279 que no utilizan dispositivos IoT con fines privados frente a unos 10.216.795 que sí lo hacen. Sin embargo, este perfilado de la población, como se ha comentado anteriormente, resulta de utilidad para los profesionales en la industria de IoT que gestionan las campañas publicitarias de estos dispositivos o se dedican a su comercialización. Al entrar en el detalle de tipologías de usuarios y de cómo agruparlos, los resultados de este trabajo han contribuido a la mejor adaptación del mercado tanto objetivo como potencial de las empresas del sector.

Por todo lo expresado anteriormente, se puede concluir que el mercado de "*consumer*" IoT dentro de España abarca diferentes niveles tanto en términos de dirección como en velocidad en dependencia de las características de cada usuario. Esto resulta en que los principales usuarios de dispositivos IoT sean por lo general los perfiles minoritarios o más bien conocidos como "early adopters", mientras que el resto del mercado potencial se encuentra o en total desconocimiento o esperando a una "mejor" definición de la oferta. Consecuentemente, se han de concentrar los esfuerzos en atacar aquellas barreras principales que impiden que la cuota de mercado aumente, las cuales también han sido recogidas en este trabajo.

5.1 Limitaciones del trabajo y Futuras líneas de investigación

Este trabajo estuvo limitado por los tipos de preguntas relacionadas con IoT incluidas en la encuesta del INE. Quizás hacer una ampliación del formato y contenido de preguntas realizadas, como, por ejemplo, incluir características de los dispositivos, pueden resultar en una mejora de los resultados del estudio. De hecho, valdría la pena incluir preguntas contenidas y analizadas por la literatura referente a este tema y que son consideradas como principales influyentes en la teoría de aceptación de tecnología. Estos pueden ser la actitud frente al uso de estos dispositivos o la facilidad percibida al usarlos.

Adicionalmente, las principales conclusiones de esta investigación se basan solo en los datos de España. Un estudio futuro debería intentar reunir datos étnicos y geográficos diversos para garantizar la generalización de los resultados.

En cuanto a metodología, el algoritmo utilizado en el presente trabajo presentó un alto grado de exigencia computacional. Futuros trabajos deberían profundizar en la validación y comparación de otros tipos de algoritmos de clustering que sean compatibles con variables tanto categóricas como numéricas y que, a la vez, no sean tan densos en términos de procesamiento como lo es el PAM.

6. Bibliografía

[1] V. Ricquebourg, D. Menga, D. Durand, B. Marhic, L. Delahoche, y C. Logé, «The Smart Home Concept : our immediate future», ene. 2007, pp. 23-28. doi: 10.1109/ICELIE.2006.347206.

[2] N. Balta-Ozkan, R. Davidson, M. Bicket, y L. Whitmarsh, «Social barriers to the adoption of smart homes», *Energy Policy*, vol. 63, pp. 363-374, dic. 2013, doi: 10.1016/j.enpol.2013.08.043.

[3] A. Adami, T. L. Hayes, y M. Pavel, «Unobtrusive Monitoring of Sleep Patterns», oct. 2003, vol. 2, pp. 1360-1363 Vol.2. doi: 10.1109/IEMBS.2003.1279555.

[4] G. Misra, V. Kumar, A. Agarwal, y K. Agarwal, «Internet of Things (IoT) – A Technological Analysis and Survey on Vision, Concepts, Challenges, Innovation Directions, Technologies, and Applications (An Upcoming or Future Generation Computer Communication System Technology)», *Am. J. Electr. Electron. Eng.*, vol. 4, pp. 23-32, feb. 2016, doi: 10.12691/ajeee-4-1-4.

[5] H. Yang, W. Lee, y H. Lee, «IoT Smart Home Adoption: The Importance of Proper Level Automation», *J. Sens.*, vol. 2018, pp. 1-11, may 2018, doi: 10.1155/2018/6464036.

[6] J. Shin, Y. Park, y D. Lee, «Who will be smart home users? An analysis of adoption and diffusion of smart homes», *Technol. Forecast. Soc. Change*, vol. 134, jun. 2018, doi: 10.1016/j.techfore.2018.06.029.

[7] «Consumer IoT Market Size by End-Use Application (In-Car Infotainment, Traffic Management, Automotive, Home Automation, Healthcare, Consumer Electronics, Wearable Devices), By Offerings (Service, Solution, Network Infrastructure, Node Component), By Phase (By Region (North America, Europe, Asia-Pacific, Rest of the World), Market Analysis Report, Forecast 2021-2026», *Marketresearch*, jun. 07, 2020. https://www.marketresearchengine.com/consumer-iot-market

[8] Telefónica, «Things Matter 2019: La experiencia del usuario de Internet de las Cosas en España», Madrid, 2019.

[9] X. Page, P. Bahirat, M. I. Safi, B. P. Knijnenburg, y P. Wisniewski, «The Internet of What?: Understanding Differences in Perceptions and Adoption for the Internet of Things», *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, n.º 4, pp. 1-22, dic. 2018, doi: 10.1145/3287061.

[10] L. Vadillo, M. L. Martín-Ruiz, I. Pau, R. Conde, y M. Á. Valero, «A Smart Telecare System at Digital Home: Perceived Usefulness, Satisfaction, and Expectations for Healthcare Professionals», *J. Sens.*, vol. 2017, pp. 1-12, 2017, doi: 10.1155/2017/8972350.

[11] W. Ben Arfi, I. Ben Nasr, T. Khvatova, y Y. Ben Zaied, «Understanding acceptance of eHealthcare by IoT natives and IoT immigrants: An integrated model of UTAUT, perceived risk, and financial cost», *Technol. Forecast. Soc. Change*, vol. 163, p. 120437, feb. 2021, doi: 10.1016/j.techfore.2020.120437.

[12] M. Q. Aldossari y A. Sidorova, «Consumer Acceptance of Internet of Things (IoT): Smart Home Context», *J. Comput. Inf. Syst.*, vol. 60, n.º 6, pp. 507-517, nov. 2020, doi: 10.1080/08874417.2018.1543000.

[13] A. Hong, C. Nam, y S. Kim, «What will be the possible barriers to consumers' adoption of smart home services?», 2020.

[14] M. Tsourela y D.-M. Nerantzaki, «An Internet of Things (IoT) Acceptance Model. Assessing Consumer's Behavior toward IoT Products and Applications», *Future Internet*, vol. 12, n.º 11, Art. n.º 11, nov. 2020, doi: 10.3390/fi12110191.

[15] Y. Kim, Y. Park, y J. Choi, «A study on the adoption of IoT smart home service: using Value-based Adoption Model», *Total Qual. Manag. Bus. Excell.*, vol. 28, pp. 1-17, abr. 2017, doi: 10.1080/14783363.2017.1310708.

