Inteligencia de Negocio - Práctica 2
Análisis Relacional mediante Segmentación

Sergio Quijano Rey - 72103503k
sergioquijano@correo.ugr.es
5º Doble Grado Ingeniería Informática y Matemáticas
Grupo de prácticas 1

10 de diciembre de 2021

Índice

1. Introducción **3**

1.1. Problema a resolver 3

1.2. Resumen de la forma de proceder 3

1.3. Observaciones y secciones previas 3

2. Caso de estudio 1 **6**

2.1. Definición del caso de estudio 6

2.2. Procesado de los datos 6

2.3. Análisis Exploratorio de datos 7

2.4. Resultados de *clustering* 10

2.5. Interpretación de la segmentación 10

3. Caso de estudio 2 **11**

4. Caso de estudio 3 **12**

5. Referencias **13**

Índice de figuras

1. Gráfico en el que se muestran todas las combinaciones de parejas de variables involucradas en el caso de estudio 7

2. Gráfico con el porcentaje de hogares que viven en entornos con vandalismo, agrupados por comunidades autónomas 8

3. Gráfico en el que se aprecia claramente la correlación entre la variable `renta_disponible_total_h` y la variable `renta_disponible_restado_alquiler` 9

4. Gráfico en el que se aprecia claramente la correlación entre la variable `renta_disponible_total_h` y la variable agregada `renta_disponible_restado_alquiler` 10

1. Introducción

1.1. Problema a resolver

En esta práctica, trabajaremos con un conjunto de datos del *"Instituto Nacional de Estadística"*. En concreto, trabajamos con la *"Encuesta de condiciones de vida, 2020"* [1]. En términos del *"INE"*, trabajamos con microdatos, lo que nos va a permitir realizar estudios en profundidad sobre la encuesta estudiada, más allá de los resúmenes basados en el estudio de las frecuencias, clásicos de las encuestas.

1.2. Resumen de la forma de proceder

Aunque más adelante, en el desarrollo de cada caso de estudio, se haga claro el procedimiento considerado, lo resumimos aquí de forma previa.

En cada caso de estudio, empezaremos definiendo el filtro que fija el caso de estudio en sí. Esto es, definimos la condición que filtrará las filas de nuestro conjunto de datos que estudiamos. Además, definimos las variables que consideramos, es decir, definimos las columnas del *dataset* que consideramos para este caso de estudio.

Realizamos un procesado básico de los datos (borrado de datos faltantes, datos `NaN`, borrado de *outliers* y normalización). A partir de este procesado, realizamos cierto análisis exploratorio de datos, para tomar ideas a la hora de atacar cada problema en concreto.

A continuación, aplicamos los algoritmos de *clustering*, mostrando gráficamente el resultado obtenido. Una vez ejecutados todos los algoritmos, calculamos un conjunto de métricas sobre los etiquetados generados, para poder comparar los resultados de los algoritmos. Visualizamos estas métricas y proyecciones del etiquetado a un espacio bidimensional (usando *PCA* y *tsne*).

Seguidamente, exploramos los parámetros de un par de algoritmos por cada caso de estudio, en busca de encontrar unos parámetros optimales para realizar la *clusterización*.

Para finalizar, en esta memoria realizamos unas conclusiones, tanto del comportamiento de los algoritmos empleados como del conocimiento extraído sobre el conjunto de datos, gracias a las técnicas que hemos empleado.

1.3. Observaciones y secciones previas

Todo el código lo hemos organizado en un único *Notebook* de *Jupyter*. Para mantener el orden, hemos agrupado el código en secciones y subsecciones. Se puede explorar cómodamente esta jerarquía empleando *Jupyter Lab* en vez de *Jupyter Notebook*, o directamente empleado *Google Colab*.

Hemos desarrollado una **sección inicial** con funcionalidades que vamos a estar usando durante todo el *notebook*, evitando así repetir excesivamente el mismo código. Además, con esto, en las secciones de los casos de estudio nos centramos más en plantear ideas y ver los resultados, evitando toda la suciedad que supondría colocar en medio de estos desarrollos todo el código con la implementación. Esta sección previa de código común se divide en las siguientes subsecciones:

1. Un decorador de `Python` para poder medir fácilmente, y sin complicar el código, las funciones que calculan los *clusters*. Esto es básico pues es una métrica que debemos estudiar, y que las librerías que estamos usando (principalmente `sklearn`), no nos aportan de forma fácil

2. Funciones para calcular *clusters*. Destaca la función `add_clustering_labels`, que toma como parámetro una función de clusterización con sus parámetros, y se encarga de calcular la clusterización y añadir las etiquetas al `dataframe` convenientemente. Además, podemos especificar qué columnas ignorar en el `dataframe` a la hora de calcular la clusterización (para no usar columnas con etiquetas de otros algoritmos, principalmente)

3. Funciones para evaluar *clusters*. En esta subsección definimos unas cuantas métricas que usaremos para comparar los resultados. La función `compute_clustering_metrics` agrupa todas estas métricas y las devuelve convenientemente en un diccionario de `Python`.

