Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

# Análisis de los factores asociados a la elección de estudios universitarios utilizando técnicas de agrupamiento

Trabajo Fin de Máster

**Máster Universitario en Gestión de la Información**

**Autor**: Jose Manuel Pérez Barberá

**Tutores**: Laura Sebastiá y Lenin Lemus

2016-2017

# Resumen

El momento de la elección de estudios universitarios es uno de los más importantes en la vida del estudiante. En este trabajo se analizan los factores asociados a la elección de estudios universitarios aplicando técnicas de agrupamiento para identificar qué perfiles de alumnos existen y cómo influyen aspectos como la empleabilidad, vocación o entorno a la hora de hacer la elección de titulación.

**Palabras clave:** agrupamiento, perfiles, elección, estudios, empleabilidad, vocación, prestigio.

# Abstract

Choosing a University degree is one of the most important moments in the student's life. This paper analyzes the associated factors with this choice applying clustering techniques to identify which student profiles exists and how aspects such as employability, vocation or environment influence in this moment.

**Keywords:** clustering, profiles, choice, studies, employability, vocation, prestige.

# Índice

# 1. Introducción

## 1.1. Introducción

El momento de la elección de estudios universitarios es uno de los más importantes en la vida del estudiante. Decidir qué estudiar y en qué profesión quieres desarrollar tu carrera laboral es una decisión difícil y en la que influyen multitud de factores. Ya en la etapa anterior, bien en bachillerato o bien en ciclos formativos de grado superior, el estudiante ha ido definiendo para qué vida laboral quiere prepararse, pero la selección de titulación universitaria es lo que marcará el futuro formativo próximo del estudiante y qué futuro laboral quiere tener.

El objetivo de este trabajo es analizar qué factores influyen en la decisión de los estudiantes, estudiar qué perfiles de alumnos existen con respecto a la influencia de unos factores u otros y ver dónde habría que centrar los esfuerzos de promoción, orientación y recomendación para que los alumnos elijan una titulación u otra.

Para realizar el análisis nos hemos basado en la tesis doctoral *"Un modelo estructural para el análisis de los factores asociados a la elección de estudios universitarios"* (Febrero de 2016) de Pedro Pablo Soriano [1] donde se definen qué factores intervienen en el momento de la selección de estudios universitarios.

Los factores estudiados en la tesis son:

- Percepción de la oferta y de la demanda
- Localización geográfica
- Entorno escolar
- Entorno próximo
- Aspectos vocacionales
- Consideración social
- Empleabilidad.

Como parte de la tesis, se define un cuestionario que recoge la información para determinar la influencia de dichos factores y se realiza una encuesta a más de 5.000 alumnos de universidades españolas.

Actualmente, en el trabajo final de máster "Determinant factors associated with the choice of university degrees in Spain through Principal Component and Factor analysis" de Regina Sandra Mollaplaza se está trabajando sobre el

cuestionario definido en la tesis anterior para intentar hacer una reducción del número de variables aplicando técnicas como ACP (Análisis de Componentes Principales) y AF (Análisis Factorial).

Sobre estos datos, cuestionario, resultados y preguntas significativas, aplicaremos técnicas de análisis de datos para entender la influencia de los factores mencionados sobre los alumnos. En concreto, utilizaremos técnicas de agrupamiento buscando estructurar los datos según su similitud y así poder predecir a qué grupo pertenecería un nuevo elemento.

## 1.2. Motivación

En la elección de este trabajo, se nos presentó la tesis doctoral mencionada anteriormente. En ella, mediante técnicas de análisis matemático, se propone un método para realizar un análisis del sistema de acceso universitario y se desarrolla un cuestionario con las variables del modelo. Asimismo, se utiliza dicho cuestionario para recabar información de gran número de alumnos de nuevo ingreso.

Nuestro trabajo se basa en utilizar los datos de las encuestas para realizar un análisis desde el punto de vista de los alumnos, es decir, qué perfiles de alumnos existen y cómo podemos caracterizarlos teniendo en cuenta las respuestas a las preguntas de la encuesta. Todo ello desde el punto de vista de qué factores influyen o no a la hora de decidir qué titulación universitaria quieren cursar.