[16] A. Alhogail, «Improving IoT Technology Adoption through Improving Consumer Trust», *Technologies*, vol. 6, p. 64, jul. 2018, doi: 10.3390/technologies6030064.
[17] C.-L. Hsu y J. Lin, «An empirical examination of consumer adoption of Internet of Things services: Network externalities and concern for information privacy perspectives», *Comput. Hum. Behav.*, vol. 62, pp. 516-527, sep. 2016, doi: 10.1016/j.chb.2016.04.023.
[18] INE, «Encuesta sobre Equipamiento y Uso de Tecnologías de Información y Comunicación en los Hogares (TIC-H)». 2020.
[19] «INE. Instituto Nacional de Estadística», *INE*. https://www.ine.es/
[20] J. Choi, S. Kim, J. Chen, y S. Dannels, «A Comparison of Maximum Likelihood and Bayesian Estimation for Polychoric Correlation Using Monte Carlo Simulation», *J. Educ. Behav. Stat. - J EDUC BEHAV STAT*, vol. 36, pp. 523-549, jul. 2011, doi: 10.3102/1076998610381398.
[21] A. Freiberg Hoffmann, J. B. Stover, G. de la Iglesia, y M. Fernández Liporace, «CORRELACIONES POLICÓRICAS Y TETRACÓRICAS EN ESTUDIOS FACTORIALES EXPLORATORIOS Y CONFIRMATORIOS», *Cienc. Psicológicas*, vol. 7, n.º 2, pp. 151-164, nov. 2013.
[22] S.-J. Cho, F. Li, y D. Bandalos, «Accuracy of the Parallel Analysis Procedure With Polychoric Correlations», *Educ. Psychol. Meas.*, vol. 69, n.º 5, pp. 748-759, oct. 2009, doi: 10.1177/0013164409332229.
[23] M. Richaud, «Desarrollos del análisis factorial para el estudio de ítem dicotomicos y ordinales», *Interdisciplinaria*, vol. 22, dic. 2005.
[24] D. R. Divgi, «Calculation of the tetrachoric correlation coefficient», *Psychometrika*, vol. 44, n.º 2, pp. 169-172, jun. 1979, doi: 10.1007/BF02293968.
[25] S. Wold, K. Esbensen, y P. Geladi, «Principal component analysis», *Chemom. Intell. Lab. Syst.*, vol. 2, n.º 1, pp. 37-52, ago. 1987, doi: 10.1016/0169-7439(87)80084-9.
[26] K. Y. Yeung y W. L. Ruzzo, «Principal component analysis for clustering gene expression data», *Bioinformatics*, vol. 17, n.º 9, pp. 763-774, sep. 2001, doi: 10.1093/bioinformatics/17.9.763.
[27] C. Ding y X. He, «*K*-means clustering via principal component analysis», en *Proceedings of the twenty-first international conference on Machine learning*, New York, NY, USA, jul. 2004, p. 29. doi: 10.1145/1015330.1015408.
[28] H. Kargupta, W. Huang, K. Sivakumar, y E. Johnson, «Distributed Clustering Using Collective Principal Component Analysis», *Knowl. Inf. Syst.*, vol. 3, n.º 4, pp. 422-448, nov. 2001, doi: 10.1007/PL00011677.
[29] J. Gower, «A General Coefficient of Similarity and Some of Its Properties», *undefined*, 1971, Accedido: may 22, 2021. [En línea]. Disponible en: /paper/A-General-Coefficient-of-Similarity-and-Some-of-Its-Gower/668bf9f3932d29f316d68276958acf62256aaac3
[30] D. Peña, *Análisis multivariante de datos*. McGraw-Hill Interamericana de España S.L., 2002.
[31] J. M. Alonso Revenga, «Análisis Clúster». Máster en Minería de Datos e Inteligencia de Negocio, Universidad Complutense de Madrid.
[32] S. Harikumar y S. Pv, «K-Medoid Clustering for Heterogeneous DataSets», *Procedia Comput. Sci.*, vol. 70, pp. 226-237, ene. 2015, doi: 10.1016/j.procs.2015.10.077.
[33] R. Joshi, A. Patidar, y S. Mishra, «Scaling k-medoid algorithm for clustering large categorical dataset and its performance analysis», en *2011 3rd International Conference on Electronics Computer Technology*, abr. 2011, vol. 2, pp. 117-121. doi: 10.1109/ICECTECH.2011.5941667.
[34] W. Budiaji y F. Leisch, «Simple K-Medoids Partitioning Algorithm for Mixed Variable Data», *Algorithms*, vol. 12, n.º 9, Art. n.º 9, sep. 2019, doi: 10.3390/a12090177.
[35] E. C. de Assis y R. M. C. R. de Souza, «A K-medoids clustering algorithm for mixed feature-type symbolic data», en *2011 IEEE International Conference on Systems, Man, and Cybernetics*, oct. 2011, pp. 527-531. doi: 10.1109/ICSMC.2011.6083737.

[36] C. Hennig y T. F. Liao, «How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification», *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 62, n.º 3, pp. 309-369, 2013, doi: https://doi.org/10.1111/j.1467-9876.2012.01066.x.
[37] L. Hunt y M. Jorgensen, «Clustering mixed data», *WIREs Data Min. Knowl. Discov.*, vol. 1, n.º 4, pp. 352-361, 2011, doi: https://doi.org/10.1002/widm.33.
[38] L. Kaufman y P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York : Wiley, ©1990. Accedido: may 23, 2021. [En línea]. Disponible en: https://www.wiley.com/en-ar/Finding+Groups+in+Data%3A+An+Introduction+to+Cluster+Analysis-p-9780471735786
[39] A. P. Reynolds, G. Richards, y V. J. Rayward-Smith, «The Application of K-Medoids and PAM to the Clustering of Rules», en *Intelligent Data Engineering and Automated Learning – IDEAL 2004*, Berlin, Heidelberg, 2004, pp. 173-178. doi: 10.1007/978-3-540-28651-6_25.
[40] D. Lei, Q. Zhu, J. Chen, H. Lin, y P. Yang, «Automatic PAM Clustering Algorithm for Outlier Detection», *J. Softw.*, vol. 7, abr. 2012, doi: 10.4304/jsw.7.5.1045-1051.
[41] L. F. Ibrahim y M. H. A. Harbi, «Using Clustering Technique M-PAM in Mobile Network Planning», Heraklion, Greece, 2008, p. 6.
[42] N. N. Mohammed y A. M. Abdulazeez, «Evaluation of Partitioning Around Medoids Algorithm with Various Distances on Microarray Data», en *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, jun. 2017, pp. 1011-1016. doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.155.
[43] H. Park, J. Lee, y C. Jun, *Abstract A K-means-like Algorithm for K-medoids Clustering and Its Performance*.
[44] H. B. Zhou y J. T. Gao, «Automatic Method for Determining Cluster Number Based on Silhouette Coefficient», *Adv. Mater. Res.*, vol. 951, pp. 227-230, 2014, doi: 10.4028/www.scientific.net/AMR.951.227.
[45] J. M. Alonso Revenga, «Análisis Factorial de Correspondencias Simples». Máster en Minería de Datos e Inteligencia de Negocio, Universidad Complutense de Madrid.
[46] T. Kotzé, O. Anderson, y K. Summerfield, «Technophobia: Gender differences in the adoption of high-technology consumer products», *South Afr. J. Bus. Manag.*, vol. 47, pp. 21-28, mar. 2016, doi: 10.4102/sajbm.v47i1.49.
[47] G. G. Gebre, H. Isoda, D. B. Rahut, Y. Amekawa, y H. Nomura, «Gender differences in the adoption of agricultural technology: The case of improved maize varieties in southern Ethiopia», *Womens Stud. Int. Forum*, vol. 76, 2019, doi: 10.1016/j.wsif.2019.102264.

7. Anexos

Anexo A: Encuesta INE.

Encuesta sobre Equipamiento y Uso de Tecnologías de Información y Comunicación en los Hogares (TIC-H). 2020.

11. ¿Cuándo fue la última vez que usó Internet?

a) En los últimos 3 meses 1

b) Hace más de 3 meses y menos de 1 año ..... 2 → pasar al Bloque VI

c) Hace más de 1 año..... 3 → pasar al Bloque XII

13. ¿Y usa Internet varias veces al día?

SÍ 1

NO..... 6

36. ¿Cuáles de las siguientes tareas relacionadas con los móviles y los ordenadores ha realizado en los últimos 12 meses?

	SI	NO
a) Transferir ficheros entre el ordenador y otros dispositivos (p.ej., cámaras digitales, teléfonos móviles, mp3 ó mp4).....	1	6
b) Instalar software o aplicaciones (apps)	1	6
c) Cambiar la configuración de cualquier software, incluidos el sistema operativo y los programas de seguridad	1	6

37. ¿Cuáles de las siguientes tareas relacionadas con la informática ha realizado en los últimos 12 meses?

	SI	NO
a) Copiar o mover ficheros o carpetas.....	1	6
b) Usar un procesador de texto	1	6
c) Crear presentaciones o documentos que integren texto, imágenes, tablas o gráficos	1	6
d) Usar hojas de cálculo	1	6

(En caso de marcar SI en la opción d) se realiza la pregunta d1)

d1) Usar sus funciones avanzadas para organizar y analizar

datos, como ordenar, filtrar, usar fórmulas, construir gráficos 1 6

e) Usar software para editar fotos, video o archivos de audio 1 6

f) Programar en un lenguaje de programación..... 1 6

40. Y, ¿ha cambiado alguna vez la configuración de su navegador de Internet para prevenir o limitar la cantidad de cookies?

SÍ..... 1

NO..... 6

41. Indique, por favor, su grado de preocupación respecto a que sus actividades online puedan estar siendo registradas para ofrecerle publicidad a medida

Muy preocupado..... 1

Algo preocupado..... 2

Nada preocupado..... 3

47. Y, en general, por favor, indique su grado de confianza en Internet

Poco o nada..... 1

Bastante..... 2

Mucho..... 3

48. ¿Ha utilizado alguno de los siguientes dispositivos o sistemas conectados a Internet con fines privados?