4. Funciones para realizar visualizaciones. Con esto es más fácil usar las funciones de `seaborn` , evitando tener que repetir código para, por ejemplo, excluir ciertas columnas del `dataframe` . Destaca la función `plot_centroids_with_size`. En esta función:

   - Calculamos con código propio, por cada *cluster*, el centroide y el radio de este, y el número de elementos en cada *cluster*
   - Aplicamos *tsne* como técnica de proyección en un espacio bidimensional de los datos de entrada
   - Con esta proyección, mostramos gráficamente los centroides obtenidos en la *clusterización*, con un tamaño proporcional al radio previamente calculado, y con un color proporcional al número de elementos de cada *cluster*

5. Funciones para procesar los *dataframes*. Con esto, evitamos repetir el código que normaliza los datos, borra datos `NaN` y borra *outliers*

Antes de pasar a los casos de estudio, debemos comentar otras dos secciones, la sección de filtrado previo del *dataset* global y la sección de variables añadidas.

En la **sección de filtrado previo del *dataset* global**, definimos el conjunto de variables que nos interesan, descartando el resto por completo. Esta selección se realizó al principio del desarrollo de la práctica, buscando una primera exploración del conjunto de datos. Por esto se seleccionan variables que más tarde no se usan para nada. Pero decidimos dejar estas variables en el *Notebook* como ilustración de cuál fue nuestro proceso a la hora de abordar el problema. Nos quedamos, en esta fase, con un total de **39 variables**.

En la **sección de variables añadidas**, definimos algunas variables agregadas (que combinan información de varias variables, o que transforman la información en bruto) a nuestro *dataset*. En concreto:

- Definimos la variable `booleana` `pidio_ayuda` que controla si un hogar pidió o no pidió ayuda, ya sea a un familiar o a otro tipo de entidad.
- Definimos la variable `gasto_transporte_total` que combina el gasto en transporte público y transporte privado
- Definimos la variable `retraso_pago`, que controla si se produjo cualquier tipo de retraso en los pagos

- Definimos la variable `ingresos_menos_gastos`, que realiza la resta de todos los ingresos de un hogar menos todos los gastos, registrados en el *dataset*, de ese hogar
- Definimos la variable ordinal `comunidad_autonoma_code`, que convierte el código en texto de una comunidad autónoma a una variable entera. Esto con la intención de poder usar esta variable en los algoritmos de *clusterización*
- Definimos la variable `habitaciones_por_persona` que calcula el cociente del número de integrantes de un hogar entre el número de habitaciones de dicho hogar

Con todo esto, pasamos a tener un total de **45 variables**. De nuevo, algunas de estas variables añadidas no se usan en los casos de estudio, pero las dejamos para ilustrar el proceso realizado en la búsqueda de casos de estudio interesantes.

A partir de este punto ya nos encontramos con las secciones propias a los casos de estudio, que pasamos a comentar en cada sección correspondiente.

2. Caso de estudio 1

2.1. Definición del caso de estudio

En este caso de estudio queremos poner nuestra atención en aquellas personas que **viven en un entorno en el que hay vandalismo**. Para ello usamos la variable `vandalismo_en_la_zona` , que se corresponde con el código `"HS190"`.

Una vez filtradas las filas que vamos a estudiar, filtramos las variables (columnas) que nos interesan. En concreto, nos quedamos con la renta disponible del hogar, los gatos mensuales de la vivienda, el gasto total en transporte y los códigos de la comunidad autónoma. Nos interesa realizar los siguientes estudios:

- Estudiar la distribución del vandalismo por cada comunidad autónoma
- Estudiar el uso del transporte en las zonas conflictivas
- Lógicamente, nos interesa el factor económico involucrado en definir zonas posiblemente conflictivas. La hipótesis básica es que en las zonas de menor renta hay más vandalismo. Pero también puede haber vandalismo en las zonas más pudientes (robos, asaltos a casas, ...)
- El gasto en vivienda puede ser interesante como otra variable para fijar el estatus socio-económico. En ciertas comunidades autónomas, un alto gasto en vivienda no significa un nivel alto de vida (precios demasiado inflados, precios en zonas urbanas mucho más altos que en zonas rurales, ...)