Este cuestionario se realizó en varias universidades españolas obteniendo casi 5.000 encuestas válidas, lo que supone disponer de una importante base de información que nos permite centrar el trabajo en el análisis. Con esto, ya podemos analizar qué técnicas y herramientas estadísticas son más adecuadas para sacar el mayor partido a los datos y poder obtener conclusiones en base a éstos.

Un aspecto importante para poder aplicar técnicas de agrupamiento con éxito es reducir el número de variables del estudio, para ello utilizaremos los resultados del trabajo de Regina Sandra Mollaplaza que identifican las preguntas significativas asociadas a cada factor.
En dicho trabajo, que parte de las 34 preguntas mencionadas anteriormente, aplicando técnicas como Análisis de componentes Principales o Análisis Factorial, se consigue reducir el número de preguntas a 12, las cuales a partir de ahora nombraremos como *“preguntas significativas”*.

## 1.3. Objetivos y metodología

El objetivo de este trabajo es realizar un análisis de los factores asociados a la elección de estudios universitarios aplicando técnicas de agrupamiento que permita identificar qué perfiles de alumnos existen y cómo influyen aspectos como la empleabilidad, vocación o entorno a la hora de hacer la elección de titulación.

Se trata de un análisis básicamente descriptivo e interpretativo para comprender qué influye a los alumnos en dicha decisión.

Para llevar a cabo este análisis partiremos de los resultados de las encuestas realizadas en la tesis de Pedro Pablo Soriano [1].

En primer lugar, se realizará una limpieza de los datos para descartar las observaciones que no nos sirvan para el análisis y homogeneizar los datos para facilitar el proceso de análisis e interpretación de los resultados.

Después de la limpieza, y antes de empezar con el análisis, exploraremos los datos para ver de qué información disponemos y ver qué posibles análisis podemos hacer con ellos. Para ello, mediante estadística descriptiva, obtendremos un resumen de las variables que disponemos para el total de observaciones.

Partiendo de los datos limpios haremos un primer análisis para ver qué perfiles existen en los datos. La base para aplicar técnicas de agrupamiento es determinar por qué variables queremos agrupar y para ello utilizaremos las variables correspondientes a las preguntas significativas de cada factor, extraídas del trabajo de Regina Sandra Mollaplaza. En base a las preguntas significativas aplicaremos técnicas de agrupamiento por esas variables para obtener los distintos grupos existentes en el conjunto total de datos. Una vez tengamos los grupos, añadiremos el resto de variables observables independientes de los factores, como por ejemplo nota de acceso, y obtendremos un resumen para caracterizar cada uno de los grupos e identificar los perfiles existentes.

El siguiente paso será aplicar esas mismas técnicas de agrupamiento enfrentando un factor con otro y estudiar qué grupos obtenemos y qué características comunes tienen los alumnos de esos grupos.

A continuación, separaremos las observaciones por ramas y repetiremos el análisis para cada rama y así ver si hay cambios en los grupos dependiendo de la rama analizada.

Además, para cada rama, realizaremos un resumen estadístico de las observaciones agrupadas por titulación para ver las diferencias entre alumnos de distintas titulaciones.

Cada uno de los grupos obtenidos en el análisis se caracterizará mediante estadística descriptiva para intentar describir cada uno de los grupos obtenidos.

# 2. Estado del arte

## 2.1. Elección de estudios universitarios

### 2.1.1. Tesis

Como hemos comentado anteriormente, para el análisis que vamos a realizar en este trabajo se utilizarán los resultados de la Tesis doctoral "Un modelo estructural para el análisis de los factores asociados a la elección de estudios universitarios" de Pedro Pablo Soriano Jiménez (2016) [1]. En este punto vamos a profundizar sobre el trabajo realizado.

El objetivo de la tesis es:

*"Conocer las razones que mueven a un estudiante (o a su familia) a solicitar una titulación en una universidad para cursar sus estudios superiores para poder diseñar acciones que pudieran optimizar la demanda y con ello el rendimiento académico en el sistema universitario".*

En el proceso para lograr el objetivo definido se realizan 5 grandes pasos:

1. Se analiza el proceso de elección de estudios y se identifican las variables que participan en la elección.
2. Se desarrolla un modelo de ecuaciones estructurales en el que aparecen los factores que intervienen en la elección y las relaciones que se dan entre ellos.
3. Se elabora un cuestionario con las preguntas que recogerán la información de las variables identificadas en el primer paso.
4. Se valida el modelo y el cuestionario.
5. Se pasa la encuesta y se analizan los resultados.

