

**UNIVERSIDAD DE GRANADA**  
**E.T.S. de Ingenierías Informática y de Telecomunicación**



**Departamento de Ciencias de la  
Computación e Inteligencia Artificial**

# **Inteligencia de Negocio**

## **Guion de Prácticas**

**Práctica 2:  
Análisis Relacional mediante Segmentación**

Curso 2019-2020

Cuarto Curso del Grado en Ingeniería Informática

# Práctica 2

## Segmentación para Análisis Empresarial

### 1. Objetivos y Evaluación

En esta segunda práctica de la asignatura Inteligencia de Negocio veremos el uso de técnicas de aprendizaje no supervisado para análisis relacional mediante segmentación. Se trabajará con un conjunto de datos sobre el que se aplicarán distintos algoritmos de agrupamiento (*clustering*). A la luz de los resultados obtenidos se deberán crear informes y análisis lo suficientemente profundos.

La práctica se calificará hasta un **máximo de 2 puntos**. Se valorará el acierto en los recursos de análisis gráficos empleados, la complejidad de los experimentos realizados, la interpretación de los resultados, la organización y redacción del informe, etc.

### 2. Descripción del problema: perfiles de la población granadina

A partir de los microdatos publicados por el Instituto Nacional de Estadística (INE) en 2018 sobre la última encuesta de fecundidad<sup>1</sup>, se dispone de un conjunto de 14.556 respuestas de mujeres a una encuesta con 463 variables sobre datos personales, datos biográficos, hogar, vivienda, padres, relaciones de pareja, hijos, fecundidad, estudios, empleo y creencias. En la web de la asignatura se incluye el conjunto de datos —procesado a partir de la fuente original— sobre el que se trabajará en esta práctica.

Muchas variables son categóricas como, por ejemplo, *estado civil* (soltera, casada, viuda, separada o divorciada). Estas variables no sirven para aplicar un análisis de agrupamiento, pero sí son útiles para fijar casos de estudio donde centrar el análisis. Hay otras variables numéricas

---

<sup>1</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736177006&menu=ultiDatos&idp=1254735573002](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177006&menu=ultiDatos&idp=1254735573002)

como, por ejemplo, *edad* que sí se pueden usar para *clustering*. Finalmente, hay también variables que, aunque no son numéricas, sí son ordinales (por ejemplo, *estudios alcanzados*) y, por tanto, también se pueden usar para *clustering*.

El objetivo de la práctica es definir algunos casos de estudio de interés (fijando condiciones en algunas variables), aplicar distintos algoritmos de clustering, analizar la calidad de las soluciones obtenidas y, finalmente, interpretar los resultados para explicar los distintos perfiles o grupos encontrados.

Con este análisis se consigue encontrar relaciones entre variables, yendo así más allá del análisis de encuesta habitual centrada en valorar la frecuencia de variables de forma independiente, como se refleja en el resumen del INE de la figura 2.1.

### 3. Tareas a Realizar

La práctica consiste en aplicar y analizar técnicas de agrupamiento para descubrir grupos en el conjunto de datos bajo estudio. El trabajo se realizará empleando bibliotecas y paquetes de Python, principalmente `numpy`, `pandas`, `scikit-learn`, `matplotlib` y `seaborn`. Se recomienda consultar los siguientes enlaces:

- <http://scikit-learn.org/stable/modules/clustering.html>
- <http://www.learn datasci.com/k-means-clustering-algorithms-python-intro/>
- [http://hdbscan.readthedocs.io/en/latest/comparing\\_clustering\\_algorithms.html](http://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html)
- <https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/>
- <http://seaborn.pydata.org/generated/seaborn.clustermap.html>

Nos interesaremos en segmentar la población seleccionando previamente grupos de interés según las variables categóricas y/u ordinales. Por ejemplo, analizar solo la población mayor de 30 años, la que tiene pareja con la que no convive, quien lleva cierto tiempo en paro, etc. Queda a elección libre del alumno escoger varios casos (al menos tres) y realizar el estudio sobre ellos. Será necesario también aplicar una normalización para que las métricas de distancia y la visualización funcionen correctamente. Deberán justificarse las decisiones tomadas respecto al tratamiento de las variables.

En cada caso de estudio se analizarán 5 algoritmos distintos de agrupamiento (siendo al menos uno de ellos K-means) obteniéndose el tiempo de ejecución y métricas de rendimiento tales como Silhouette y el índice Calinski-Harabaz. Además, se analizará el efecto de algunos parámetros determinantes (por ejemplo, el valor de  $k$  si el algoritmo necesita fijarlo *a priori*) en al menos 2 algoritmos distintos para cada caso de estudio.