Además, en el *dataframe* que usamos para almacenar estos datos (`df_study_case`), guardamos el factor de elevación, que usaremos para ponderar los ejemplos, en aquellos algoritmos que acepten este parámetro de ponderación.

Tras el filtrado, obtenemos un *dataset* con 2081 filas (objetos a *clusterizar*) y 5 columnas (4 + 1 para el factor de elevación).

2.2. Procesado de los datos

Como ya hemos comentado previamente, realizaremos el siguiente pre-procesado de los datos:

- Borrado de las filas que contengan algún valor `NaN`
- Borrado de *outliers* usando la regla $3 \cdot IQR$, variable por variable
- Normalización del rango de las variables al intervalo $[0, 1]$

Tras aplicar esto, nos quedamos con un *dataset* de 1947 filas.

2.3. Análisis Exploratorio de datos

Comenzamos haciendo un *plot* por pares de variables, que mostramos en la siguiente figura:

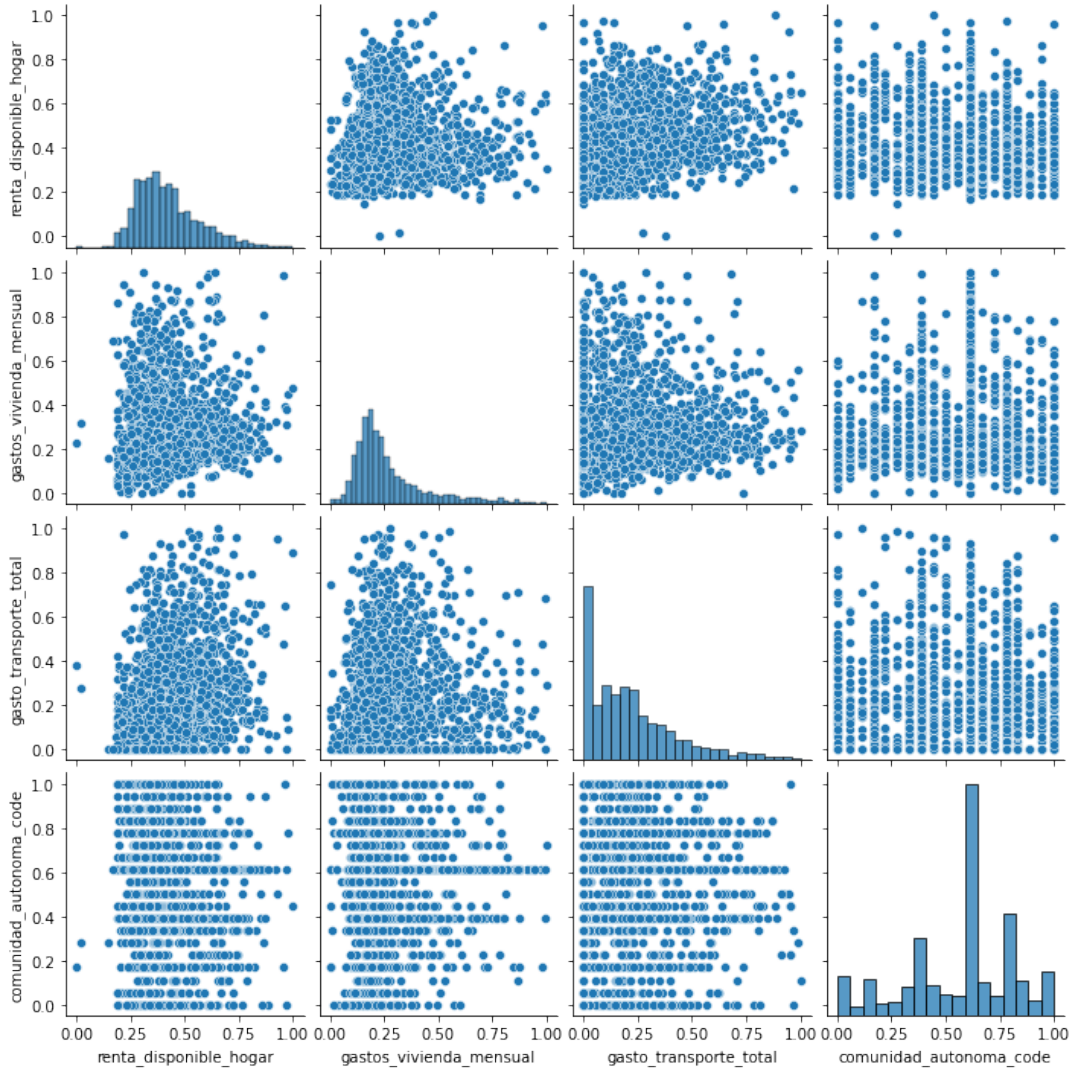

Figura 1: Gráfico en el que se muestran todas las combinaciones de parejas de variables involucradas en el caso de estudio

En este gráfico tenemos las distribuciones al poner en un eje una variable y en el otro eje otra variable, considerando todas las parejas posibles. En la diagonal tenemos la distribución de cada variable, por sí sola.