SÍ NO

a) Sistemas para la administración de energía en su hogar,

termostatos,luces, enchufes u otras soluciones …… …….. 1 6

b) Sistemas de seguridad para el hogar, como

alarmas de detector de humo,

cámara de seguridad, cerraduras de las puertas 1 6

c) Electrodomésticos conectados, como

aspiradoras de robot, frigoríficos,

hornos, máquinas de café…….. 1 6

d) Un asistente virtual en forma altavoz inteligente o de app,

como Alexa,Siri, Google Home, Echo, Computer, Google Assistant, Cortana, Bixby …….. 1 6

49. ¿Conoce la existencia de tales dispositivos o sistemas conectados a Internet?

SÍ..... 1 pasar a pregunta 50

NO..... 6 pasar a pregunta 51

50. ¿Por cuáles de los siguientes motivos no ha utilizado dispositivos o sistemas conectados a Internet con fines privados?

SÍ NO

a) No tuve necesidad de usar esos dispositivos o sistemas conectados……. 1 6

b) Costes demasiado altos…… 1 6

c) Falta de compatibilidad con otros dispositivos o sistema…… 1 6

d) Falta de habilidades para utilizar esos dispositivos o sistemas…… 1 6

e) Preocupaciones sobre la privacidad y protección de datos personales

generados por esos dispositivos o sistemas…… 1 6

f) Preocupaciones sobre la seguridad (por ej. que sea pirateado) …… 1 6

g) Preocupaciones sobre la seguridad o la salud (p. ej. pueda provocar

un accidente o un problema de salud) …….. 1 6

h) Otras razones…….. 1 6

XII.- CARACTERÍSTICAS SOCIOECONÓMICAS DE LA PERSONA SELECCIONADA

Soltero/a..... **1**
Casado/a..... **2**
Viudo/a..... **3**
Separado/a..... **4**
Divorciado/a..... **5**

57. Con independencia de su situación legal, ¿convive actualmente en pareja?

SÍ..... **1**
NO..... **6**

58. ¿Cuáles son sus estudios terminados de más alto nivel?

Entrevistador: *Anote en el literal del estudio terminado y señale la opción que corresponda. El informante debe especificar en su respuesta lo suficiente para que se pueda codificar correctamente.*

Analfabetos y estudios primarios incompletos..... **0**
Educación Primaria..... **1**
Primera etapa de la Educación Secundaria y similar (Bachiller elemental, EGB, ESO…) **2**

Segunda etapa de la Educación Secundaria y similar (Bachiller Superior, BUP, Bachillerato, FPI, FP de Grado Medio...) …… **3**
Educación postsecundaria no superior (CdPN3…… **4**
Formación Profesional de Grado Superior (FPII) y títulos propios de universidades de duración igual o superior a 2 años…… **5**
Grados universitarios de 240 créditos ECTS (Bolonia), diplomados universitarios, títulos propios universitarios de experto o especialista y similares……..... **6**
Grados universitarios de más de 240 créditos ECTS (Bolonia), licenciados, másteres y especialidades en Ciencias de la Salud por el sistema de residencia y similares..... **7**
Título de Doctorado..... **8**
No se puede codificar..... **9**

59. ¿En cuál de las siguientes situaciones en relación con la actividad se encontraba la semana pasada? Si se encontraba en varias, indique solo la que considere principal.

Trabajando por cuenta ajena con contrato indefinido (o relación laboral permanente)..... **1**
Trabajando por cuenta ajena con contrato temporal.. **2**
Trabajando por cuenta propia (se incluye ayuda familiar)..... **3**
Parado..... **4**
Estudiante..... **5**
Jubilado ó prejubilado..... **6**
Incapacitado permanente..... **7** → *pasar a pregunta 64*
Realizando tareas de voluntariado social..... **8**
Labores del hogar..... **9**
Otra situación **0**

60. ¿Qué tipo de jornada tiene en su trabajo principal?

A tiempo completo..... **1**
A tiempo parcial..... **2**

62. Respecto al trabajo desarrollado la semana pasada, ¿cuál era su ocupación principal?

63 (b) ***Entrevistador:*** **La ocupación del trabajador es:**

Trabajador TIC..... **1**
Otros trabajadores..... **6**

64. Por último, ¿me podría indicar, aproximadamente, el intervalo en el que se encuentran los ingresos mensuales netos de su hogar (es decir, después de las retenciones a cuenta por impuestos, cotizaciones sociales y otros pagos asimilados)? Incluya, por favor, todas las fuentes de ingreso (en el caso de existir más de una), considerando (para los ingresos del trabajo por cuenta ajena) la parte proporcional de las pagas extraordinarias y otros ingresos extraordinarios percibidos regularmente.

Entrevistador: *anote el intervalo declarado. Se deberá contabilizar* ***la suma*** *de los ingresos*

Menos de 900 euros..... **1**
De 900 a menos de 1.600 euros..... **2**
De 1.600 a menos de 2.500 euros..... **3**
De 2.500 a menos de 3.000 euros..... **4**
3.000 ó más euros..... **5**
NS/NR..... **6**

Anexo B: Códigos INE.

56. ¿Cuál es su estado civil legal?

NIVEL DE ESTUDIOS TERMINADOS
0 Analfabetos y estudios primarios incompletos
1 Educación Primaria
2 Primera etapa de la Educación Secundaria y similar
3 Segunda etapa de la Educación Secundaria y similar (Bachillerato y FP de Grado Medio)
4 Educación postsecundaria no superior
5 Formación Profesional de Grado Superior y títulos propios de universidades de duración igual o superior a 2 años.
6 Grados universitarios de 240 créditos ECTS (Bolonia), diplomados universitarios, títulos propios universitarios de experto o especialista y similares
7 Grados universitarios de más de 240 créditos ECTS (Bolonia), licenciados, másteres y especialidades en Ciencias de la Salud por el sistema de residencia y similares
8 Título de Doctorado
9 No se puede codificar

CODIFICACIÓN DE LA OCUPACIÓN
Esta variable se va a clasificar bajo una doble perspectiva: ocupaciones manuales/no manuales y ocupaciones TIC/ No TIC. Se va a utilizar para ello la Clasificación Nacional de Ocupaciones 2011 (CNO-11) versión nacional de la clasificación ISCO-08 (sobre la que, según Eurostat, debe establecerse la codificación).
2) Ocupaciones TIC/ Ocupaciones no TIC
En este caso, y siguiendo de igual forma las indicaciones de Eurostat, se procede a una definición por exclusión: se precisan aquellas ocupaciones que configuran los denominados "Trabajadores TIC" considerando "Otros Trabajadores" aquellos no incluidos en lo anterior. Dados los posibles matices que algunos de los textos que se enumeran a continuación puedan tener, se incluyen los correspondientes códigos CNO-11 a fin de que, en su caso, puedan ser consultados si existieran dudas.

Anexo C: Material complementario resultante de la depuración de datos.

Nodo DMDB


Variables de intervalo

Inputs ordenados	Rol de los datos	Variable	Mediana	Ausente	No ausente	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis	Rol	Etiqueta	Valor absoluto del c.v.	Coeficiente de variación	Signo
1	TRAIN	tot_may74_	0	0	11394	0	3	0.062313	0.267585	4.631219	23.11983	INPUT	tot_may74_	4.294171	4.294171	+
2	TRAIN	total10_15_	0	0	11394	0	4	0.185975	0.465704	2.603288	6.615375	INPUT	total10_15_	2.504122	2.504122	+
3	TRAIN	tot_16_24_...	0	0	11394	0	4	0.191417	0.478221	2.65282	7.155234	INPUT	tot_16_24_...	2.498324	2.498324	+
4	TRAIN	tot_men16_	0	0	11394	0	6	0.415043	0.748459	1.775041	2.67337	INPUT	tot_men16_	1.80333	1.80333	+
5	TRAIN	tot_16_64_...	1	0	11394	0	6	0.75531	0.900657	1.188751	1.29265	INPUT	tot_16_64_...	1.192434	1.192434	+
6	TRAIN	tot_16_64_t_	1	0	11394	0	5	1.060646	0.860144	0.413975	-0.20006	INPUT	tot_16_64_t_	0.810963	0.810963	+
7	TRAIN	tot_mh_	2	0	11394	1	13	2.575215	1.207369	0.602831	0.740091	INPUT	tot_mh_	0.468842	0.468842	+
8	TRAIN	edad_	49	0	11394	16	74	48.11102	14.8808	-0.32679	-0.6847	INPUT	edad_	0.309301	0.309301	+