El análisis deberá apoyarse en visualizaciones tales como nubes de puntos (*scatter matrix*), dendrogramas y mapas de temperatura (*heatmap*). Por ejemplo, en la figura 2.2 se incluye un scatter matrix de un conjunto de variables ordinales obtenido por K-means ( $k = 5$ ) en mujeres entre 20 y 40 años. Se recomienda que sobre estas visualizaciones se construyan tablas que

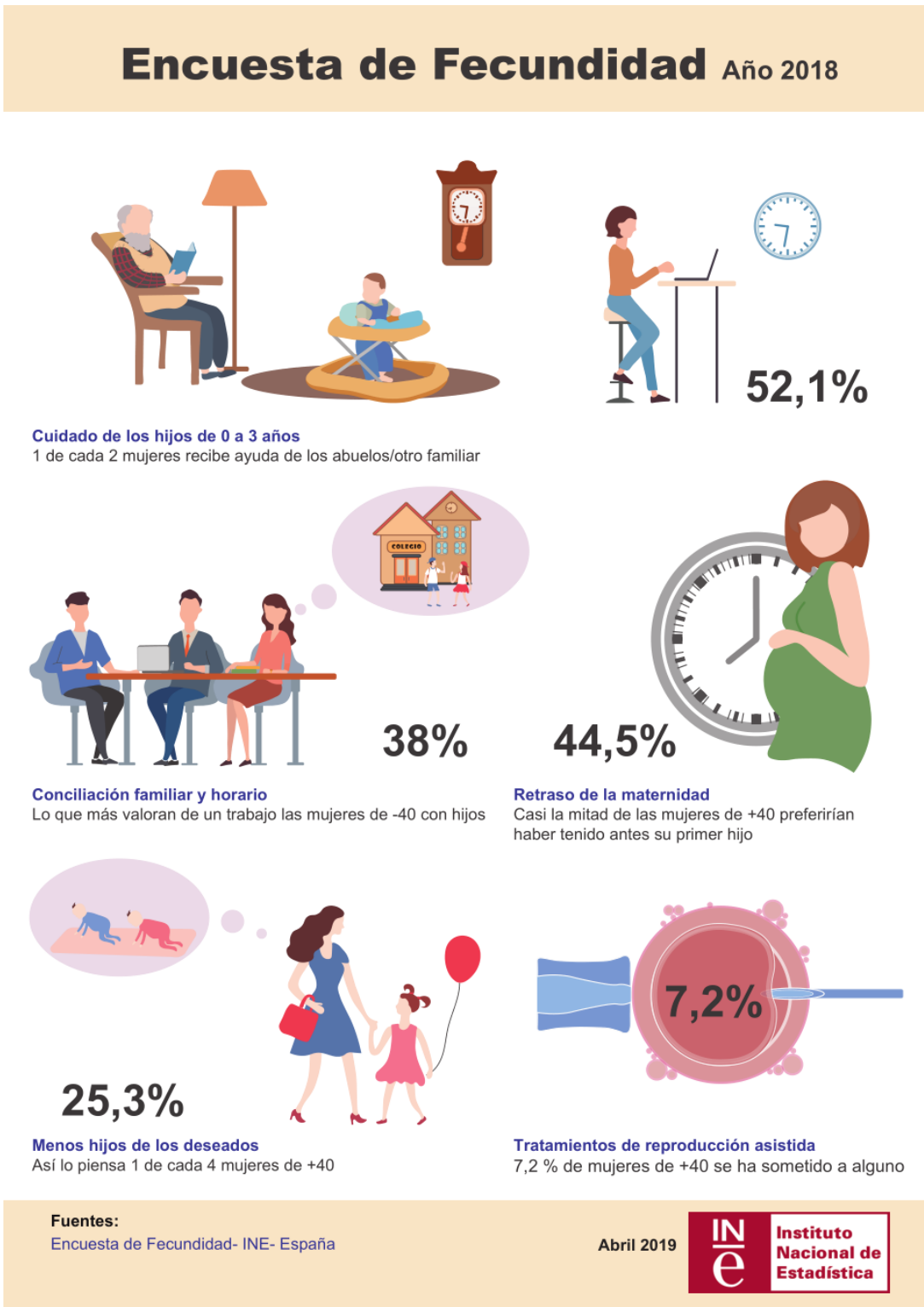


Figura 2.1: Resumen del INE sobre los resultados de la encuesta

caractericen aproximadamente cada grupo observando las agrupaciones realizadas. También pueden generarse gráficas de los centroides como la de la figura 2.3 para ayudar a interpretar el significado de cada grupo. En la web de la asignatura se incluyen *scripts* de ejemplo que pueden servir como punto de partida para realizar la práctica.