Lo que más llama la atención de este gráfico es que, en la distribución de las comunidades autónomas, hay una comunidad que destaca claramente sobre otras. Esto podría ser porque dicha comunidad autónoma tiene una mayor representación en la encuesta, o no. Por tanto, es necesario visualizar dicha distribución de forma porcentual, considerando como base para el cálculo toda la población.

El resto de distribuciones tienen una forma parecida a la normal, lo que era de esperar. Salvo la distribución del gasto en transporte, donde el gasto mínimo (puede que no sea nulo, pues hemos normalizado las variables al rango $[0, 1]$) destaca sobre el resto de valores, rompiendo con

una distribución normal que esperábamos.

Se puede apreciar cierta correlación positiva entre la renta disponible en el hogar y el gasto en vivienda mensual. Pero parece ser una correlación muy débil y que no destaca nada interesante (es lógico que a mayor dinero disponible, generalmente se gaste más en vivienda).

La correlación entre gasto en vivienda y gasto en transporte no parece aportar nada nuevo al estudio que queremos realizar.

Por la naturaleza ordinal del código de la comunidad autónoma, no podemos extraer mucha información al respecto usando esta gráfica y esta variable en concreto.

Como comentábamos previamente, destaca los picos de ciertas comunidades autónomas en el número de hogares que viven en un entorno de vandalismo. Pero esto puede deberse a diferencias en tamaños de las comunidades. Así que pasamos a mostrar el porcentaje de hogares que viven en entornos con vandalismo, agrupados por comunidades autónomas:

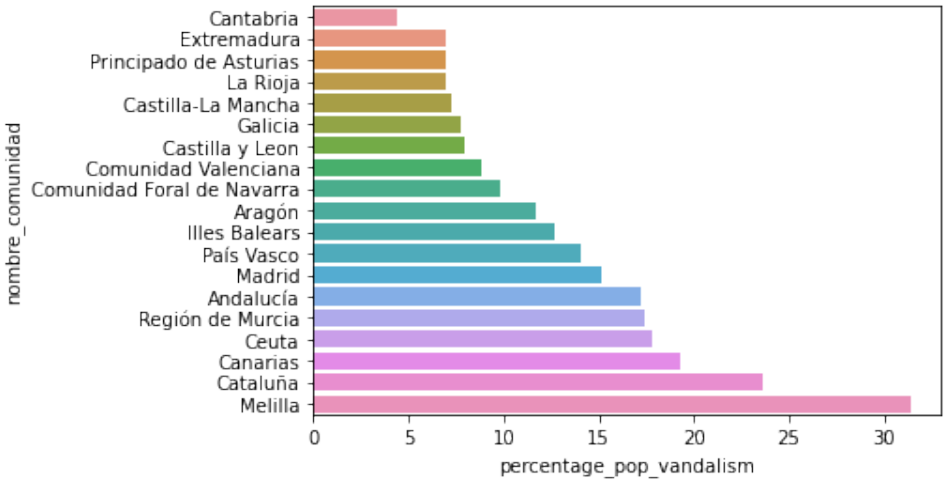

Figura 2: Gráfico con el porcentaje de hogares que viven en entornos con vandalismo, agrupados por comunidades autónomas

En este gráfico podemos ver que, como suponíamos, hay ciertas comunidades autónomas con un porcentaje de hogares en zonas con vandalismo mucho mayores que otras comunidades. Destaca la ciudad autónoma de Melilla, que se distancia de la segunda peor comunidad en esta métrica prácticamente un 10 %. Además, a la vista de estos datos, y sin necesidad de introducir datos externos sobre poblaciones de las comunidades autónomas, parece que las zonas más rurales y con menos concentración de población (Cantabria, Extremadura, Asturias, ...) tienen porcentajes de vandalismo significativamente mayor que comunidades más densamente pobladas (Cataluña, Andalucía, Madrid, ...). Destacan, saliéndose de lo que acabamos de comentar, las dos ciudades autónomas.