Rol de los datos	Nombre de la variable	Rol	Número de niveles	Ausente	Moda	Porcentaje moda	Moda2	Porcentaje Moda2
TRAIN	activ_	INPUT	12	0	NA	47.73	Administración pública, defensa,	16.80
TRAIN	confint_	INPUT	3	0	Bastante	53.12	Poco o nada	43.06
TRAIN	conv_	INPUT	2	0	Si	52.28	No	47.72
TRAIN	convpar_	INPUT	3	0	No conviviendo en pareja	47.72	Conviviendo con su cónyuge	45.35
TRAIN	cpro_	INPUT	19	0	Madrid	11.17	Andalucía	8.95
TRAIN	disp48_1_	INPUT	2	0	No	92.14	Si	7.86
TRAIN	disp48_2_	INPUT	2	0	No	91.54	Si	8.46
TRAIN	disp48_3_	INPUT	2	0	No	90.81	Si	9.19
TRAIN	disp48_4_	INPUT	2	0	No	83.75	Si	16.25
TRAIN	disp51_1_	INPUT	2	0	Si	66.54	No	33.46
TRAIN	disp51_2_	INPUT	2	0	No	73.70	Si	26.30
TRAIN	disp51_3_	INPUT	2	0	No	84.83	Si	15.17
TRAIN	disp52_1_	INPUT	2	0	No	76.99	Si	23.01
TRAIN	disp52_2_	INPUT	2	0	No	93.06	Si	6.94
TRAIN	disp52_3_	INPUT	2	0	No	97.31	Si	2.69
TRAIN	disp52_4_	INPUT	2	0	No	93.43	Si	6.57
TRAIN	estc_	INPUT	5	0	Casado/a	50.49	Soltero/a	35.13
TRAIN	frec_int_	INPUT	3	0	Diariamente(5 días por semana)	86.40	Todas las semanas pero no diaria	10.70
TRAIN	habitat_estrato_	INPUT	7	0	Municipios capitales de provinci	24.20	Municipios con menos de 10.000 h	19.57
TRAIN	habitat_eur_	INPUT	3	0	Densely populated area	55.88	Intermediate density area	29.20
TRAIN	indol1_	INPUT	2	0	Si	80.28	No	19.72
TRAIN	indol2_	INPUT	2	0	No	59.18	Si	40.82
TRAIN	indol3_	INPUT	2	0	No	89.08	Si	10.92
TRAIN	indol4_	INPUT	2	0	No	77.50	Si	22.50
TRAIN	indol5_	INPUT	2	0	Si	63.15	No	36.85
TRAIN	indol6_	INPUT	2	0	No	66.65	Si	33.35
TRAIN	indol7_	INPUT	2	0	No	78.20	Si	21.80
TRAIN	ing_hog_	INPUT	6	0	De 900 a menos de 1.600 euros	28.87	De 1.600 a menos de 2.500 euros	21.43
TRAIN	nacionalidad_	INPUT	4	0	Española	94.24	Extranjera	3.95
TRAIN	nivelest_	INPUT	10	0	Segunda etapa de la Educación Se	24.21	Primera etapa de la Educación Se	22.70
TRAIN	nodisp1_	INPUT	3	0	NA	45.06	Si	41.77
TRAIN	nodisp2_	INPUT	3	0	NA	45.06	No	35.48
TRAIN	nodisp3_	INPUT	3	0	NA	45.06	No	43.35
TRAIN	nodisp4_	INPUT	3	0	NA	45.06	No	40.03
TRAIN	nodisp5_	INPUT	3	0	NA	45.06	No	35.08
TRAIN	nodisp6_	INPUT	3	0	NA	45.06	No	34.99
TRAIN	nodisp7_	INPUT	3	0	NA	45.06	No	43.69
TRAIN	nodisp8_	INPUT	3	0	NA	45.06	No	43.89
TRAIN	ocupacion1_	INPUT	3	0	NA	47.73	No manual	38.70
TRAIN	ocupacion2_	INPUT	3	0	No TIC	50.35	NA	47.73
TRAIN	pnacto_	INPUT	3	0	España	92.34	De un país fuera UE	5.78
TRAIN	preopub_	INPUT	3	0	Algo preocupado	49.87	Muy preocupado	25.33
TRAIN	prevcook_	INPUT	2	0	No	71.58	Si	28.42
TRAIN	sexo_	INPUT	2	0	Mujer	52.79	Hombre	47.21
TRAIN	sit_lab_	INPUT	10	0	Trabajando por cuenta ajena con	37.11	Jubilado ó prejubilado	15.58
TRAIN	softras_	INPUT	2	0	No	86.05	Si	13.95
TRAIN	tareainf1_	INPUT	2	0	Si	61.50	No	38.50
TRAIN	tareainf2_	INPUT	2	0	Si	54.78	No	45.22
TRAIN	tareainf3_	INPUT	2	0	No	58.57	Si	41.43
TRAIN	tareainf4_	INPUT	2	0	No	60.49	Si	39.51
TRAIN	tareainf4_1_	INPUT	3	0	NA	60.49	Si	25.86
TRAIN	tareainf5_	INPUT	2	0	No	59.61	Si	40.39
TRAIN	tareainf6_	INPUT	2	0	No	93.85	Si	6.15
TRAIN	tip_h_	INPUT	5	0	Pareja con hijos que convivan en	37.93	Hogar unipersonal	21.20
TRAIN	tip_jor_	INPUT	3	0	NA	47.73	A tiempo completo	45.35
TRAIN	tmor1_	INPUT	2	0	Si	54.97	No	45.03
TRAIN	tmor2_	INPUT	2	0	Si	58.88	No	41.12
TRAIN	tmor3_	INPUT	2	0	No	74.79	Si	25.21
TRAIN	ult_int_	INPUT	1	0	En los últimos 3 meses	100.0		0.00
TRAIN	uso_int_	INPUT	1	0	Si	100.0		0.00
TRAIN	usomov_	INPUT	2	0	Si	99.97	No	0.03
TRAIN	vintd_	INPUT	3	0	Si	83.66	NA	13.60
TRAIN	viv_inter_	INPUT	2	0	Si	99.76	No	0.24

Variables de clase

Rol de los datos	Nombre de la variable	Nivel	CODE	Número de ocurrencias	Tipo	Porcentaje	Índice de nivel
TRAIN	activ_	NA	2	5438	C	47.72687	9
TRAIN	activ_	Administración pública, defe...	0	1914	C	16.79831	3
TRAIN	activ_	Comercio al por mayor y al p...	6	1187	C	10.41776	5
TRAIN	activ_	Industrias extractivas, manuf...	1	794	C	6.96858	7
TRAIN	activ_	Otros servicios	7	522	C	4.581359	11
TRAIN	activ_	Servicios empresariales	5	401	C	3.519396	12
TRAIN	activ_	Construcción	8	356	C	3.124451	6
TRAIN	activ_	Información y comunicaciones	3	262	C	2.299456	8
TRAIN	activ_	Agricultura, silvicultura y pesca	9	260	C	2.281903	4
TRAIN	activ_	Actividades financieras y de ...	4	192	C	1.685097	1
TRAIN	activ_	Actividades inmobiliarias	10	37	C	0.324732	2
TRAIN	activ_	No se puede codificar	11	31	C	0.272073	10
TRAIN	confint_	Bastante	0	6052	C	53.11567	1
TRAIN	confint_	Poco o nada	1	4906	C	43.05775	3
TRAIN	confint_	Mucho	2	436	C	3.826575	2
TRAIN	conv_	Si	1	5957	C	52.2819	2
TRAIN	conv_	No	0	5437	C	47.7181	1
TRAIN	convpar_	No conviviendo en pareja	0	5437	C	47.7181	3
TRAIN	convpar_	Conviviendo con su cónyuge	1	5167	C	45.34843	1
TRAIN	convpar_	Conviviendo con una pareja ...	2	790	C	6.933474	2
TRAIN	cpro_	Madrid	11	1273	C	11.17255	15
TRAIN	cpro_	Andalucía	3	1020	C	8.95208	1
TRAIN	cpro_	Cataluña	7	991	C	8.69756	8
TRAIN	cpro_	Comunitat Valenciana	2	756	C	6.635071	10
TRAIN	cpro_	País Vasco	0	718	C	6.301562	19
TRAIN	cpro_	Galicia	8	693	C	6.082148	12
TRAIN	cpro_	Castilla y León	4	644	C	5.652098	6
TRAIN	cpro_	Aragón	9	609	C	5.344918	2
TRAIN	cpro_	Castilla-La Mancha	1	604	C	5.301036	7
TRAIN	cpro_	Navarra	13	597	C	5.2396	18
TRAIN	cpro_	La Rioja	10	595	C	5.222047	14
TRAIN	cpro_	Asturias	14	577	C	5.064069	3
TRAIN	cpro_	Canarias	15	487	C	4.274179	4
TRAIN	cpro_	Cantabria	16	450	C	3.949447	5
TRAIN	cpro_	Extremadura	5	447	C	3.923117	11
TRAIN	cpro_	Murcia	12	444	C	3.896788	17
TRAIN	cpro_	Islas Baleares	6	380	C	3.335089	13
TRAIN	cpro_	Melilla	18	63	C	0.552923	16
TRAIN	cpro_	Ceuta	17	46	C	0.403721	9
TRAIN	disp48_1_	No	0	10498	C	92.13621	1
TRAIN	disp48_1_	Si	1	896	C	7.863788	2
TRAIN	disp48_2_	No	1	10430	C	91.53941	1
TRAIN	disp48_2_	Si	0	964	C	8.460593	2
TRAIN	disp48_3_	No	0	10347	C	90.81095	1
TRAIN	disp48_3_	Si	1	1047	C	9.189047	2
TRAIN	disp48_4_	No	1	9543	C	83.75461	1
TRAIN	disp48_4_	Si	0	1851	C	16.24539	2
TRAIN	disp51_1_	Si	0	7581	C	66.53502	2
TRAIN	disp51_1_	No	1	3813	C	33.46498	1
TRAIN	disp51_2_	No	0	8397	C	73.69668	1
TRAIN	disp51_2_	Si	1	2997	C	26.30332	2
TRAIN	disp51_3_	No	1	9666	C	84.83412	1
TRAIN	disp51_3_	Si	0	1728	C	15.16588	2
TRAIN	disp52_1_	No	0	8772	C	76.98789	1
TRAIN	disp52_1_	Si	1	2622	C	23.01211	2
TRAIN	disp52_2_	No	0	10603	C	93.05775	1
TRAIN	disp52_2_	Si	1	791	C	6.94225	2