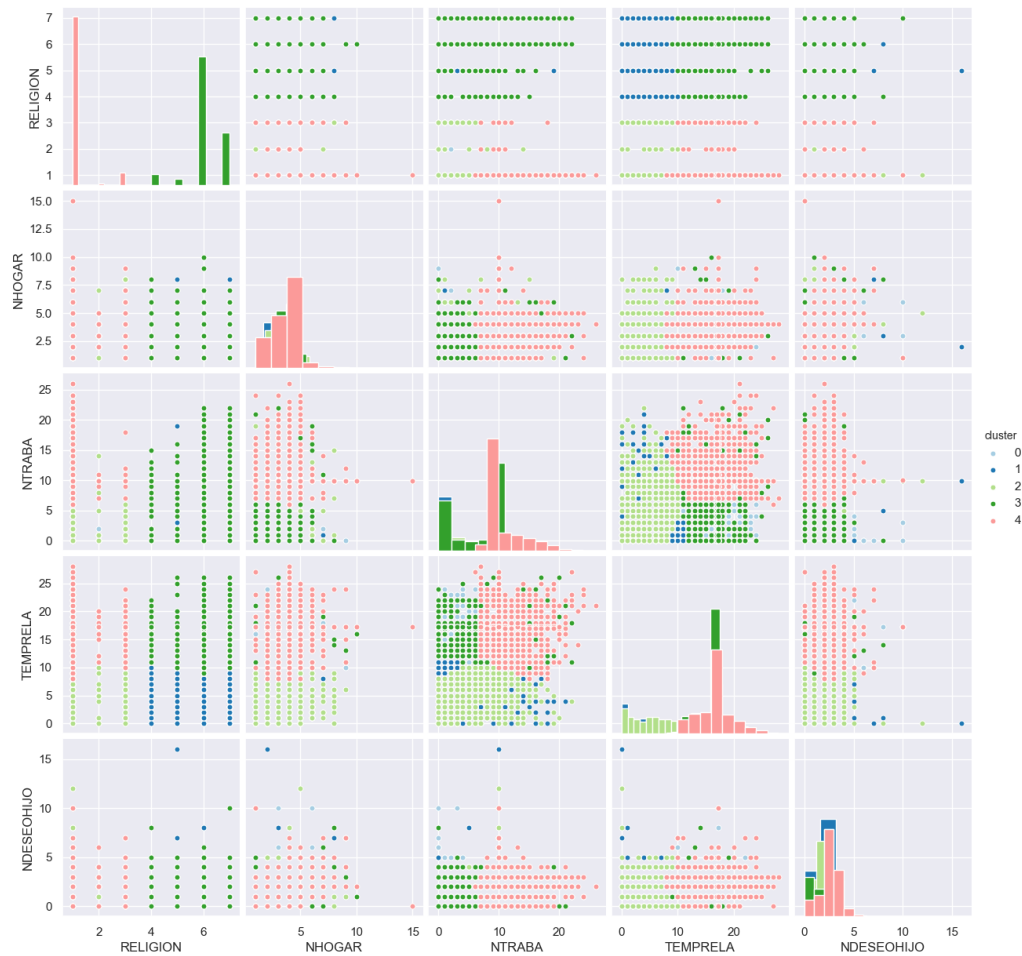


Figura 2.2: Ejemplo de resultado de K-means con  $k = 5$  relacionando variables como religión, número de personas en el hogar, duración en el trabajo actual, duración de pareja y número de hijos deseados

A partir de los resultados obtenidos se deberán extraer conclusiones sobre los grupos de población. Se valorará el acierto en la selección de casos de estudio que mejor reflejen los grupos encontrados en los datos.