Como resultado del punto 1, se detectaron los siguientes factores que, posteriormente, estudia la encuesta.

- Factores internos: Oferta plazas y percepción de oferta/demanda
- Factores externos
    - Factores individuales: Vía de acceso, género, nota de acceso, aspectos vocacionales, influencia del entorno próximo y localización geográfica
    - Factores sociales: Consideración social de la Universidad, empleabilidad percibida y consideración social de la titulación.

La encuesta resultante del trabajo se pasó en 31 universidades del Sistema Universitario Público Español (SUPE), obteniendo un total de 5111 resultados válidos.

A continuación se presentan los resultados agrupados por ramas para cada factor analizado. En el gráfico de la izquierda cada línea representa la distribución de las respuestas agrupadas del factor en cuestión, y en el histograma se muestra la suma media de respuestas.

*AH: Artes y Humanidades.* *CS: Ciencias de la Salud.* *SJ: C. Sociales y Jurídicas*
*C: Ciencias.* *IA: Ing. y Arquitectura.*

<u>Factores Sociales (11 preguntas)</u>

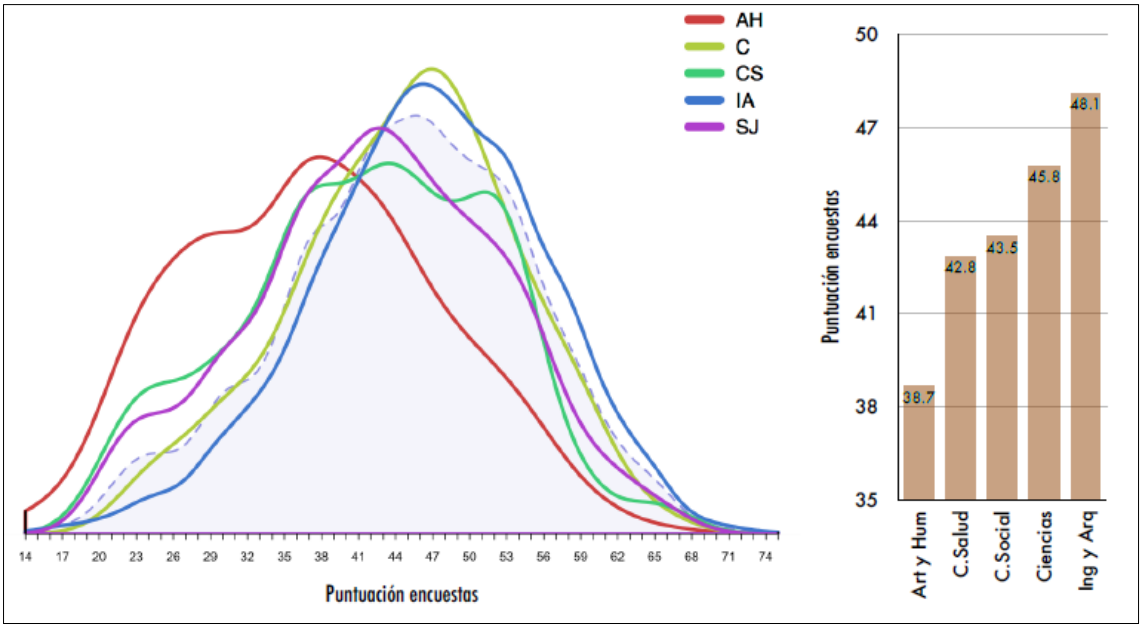


**Figura 1**

En la figura 1 observamos cómo la influencia de los factores sociales entre las ramas de Ingeniería y Arquitectura y la rama de Artes y Humanidades difieren en casi 10 puntos y como el resto de ramas tienen distribuciones y valores medios similares.