Rol de los datos	Nombre de la variable	Nivel	CODE	Número de ocurrencias	Tipo	Porcentaje	Índice de nivel
TRAIN	disp52_3_	No	0	11088	C	97.31438	1
TRAIN	disp52_3_	Si	1	306	C	2.685624	2
TRAIN	disp52_4_	No	0	10645	C	93.42636	1
TRAIN	disp52_4_	Si	1	749	C	6.573635	2
TRAIN	estc_	Casado/a	1	5753	C	50.49149	1
TRAIN	estc_	Soltero/a	0	4003	C	35.13253	4
TRAIN	estc_	Divorciado/a	2	903	C	7.925224	2
TRAIN	estc_	Viudo/a	3	513	C	4.50237	5
TRAIN	estc_	Separado/a	4	222	C	1.948394	3
TRAIN	frec_int_	Diariamente(5 días por sem...	0	9844	C	86.39635	1
TRAIN	frec_int_	Todas las semanas pero no ...	1	1219	C	10.69861	3
TRAIN	frec_int_	Menos de una vez a la sema...	2	331	C	2.905038	2
TRAIN	habitat_estrato_	Municipios capitales de provi...	0	2757	C	24.19695	1
TRAIN	habitat_estrato_	Municipios con menos de 10...	2	2230	C	19.5717	3
TRAIN	habitat_estrato_	Municipios con 500.000 ó m...	6	1518	C	13.3228	2
TRAIN	habitat_estrato_	Municipios entre 20.000 y m...	5	1491	C	13.08583	6
TRAIN	habitat_estrato_	Municipios entre 10.000 y m...	1	1228	C	10.7776	4
TRAIN	habitat_estrato_	Municipios entre 50.000 y m...	3	1188	C	10.42654	7
TRAIN	habitat_estrato_	Municipios entre 100.000 y ...	4	982	C	8.618571	5
TRAIN	habitat_eur_	Densely populated area	0	6367	C	55.88029	1
TRAIN	habitat_eur_	Intermediate density area	1	3327	C	29.19958	2
TRAIN	habitat_eur_	Thinly populated area	2	1700	C	14.92013	3
TRAIN	indol1_	Si	0	9147	C	80.27909	2
TRAIN	indol1_	No	1	2247	C	19.72091	1
TRAIN	indol2_	No	1	6743	C	59.18027	1
TRAIN	indol2_	Si	0	4651	C	40.81973	2
TRAIN	indol3_	No	0	10150	C	89.08197	1
TRAIN	indol3_	Si	1	1244	C	10.91803	2
TRAIN	indol4_	No	1	8830	C	77.49693	1
TRAIN	indol4_	Si	0	2564	C	22.50307	2
TRAIN	indol5_	Si	0	7195	C	63.14727	2
TRAIN	indol5_	No	1	4199	C	36.85273	1
TRAIN	indol6_	No	0	7594	C	66.64911	1
TRAIN	indol6_	Si	1	3800	C	33.35089	2
TRAIN	indol7_	No	1	8910	C	78.19905	1
TRAIN	indol7_	Si	0	2484	C	21.80095	2
TRAIN	ing_hog_	De 900 a menos de 1.600 e...	2	3289	C	28.86607	4
TRAIN	ing_hog_	De 1.600 a menos de 2.500 ...	0	2442	C	21.43233	2
TRAIN	ing_hog_	NS/NR	5	2019	C	17.71985	6
TRAIN	ing_hog_	Menos de 900 euros	3	1441	C	12.64701	5
TRAIN	ing_hog_	3.000 ó más euros	1	1274	C	11.18132	1
TRAIN	ing_hog_	De 2.500 a menos de 3.000 ...	4	929	C	8.153414	3
TRAIN	nacionalidad_	Española	0	10738	C	94.24258	1
TRAIN	nacionalidad_	Extranjera	1	450	C	3.949447	3
TRAIN	nacionalidad_	Española y otra	2	204	C	1.790416	2
TRAIN	nacionalidad_	Ninguna	3	2	C	0.017553	4
TRAIN	nivelest_	Segunda etapa de la Educac...	5	2758	C	24.20572	9
TRAIN	nivelest_	Primera etapa de la Educaci...	3	2586	C	22.69616	8
TRAIN	nivelest_	Grados universitarios de má...	2	1736	C	15.23609	6
TRAIN	nivelest_	Formación Profesional de Gr...	1	1416	C	12.42759	4
TRAIN	nivelest_	Grado universitario de 240 cr...	0	1359	C	11.92733	5
TRAIN	nivelest_	Educación Primaria	4	1124	C	9.864841	2
TRAIN	nivelest_	Analfabetos y estudios prima...	6	251	C	2.202914	1
TRAIN	nivelest_	Título de Doctorado	7	128	C	1.123398	10
TRAIN	nivelest_	No se puede codificar	9	19	C	0.166754	7
TRAIN	nivelest_	Educación postsecundaria n...	8	17	C	0.149201	3
TRAIN	nodisp1_	NA	0	5134	C	45.0588	1