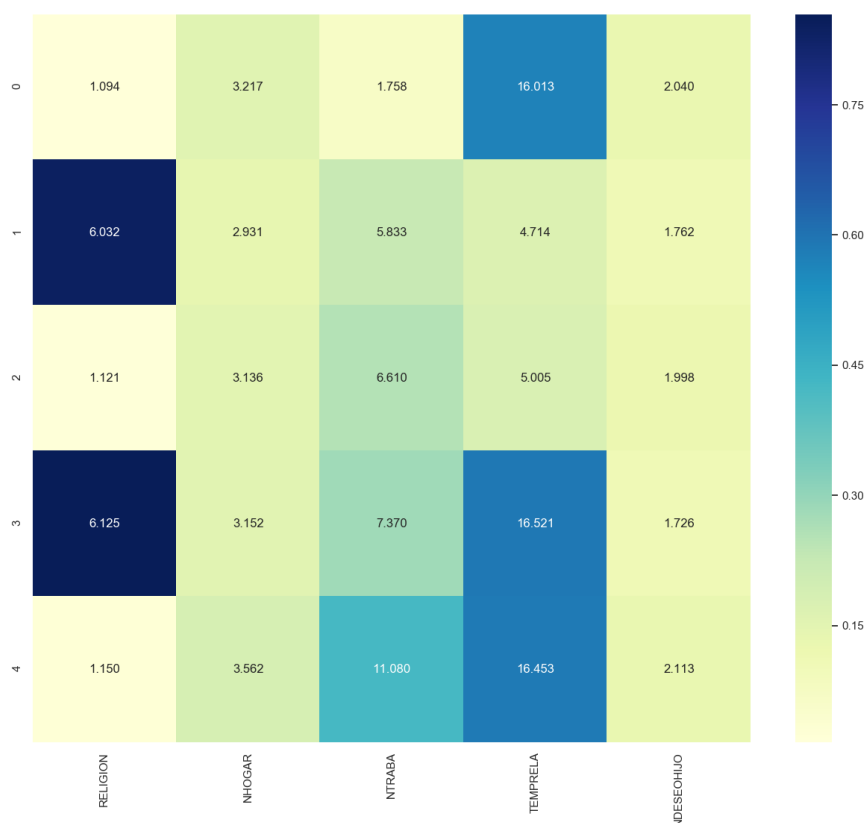


Figura 2.3: Centros de los grupos de la figura 2.2

## 4. Esquema de la Documentación

La documentación entregada deberá ajustarse al siguiente esquema (debe respetarse la numeración y nombre de las secciones):

1. **Introducción:** se hablará sobre el problema abordado y todas las consideraciones generales que se deseen indicar.
2. **Caso de estudio X:** se incluirá una sección por cada caso de estudio analizado (el epígrafe describirá el subconjunto de datos bajo estudio). En ella se explicará en detalle en un primer apartado qué caso se analiza y por qué (deberá indicarse el número de datos que representa el caso de estudio). Se incluirá una tabla comparativa con los resultados de los algoritmos de *clustering* (que incluirá, al menos, el número de *clusters* obtenidos, el valor de las métricas y el tiempo de ejecución en cada algoritmo) y tantas otras tablas para el análisis de los parámetros (una tabla por algoritmo). Cada sección contendrá

las visualizaciones necesarias para analizar el problema y junto a cada visualización se incluirá una tabla que caracterice cada *cluster*. Se añadirá un apartado final titulado “Interpretación de la segmentación” que incluirá las conclusiones generales a las que haya llegado el alumno a la luz de los resultados en el correspondiente caso de estudio. En cada sección deberán incluirse extractos de los *scripts* que el alumno considere relevantes para destacar el trabajo realizado.

3. **Contenido adicional:** opcionalmente, cualquier tarea adicional a las descritas en este guion puede presentarse en esta sección.
4. **Bibliografía:** referencias y material consultado para la realización de la práctica.

Las tablas de resultados no deberán ser capturas de pantalla, sino tablas creadas en el procesador de texto empleado. No se aceptarán otras secciones distintas de estas. Además, la primera página de la documentación incluirá una portada con el nombre completo del alumno, grupo de prácticas y dirección email. También se incluirá una segunda página con el índice del documento donde las diferentes secciones y páginas estarán enlazadas en el pdf.

## 5. Entrega

La fecha límite de entrega será el lunes **2 de diciembre** de 2019 hasta las **23:59**. La entrega se realizará a través de la web de la asignatura en <https://decsai.ugr.es/intranet>. En un único fichero zip se incluirá la documentación, los *scripts* de Python empleados y cualquier otro archivo que el alumno considere relevante. El nombre del archivo zip será el siguiente (sin espacios): P2-apellido1-apellido2-nombre.zip. La documentación tendrá el mismo nombre pero con extensión pdf. Es decir, la alumna “María Teresa del Castillo Gómez” subirá el archivo P2-delCastillo-Gómez-MaríaTeresa.zip que contendrá, entre otros, el archivo P2-delCastillo-Gómez-MaríaTeresa.pdf.

Si, por alguna razón justificada, el archivo zip fuera demasiado grande para subirse en la plataforma, se podrá incluir en la documentación (visible en la primera página de portada) un enlace a <http://consigna.ugr.es> con el material adicional. En tal caso, la permanencia en consigna deberá ser de 1 mes y deberá comunicarse por email al profesor de prácticas. En ningún caso se aceptan otros tipos de enlaces como Dropbox, Google Drive o similares.