Factores Sociales: Consideración social (3 preguntas)

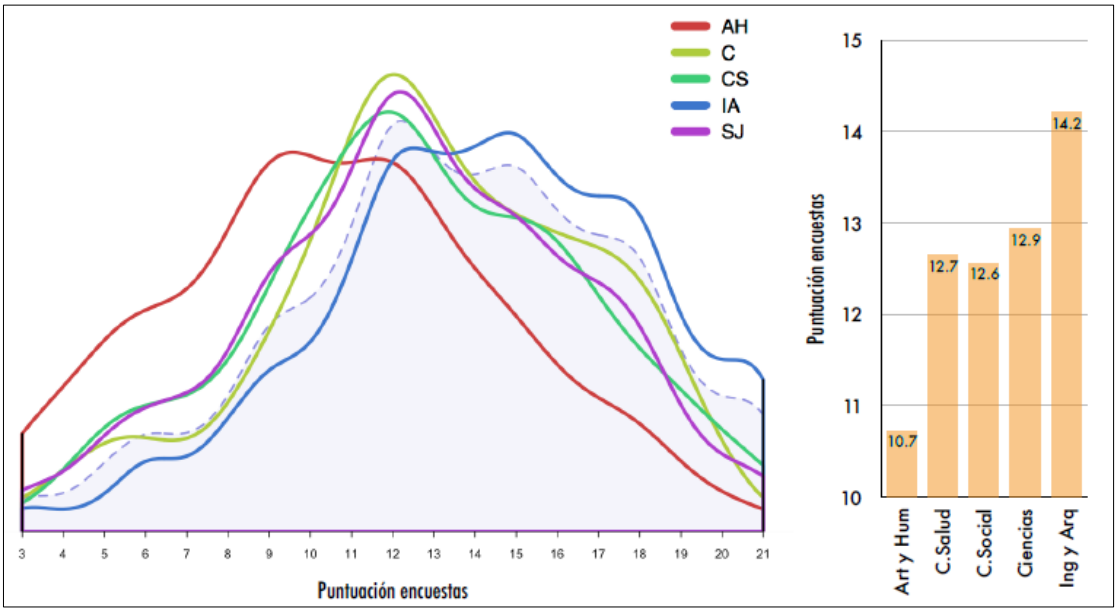


**Figura 2**

Factores Sociales: Consideración Universidad (5 preguntas)

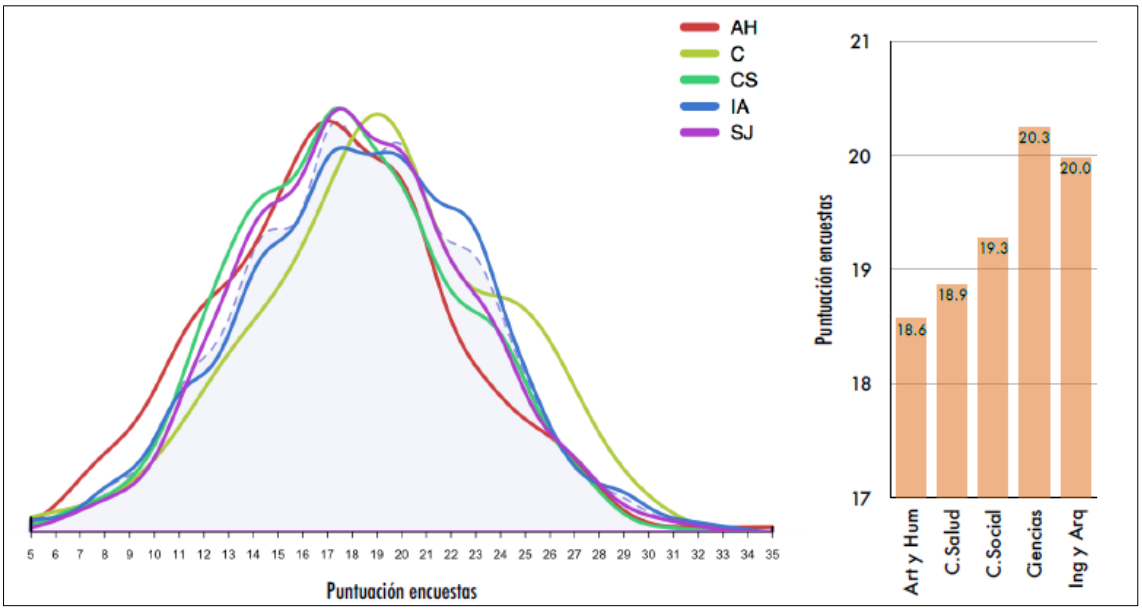


**Figura 3**

Factores Sociales: Utilidad Percibida (3 preguntas)