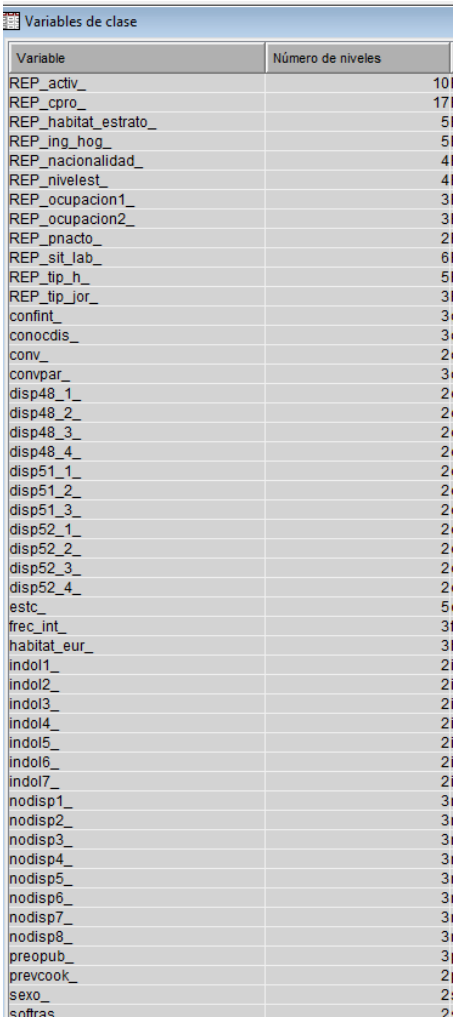
Variables de clase

Rol de los datos	Nombre de la variable	Nivel	CODE	Número de ocurrencias	Tipo	Porcentaje	Índice de nivel
TRAIN	nodisp1_	Si	1	4759	C	41.7676	3
TRAIN	nodisp1_	No	2	1501	C	13.1736	2
TRAIN	nodisp2_	NA	0	5134	C	45.0588	1
TRAIN	nodisp2_	No	2	4043	C	35.48359	2
TRAIN	nodisp2_	Si	1	2217	C	19.45761	3
TRAIN	nodisp3_	NA	0	5134	C	45.0588	1
TRAIN	nodisp3_	No	2	4939	C	43.34738	2
TRAIN	nodisp3_	Si	1	1321	C	11.59382	3
TRAIN	nodisp4_	NA	0	5134	C	45.0588	1
TRAIN	nodisp4_	No	1	4561	C	40.02984	2
TRAIN	nodisp4_	Si	2	1699	C	14.91136	3
TRAIN	nodisp5_	NA	0	5134	C	45.0588	1
TRAIN	nodisp5_	No	1	3997	C	35.07987	2
TRAIN	nodisp5_	Si	2	2263	C	19.86133	3
TRAIN	nodisp6_	NA	0	5134	C	45.0588	1
TRAIN	nodisp6_	No	1	3987	C	34.9921	2
TRAIN	nodisp6_	Si	2	2273	C	19.9491	3
TRAIN	nodisp7_	NA	0	5134	C	45.0588	1
TRAIN	nodisp7_	No	1	4978	C	43.68966	2
TRAIN	nodisp7_	Si	2	1282	C	11.25154	3
TRAIN	nodisp8_	NA	0	5134	C	45.0588	1
TRAIN	nodisp8_	No	1	5001	C	43.89152	2
TRAIN	nodisp8_	Si	2	1259	C	11.04968	3
TRAIN	ocupacion1_	NA	2	5438	C	47.72687	2
TRAIN	ocupacion1_	No manual	0	4410	C	38.70458	3
TRAIN	ocupacion1_	Manual	1	1546	C	13.56854	1
TRAIN	ocupacion2_	No TIC	0	5737	C	50.35106	2
TRAIN	ocupacion2_	NA	1	5438	C	47.72687	1
TRAIN	ocupacion2_	TIC	2	219	C	1.922064	3
TRAIN	pnacto_	España	0	10521	C	92.33807	2
TRAIN	pnacto_	De un país fuera UE	1	659	C	5.783746	1
TRAIN	pnacto_	Otro país UE	2	214	C	1.878181	3
TRAIN	preopub_	Algo preocupado	1	5682	C	49.86835	1
TRAIN	preopub_	Muy preocupado	2	2886	C	25.32912	2
TRAIN	preopub_	Nada preocupado	0	2826	C	24.80253	3
TRAIN	prevcook_	No	0	8156	C	71.58153	1
TRAIN	prevcook_	Si	1	3238	C	28.41847	2
TRAIN	sexo_	Mujer	1	6015	C	52.79094	2
TRAIN	sexo_	Hombre	0	5379	C	47.20906	1
TRAIN	sit_lab_	Trabajando por cuenta ajena...	1	4228	C	37.10725	8
TRAIN	sit_lab_	Jubilado ó prejubilado	4	1775	C	15.57837	3
TRAIN	sit_lab_	Parado	2	1485	C	13.03318	6
TRAIN	sit_lab_	Trabajando por cuenta propia	6	938	C	8.232403	10
TRAIN	sit_lab_	Estudiante	7	845	C	7.416184	1
TRAIN	sit_lab_	Trabajando por cuenta ajena...	0	790	C	6.933474	9
TRAIN	sit_lab_	Otra situación	3	603	C	5.292259	5
TRAIN	sit_lab_	Labores del hogar	5	503	C	4.414604	4
TRAIN	sit_lab_	Incapacitado permanente	8	220	C	1.930841	2
TRAIN	sit_lab_	Realizando tareas de voluntа...	9	7	C	0.061436	7
TRAIN	softras_	No	0	9804	C	86.04529	1
TRAIN	softras_	Si	1	1590	C	13.95471	2
TRAIN	tareainf1_	Si	0	7007	C	61.49728	2
TRAIN	tareainf1_	No	1	4387	C	38.50272	1
TRAIN	tareainf2_	Si	0	6242	C	54.78322	2
TRAIN	tareainf2_	No	1	5152	C	45.21678	1
TRAIN	tareainf3_	No	1	6674	C	58.57469	1
TRAIN	tareainf3_	Si	0	4720	C	41.42531	2
TRAIN	tareainf4_	Si	0	4502	C	39.51202	2
TRAIN	tareainf4_1_	NA	2	6892	C	60.48798	1
TRAIN	tareainf4_1_	Si	1	2947	C	25.86449	3
TRAIN	tareainf4_1_	No	0	1555	C	13.64753	2
TRAIN	tareainf5_	No	1	6792	C	59.61032	1
TRAIN	tareainf5_	Si	0	4602	C	40.38968	2
TRAIN	tareainf6_	No	0	10693	C	93.84764	1
TRAIN	tareainf6_	Si	1	701	C	6.152361	2
TRAIN	tip_h_	Pareja con hijos que conviva...	1	4322	C	37.93225	4
TRAIN	tip_h_	Hogar unipersonal	0	2416	C	21.20414	1
TRAIN	tip_h_	Pareja sin hijos que conviva...	3	2067	C	18.14113	5
TRAIN	tip_h_	Padre o madre solo/a que co...	4	1507	C	13.22626	3
TRAIN	tip_h_	Otro tipo de hogar	2	1082	C	9.496226	2
TRAIN	tip_jor_	NA	2	5438	C	47.72687	3
TRAIN	tip_jor_	A tiempo completo	1	5167	C	45.34843	1
TRAIN	tip_jor_	A tiempo parcial	0	789	C	6.924697	2
TRAIN	tmor1_	Si	0	6263	C	54.96753	2
TRAIN	tmor1_	No	1	5131	C	45.03247	1
TRAIN	tmor2_	Si	0	6709	C	58.88187	2
TRAIN	tmor2_	No	1	4685	C	41.11813	1
TRAIN	tmor3_	No	0	8522	C	74.79375	1
TRAIN	tmor3_	Si	1	2872	C	25.20625	2
TRAIN	usomov_	Si	0	11391	C	99.97367	2
TRAIN	usomov_	No	1	3	C	0.02633	1
TRAIN	vintd_	Si	0	9532	C	83.65807	3
TRAIN	vintd_	NA	1	1550	C	13.60365	1
TRAIN	vintd_	No	2	312	C	2.738283	2
TRAIN	viv_inter_	Si	0	11367	C	99.76303	2
TRAIN	viv_inter_	No	1	27	C	0.236967	1

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles ▼
cpro_	cpro_	C	19
activ_	activ_	C	12
nivelest_	nivelest_	C	10
sit_lab_	sit_lab_	C	9
habitat_estrato_	habitat_estrato_	C	7
ing_hog_	ing_hog_	C	6
estc_	estc_	C	5
tip_h_	tip_h_	C	5
nacionalidad_	nacionalidad_	C	4
confint_	confint_	C	3
conocdis_	conocdis_	C	3
convpar_	convpar_	C	3
frec_int_	frec_int_	C	3
habitat_eur_	habitat_eur_	C	3
nodisp1_	nodisp1_	C	3
nodisp2_	nodisp2_	C	3
nodisp3_	nodisp3_	C	3
nodisp4_	nodisp4_	C	3
nodisp5_	nodisp5_	C	3
nodisp6_	nodisp6_	C	3
nodisp7_	nodisp7_	C	3
nodisp8_	nodisp8_	C	3
ocupacion1_	ocupacion1_	C	3
ocupacion2_	ocupacion2_	C	3
pnacto_	pnacto_	C	3
preopub_	preopub_	C	3
tareainf4_1_	tareainf4_1_	C	3
tip_jor_	tip_jor_	C	3
vintd_	vintd_	C	3
conv_	conv_	C	2
disp48_1_	disp48_1_	C	2
disp48_2_	disp48_2_	C	2
disp48_3_	disp48_3_	C	2
disp48_4_	disp48_4_	C	2
disp51_1_	disp51_1_	C	2
disp51_2_	disp51_2_	C	2
disp51_3_	disp51_3_	C	2
disp52_1_	disp52_1_	C	2
disp52_2_	disp52_2_	C	2
disp52_3_	disp52_3_	C	2
disp52_4_	disp52_4_	C	2
indol1_	indol1_	C	2
indol2_	indol2_	C	2
indol3_	indol3_	C	2
indol4_	indol4_	C	2
indol5_	indol5_	C	2
indol6_	indol6_	C	2
indol7_	indol7_	C	2
prevcook_	prevcook_	C	2
sexo_	sexo_	C	2
softras_	softras_	C	2
tareainf1_	tareainf1_	C	2
tareainf2_	tareainf2_	C	2
tareainf3_	tareainf3_	C	2
tareainf4_	tareainf4_	C	2
tareainf5_	tareainf5_	C	2
tareainf6_	tareainf6_	C	2

Al haber recategorizado y agrupado algunas categorías, no solo disminuyeron la cantidad de niveles, sino que además, cada uno se encontró mejor representado.