Aprovechamos la sección de exploración de datos de este caso de estudio para realizar la siguiente observación, que tendrá impacto en el resto de casos de estudio. En un primer momento, consideramos usar las variables `renta_disponible_total_hogar` y `renta_disponible_restado_alq` . Sin embargo, como mostramos a continuación, están muy correladas, por lo que no aportan información interesante:

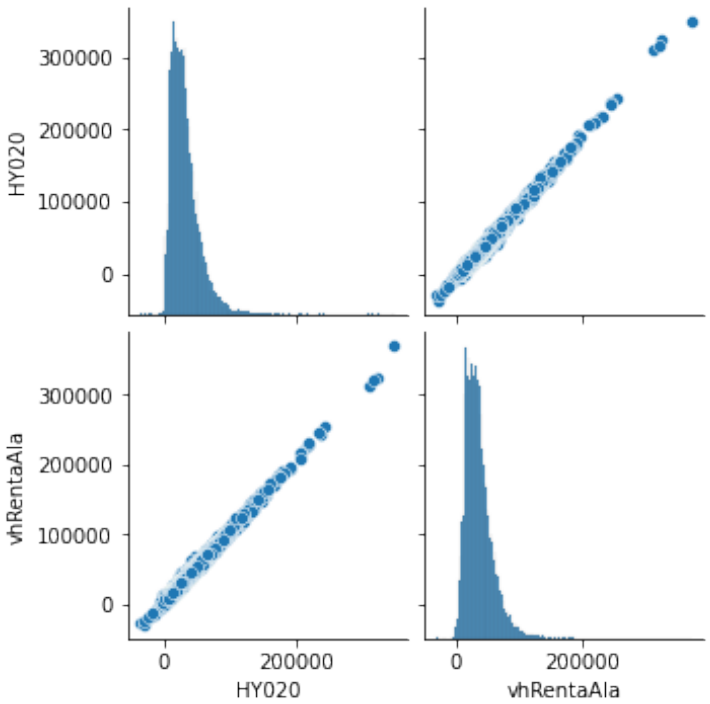

Figura 3: Gráfico en el que se aprecia claramente la correlación entre la variable `renta_disponible_total_hogar` y la variable `renta_disponible_restado_alquiler`

Lo mismo ocurre con la variable agregada por nosotros, `ingresos_menos_gastos`, y la variable ya presente en los datos originales, `renta_disponible_total_hogar`. Esto se muestra claramente en la siguiente gráfica:

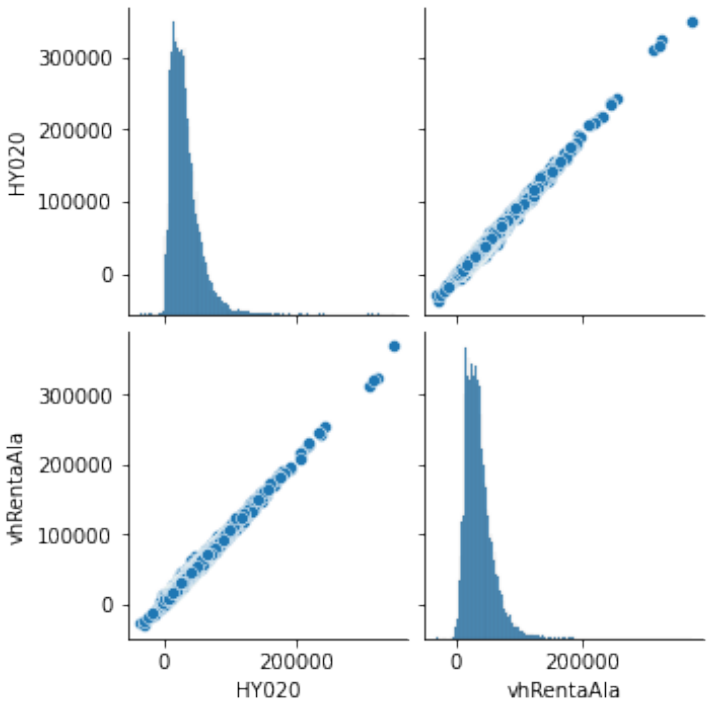

Figura 4: Gráfico en el que se aprecia claramente la correlación entre la variable `renta_disponible_total_hogar` y la variable agregada `renta_disponible_restado_alquiler`

Por tanto, no tiene sentido usar ambas variables en un mismo caso de estudio, pues apenas aportan información nueva al problema, debido a la alta correlación entre ambas.

2.4. Resultados de *clustering*

2.5. Interpretación de la segmentación

3. Caso de estudio 2

4. Caso de estudio 3

5. Referencias

[1] "Inebase / nivel y condiciones de vida (ipc) /condiciones de vida /encuesta de condiciones de vida / resultados." https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176807&menu=resultados&idp=1254735976608. (Accessed on 12/09/2021).