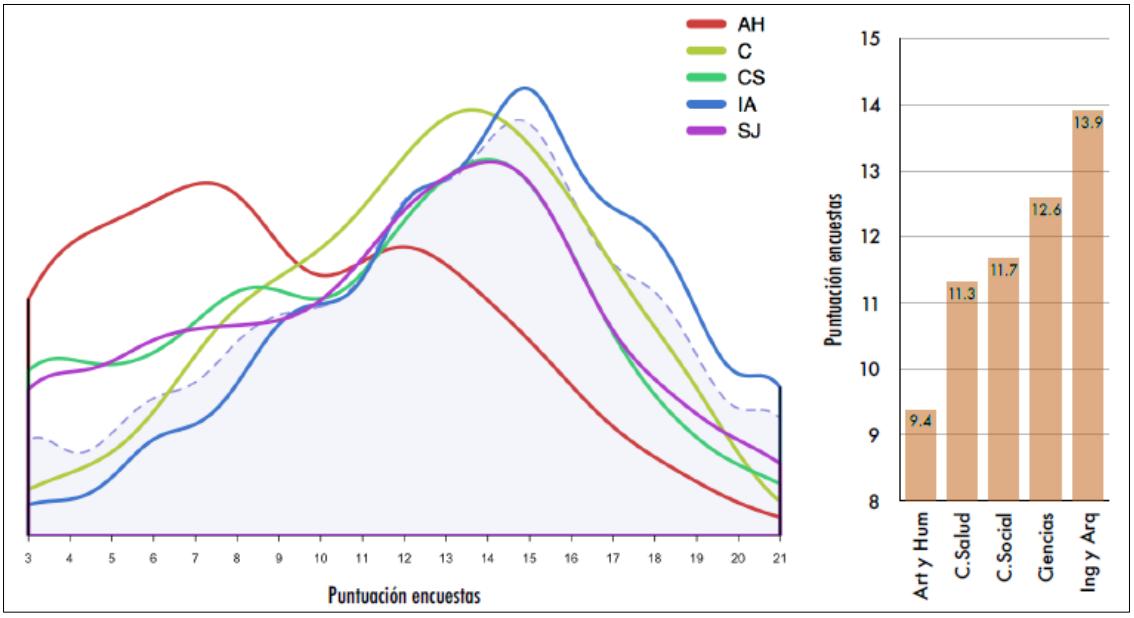


**Figura 4**

Como vemos en las figuras 2, 3 y 4, si entramos un poco más en detalle y analizamos cada factor social de forma individual, vemos que la tendencia observada para todos los factores se repite para cada factor.

Sobretodo llaman la atención los resultados de los factores Consideración social y Utilidad percibida, donde se observan diferencias claras entre ramas.

Factores Individuales (14 preguntas)