Variables de clase

Variable	Número de niveles
REP_activ_	10
REP_cpro_	17
REP_habitat_estrato_	5
REP_ing_hog_	5
REP_nacionalidad_	4
REP_nivelest_	4
REP_ocupacion1_	3
REP_ocupacion2_	3
REP_pnacto_	2
REP_sit_lab_	6
REP_tip_h_	5
REP_tip_jor_	3
confint_	3
conocdis_	3
conv_	2
convpar_	3
disp48_1_	2
disp48_2_	2
disp48_3_	2
disp48_4_	2
disp51_1_	2
disp51_2_	2
disp51_3_	2
disp52_1_	2
disp52_2_	2
disp52_3_	2
disp52_4_	2
estc_	5
frec_int_	3
habitat_eur_	3
indol1_	2
indol2_	2
indol3_	2
indol4_	2
indol5_	2
indol6_	2
indol7_	2
nodisp1_	3
nodisp2_	3
nodisp3_	3
nodisp4_	3
nodisp5_	3
nodisp6_	3
nodisp7_	3
nodisp8_	3
preopub_	3
prevcook_	2
sexo_	2
softras	2

Anexo D: Resultados complementarios del análisis clúster.

Clúster 1

Variable	**Medida**	**Valor**
*sexo_*	Hombre	901
	Mujer	0
*tot_mh_*	Min.	1
	1st Qu.	1
	Median	2
	Mean	2.393
	3rd Qu.	3
	Max.	7
*REP_ing_hog_*	Más de 2.500	0
	De 1.600 a menos de 2.500 euros	663
	De 900 a menos de 1.600 euros	69
	Menos de 900 euros	41
	NS/NR	128
*REP_nivelest_*	Sin estudios o primaria	14
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	86

	Segunda etapa de educación secundaria y educación postsecundaria no superior	226
	Educación Superior	575
*REP_ocupacion2_*	No Aplica	1
	No TIC	843
	TIC	57
*REP_sit_lab_*	Estudiante	0
	Jubilado ó prejubilado	0
	Parado	0
	Otros	0
	Labores del hogar	1
	Trabajando	900
*REP_tip_jor_*	A tiempo completo	849
	A tiempo parcial	51
	No Aplica	1
CPConInf_UsoInt	Min.	-1.29835
	1st Qu.	-0.08079
	Median	0.53098
	Mean	0.38858
	3rd Qu.	0.9873
	Max.	1.32601
CPConfSegInt	Min.	-2.4299
	1st Qu.	-0.5606
	Median	0.3045
	Mean	0.116
	3rd Qu.	0.7867
	Max.	1.7588
*edad_*	Min.	17
	1st Qu.	38
	Median	45
	Mean	45.03
	3rd Qu.	53
	Max.	71

Clúster 2

Variable	**Medida**	**Valor**
*sexo_*	Hombre	0
	Mujer	851
*tot_mh_*	Min.	1
	1st Qu.	2
	Median	3
	Mean	3.068
	3rd Qu.	4
	Max.	13

*REP_ing_hog_*	Más de 2.500	698
	De 1.600 a menos de 2.500 euros	0
	De 900 a menos de 1.600 euros	17
	Menos de 900 euros	35
	NS/NR	101
*REP_nivelest_*	Sin estudios o primaria	9
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	42
	Segunda etapa de educación secundaria y educación postsecundaria no superior	103
	Educación Superior	697
*REP_ocupacion2_*	No Aplica	2
	No TIC	816
	TIC	33
*REP_sit_lab_*	Estudiante	0
	Jubilado ó prejubilado	0
	Parado	0
	Otros	0
	Labores del hogar	2
	Trabajando	849
*REP_tip_jor_*	A tiempo completo	724
	A tiempo parcial	125
	No Aplica	2
CPConInf_UsoInt	Min.	-1.2984
	1st Qu.	0.2797
	Median	0.7731
	Mean	0.5516
	3rd Qu.	1.0228
	Max.	1.326
CPConfSegInt	Min.	-2.42441
	1st Qu.	-0.55463
	Median	0.29898
	Mean	0.09011
	3rd Qu.	0.7963
	Max.	1.74729
*edad_*	Min.	18
	1st Qu.	39
	Median	45
	Mean	45.24
	3rd Qu.	53

	Max.	68

Clúster 3

Variable	**Medida**	**Valor**
*sexo_*	Hombre	0
	Mujer	815
*tot_mh_*	Min.	1
	1st Qu.	2
	Median	3
	Mean	2.625
	3rd Qu.	4
	Max.	6
*REP_ing_hog_*	Más de 2.500	0
	De 1.600 a menos de 2.500 euros	677
	De 900 a menos de 1.600 euros	0
	Menos de 900 euros	28
	NS/NR	110
*REP_nivelest_*	Sin estudios o primaria	14
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	77
	Segunda etapa de educación secundaria y educación postsecundaria no superior	135
	Educación Superior	589
*REP_ocupacion2_*	No Aplica	0
	No TIC	803
	TIC	12
*REP_sit_lab_*	Estudiante	0
	Jubilado ó prejubilado	0
	Parado	0
	Otros	0
	Labores del hogar	0
	Trabajando	815
*REP_tip_jor_*	A tiempo completo	647
	A tiempo parcial	168
	No Aplica	0
CPConInf_UsoInt	Min.	-1.2984
	1st Qu.	-0.2099
	Median	0.4484
	Mean	0.2586
	3rd Qu.	0.7969
	Max.	1.326
CPConfSegInt	Min.	-2.46851
	1st Qu.	-0.59322

	Median	0.28747
	Mean	0.04229
	3rd Qu.	0.77935
	Max.	1.75878
*edad_*	Min.	17
	1st Qu.	39
	Median	45
	Mean	45.71
	3rd Qu.	53
	Max.	71

Clúster 4

Variable	**Medida**	**Valor**
*sexo_*	Hombre	1354
	Mujer	0
*tot_mh_*	Min.	1
	1st Qu.	2
	Median	2
	Mean	2.195
	3rd Qu.	3
	Max.	10
*REP_ing_hog_*	Más de 2.500	209
	De 1.600 a menos de 2.500 euros	335
	De 900 a menos de 1.600 euros	547
	Menos de 900 euros	103
	NS/NR	160
	NA	
*REP_nivelest_*	Sin estudios o primaria	264
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	309
	Segunda etapa de educación secundaria y educación postsecundaria no superior	354
	Educación Superior	427
	No Aplica	1354
*REP_ocupacion2_*	No TIC	0
	TIC	0
*REP_sit_lab_*	Estudiante	14
	Jubilado ó prejubilado	802
	Parado	305
	Otros	225
	Labores del hogar	8
	Trabajando	0
*REP_tip_jor_*	A tiempo completo	0

	A tiempo parcial	0
	No Aplica	1354
CPConInf_UsoInt	Min.	-1.2984
	1st Qu.	-0.9816
	Median	-0.5543
	Mean	-0.351
	3rd Qu.	0.2806
	Max.	1.326
CPConfSegInt	Min.	-2.40599
	1st Qu.	-0.612396
	Median	0.231406
	Mean	0.002282
	3rd Qu.	0.723216
	Max.	1.758784
*edad_*	Min.	16
	1st Qu.	55
	Median	64
	Mean	60.16
	3rd Qu.	69
	Max.	74

Clúster 5

Variable	**Medida**	**Valor**
*sexo_*	Hombre	0
	Mujer	1176
*tot_mh_*	Min.	1
	1st Qu.	1
	Median	2
	Mean	2.353
	3rd Qu.	3
	Max.	7
*REP_ing_hog_*	Más de 2.500	6
	De 1.600 a menos de 2.500 euros	0
	De 900 a menos de 1.600 euros	810
	Menos de 900 euros	171
	NS/NR	189
*REP_nivelest_*	Sin estudios o primaria	123
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	286
	Segunda etapa de educación secundaria y educación postsecundaria no superior	353

	Educación Superior	414
*REP_ocupacion2_*	No Aplica	0
	No TIC	1169
	TIC	7
*REP_sit_lab_*	Estudiante	0
	Jubilado ó prejubilado	0
	Parado	0
	Otros	0
	Labores del hogar	0
	Trabajando	1176
*REP_tip_jor_*	A tiempo completo	876
	A tiempo parcial	300
	No Aplica	0
CPConInf_UsoInt	Min.	-1.2984
	1st Qu.	-0.9312
	Median	-0.2158
	Mean	-0.1509
	3rd Qu.	0.509
	Max.	1.326
CPConfSegInt	Min.	-2.3674
	1st Qu.	-0.9453
	Median	-0.2904
	Mean	-0.2049
	3rd Qu.	0.6822
	Max.	1.7703
*edad_*	Min.	17
	1st Qu.	40
	Median	47
	Mean	46.91
	3rd Qu.	55
	Max.	72

Clúster 26

Variable	**Medida**	**Valor**
*sexo_*	Hombre	0
	Mujer	1633
*tot_mh_*	Min.	1
	1st Qu.	2
	Median	2
	Mean	2.491
	3rd Qu.	3
	Max.	7

*REP_ing_hog_*	Más de 2.500	243
	De 1.600 a menos de 2.500 euros	291
	De 900 a menos de 1.600 euros	879
	Menos de 900 euros	0
	NS/NR	220
*REP_nivelest_*	Sin estudios o primaria	216
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	284
	Segunda etapa de educación secundaria y educación postsecundaria no superior	524
	Educación Superior	609
*REP_ocupacion2_*	No Aplica	1633
	No TIC	0
	TIC	0
*REP_sit_lab_*	Estudiante	103
	Jubilado ó prejubilado	488
	Parado	475
	Otros	311
	Labores del hogar	256
	Trabajando	0
*REP_tip_jor_*	A tiempo completo	0
	A tiempo parcial	0
	No Aplica	1633
CPConInf_UsoInt	Min.	-1.2984
	1st Qu.	-0.9312
	Median	-0.2895
	Mean	-0.215
	3rd Qu.	0.5061
	Max.	1.326
CPConfSegInt	Min.	-2.44142
	1st Qu.	-0.50128
	Median	0.24291
	Mean	0.06667
	3rd Qu.	0.754
	Max.	1.78588
*edad_*	Min.	16
	1st Qu.	41
	Median	55
	Mean	52.5
	3rd Qu.	66
	Max.	74

Clúster 7

Variable	Medida	Valor
*sexo_*	Hombre	225
	Mujer	1192
*tot_mh_*	Min.	1
	1st Qu.	1
	Median	2
	Mean	2.079
	3rd Qu.	3
	Max.	8
*REP_ing_hog_*	Más de 2.500	69
	De 1.600 a menos de 2.500 euros	222
	De 900 a menos de 1.600 euros	0
	Menos de 900 euros	843
	NS/NR	283
*REP_nivelest_*	Sin estudios o primaria	501
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	559
	Segunda etapa de educación secundaria y educación postsecundaria no superior	196
	Educación Superior	161
	No Aplica	1401
*REP_ocupacion2_*	No TIC	16
	TIC	0
*REP_sit_lab_*	Estudiante	28
	Jubilado ó prejubilado	458
	Parado	496
	Otros	187
	Labores del hogar	232
	Trabajando	16
*REP_tip_jor_*	A tiempo completo	0
	A tiempo parcial	16
	No Aplica	1401
CPConInf_UsoInt	Min.	-1.2984
	1st Qu.	-1.0589
	Median	-0.9579
	Mean	-0.7108
	3rd Qu.	-0.5189
	Max.	1.326
CPConfSegInt	Min.	-2.42992
	1st Qu.	-0.88459
	Median	-0.06871
	Mean	-0.11796
	3rd Qu.	0.5109
	Max.	1.77082
*edad_*	Min.	16

1st Qu.	49
Median	60
Mean	56.64
3rd Qu.	67
Max.	74

Clúster 8

Variable	**Medida**	**Valor**
*sexo_*	Hombre	999
	Mujer	0
*tot_mh_*	Min.	1
	1st Qu.	2
	Median	3
	Mean	3.004
	3rd Qu.	4
	Max.	8
*REP_ing_hog_*	Más de 2.500	753
	De 1.600 a menos de 2.500 euros	0
	De 900 a menos de 1.600 euros	103
	Menos de 900 euros	27
	NS/NR	116
*REP_nivelest_*	Sin estudios o primaria	3
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	54
	Segunda etapa de educación secundaria y educación postsecundaria no superior	167
	Educación Superior	775
*REP_ocupacion2_*	No Aplica	2
	No TIC	897
	TIC	100
*REP_sit_lab_*	Estudiante	0
	Jubilado ó prejubilado	0
	Parado	0
	Otros	0
	Labores del hogar	2
	Trabajando	997
*REP_tip_jor_*	A tiempo completo	949
	A tiempo parcial	48
	No Aplica	2
CPConInf_UsoInt	Min.	-1.2211
	1st Qu.	0.5847

	Median	0.9605
	Mean	0.7935
	3rd Qu.	1.1983
	Max.	1.326
CPConfSegInt	Min.	-2.4207
	1st Qu.	-0.6373
	Median	0.2664
	Mean	-0.0445
	3rd Qu.	0.7568
	Max.	1.7441
*edad_*	Min.	16
	1st Qu.	39
	Median	46
	Mean	45.51
	3rd Qu.	53
	Max.	70

Clúster 9

Variable	**Medida**	**Valor**
*sexo_*	Hombre	697
	Mujer	348
*tot_mh_*	Min.	1
	1st Qu.	3
	Median	3
	Mean	3.433
	3rd Qu.	4
	Max.	9
*REP_ing_hog_*	Más de 2.500	186
	De 1.600 a menos de 2.500 euros	188
	De 900 a menos de 1.600 euros	53
	Menos de 900 euros	105
	NS/NR	513
*REP_nivelest_*	Sin estudios o primaria	41
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	359
	Segunda etapa de educación secundaria y educación postsecundaria no superior	396
	Educación Superior	249
*REP_ocupacion2_*	No Aplica	1045
	No TIC	0
	TIC	0

*REP_sit_lab_*	Estudiante	700
	Jubilado ó prejubilado	27
	Parado	209
	Otros	107
	Labores del hogar	2
	Trabajando	0
*REP_tip_jor_*	A tiempo completo	0
	A tiempo parcial	0
	No Aplica	1045
CPConInf_UsoInt	Min.	-1.2984
	1st Qu.	0.4956
	Median	0.8185
	Mean	0.6586
	3rd Qu.	1.0248
	Max.	1.326
CPConfSegInt	Min.	-2.3674
	1st Qu.	-0.5546
	Median	0.334
	Mean	0.1778
	3rd Qu.	0.8404
	Max.	1.7648
*edad_*	Min.	16
	1st Qu.	18
	Median	21
	Mean	25.79
	3rd Qu.	28
	Max.	74

Clúster 10

Variable	**Medida**	**Valor**
*sexo_*	Hombre	1203
	Mujer	0
*tot_mh_*	Min.	1
	1st Qu.	1
	Median	3
	Mean	2.573
	3rd Qu.	4
	Max.	10
*REP_ing_hog_*	Más de 2.500	39
	De 1.600 a menos de 2.500 euros	66
	De 900 a menos de 1.600 euros	811

	Menos de 900 euros	88
	NS/NR	199
*REP_nivelest_*	Sin estudios o primaria	209
	Primera etapa de la Educación Secundaria y similar(Bachiller elemental, EGB, ESO…)	530
	Segunda etapa de educación secundaria y educación postsecundaria no superior	321
	Educación Superior	143
*REP_ocupacion2_*	No Aplica	0
	No TIC	1193
	TIC	10
*REP_sit_lab_*	Estudiante	0
	Jubilado ó prejubilado	0
	Parado	0
	Otros	0
	Labores del hogar	0
	Trabajando	1203
*REP_tip_jor_*	A tiempo completo	1122
	A tiempo parcial	81
	No Aplica	0
CPConInf_UsoInt	Min.	-1.29835
	1st Qu.	-0.98163
	Median	-0.67577
	Mean	-0.41571
	3rd Qu.	0.04921
	Max.	1.32601
CPConfSegInt	Min.	-2.4478
	1st Qu.	-0.71184
	Median	-0.06871
	Mean	-0.05059
	3rd Qu.	0.72322
	Max.	1.75932
*edad_*	Min.	17
	1st Qu.	40
	Median	48
	Mean	47.23
	3rd Qu.	55
	Max.	69