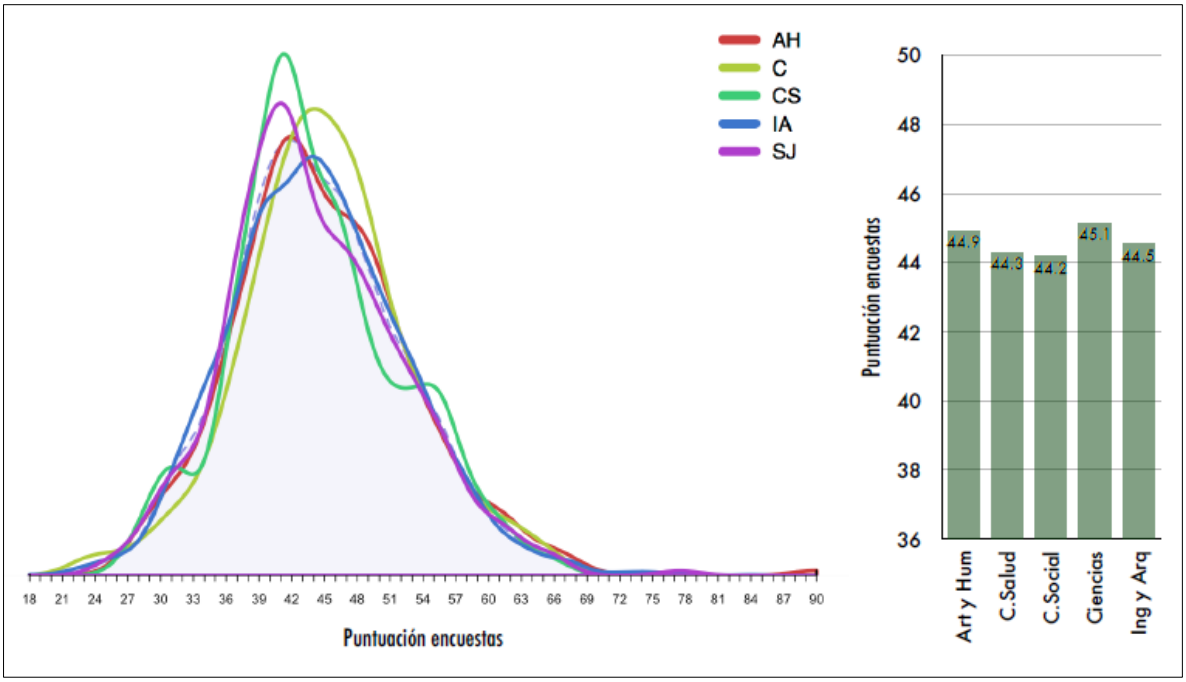


**Figura 5**

En cuanto a los factores individuales, vemos que en el estudio agrupado, figura 5, no hay prácticamente diferencias entre las ramas de estudio. Es en el análisis por cada factor donde sí que encontramos mayores diferencias.

Factores Individuales: Aspectos Vocacionales (4 preguntas)

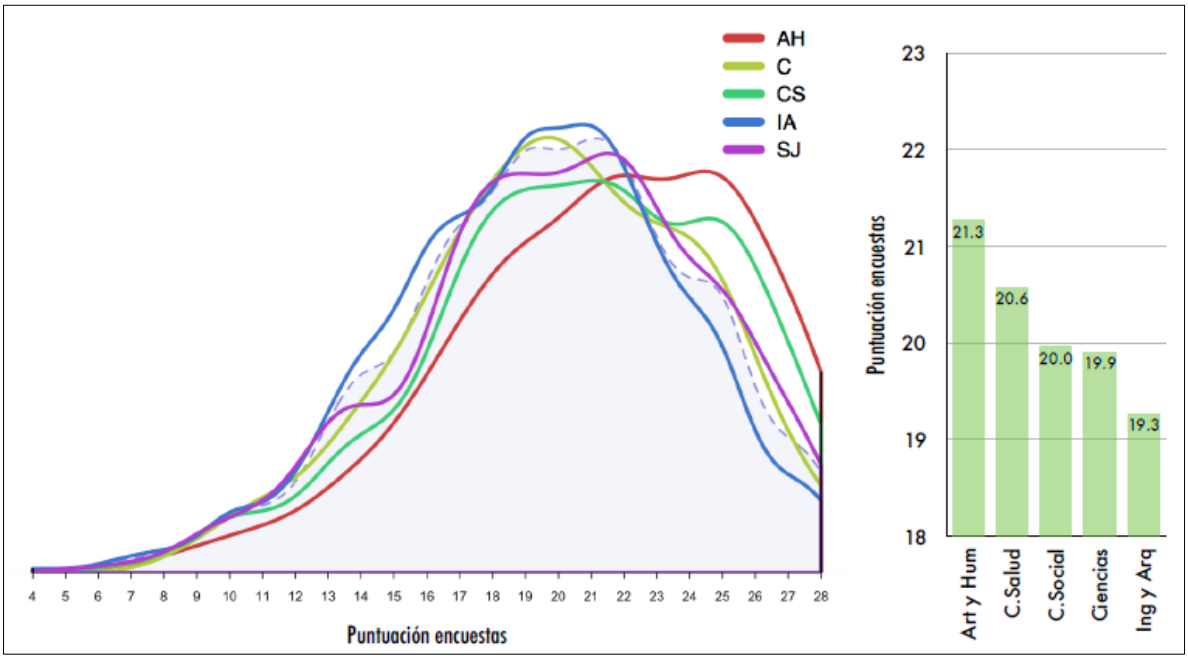


**Figura 6**

En esta figura se observa que los Aspectos Vocacionales sí que tienen una influencia importante según los datos. Vemos además que sí que existe cierta diferencia entre ramas en este aspecto.

Factores Individuales: Influencia Entorno (8 preguntas)

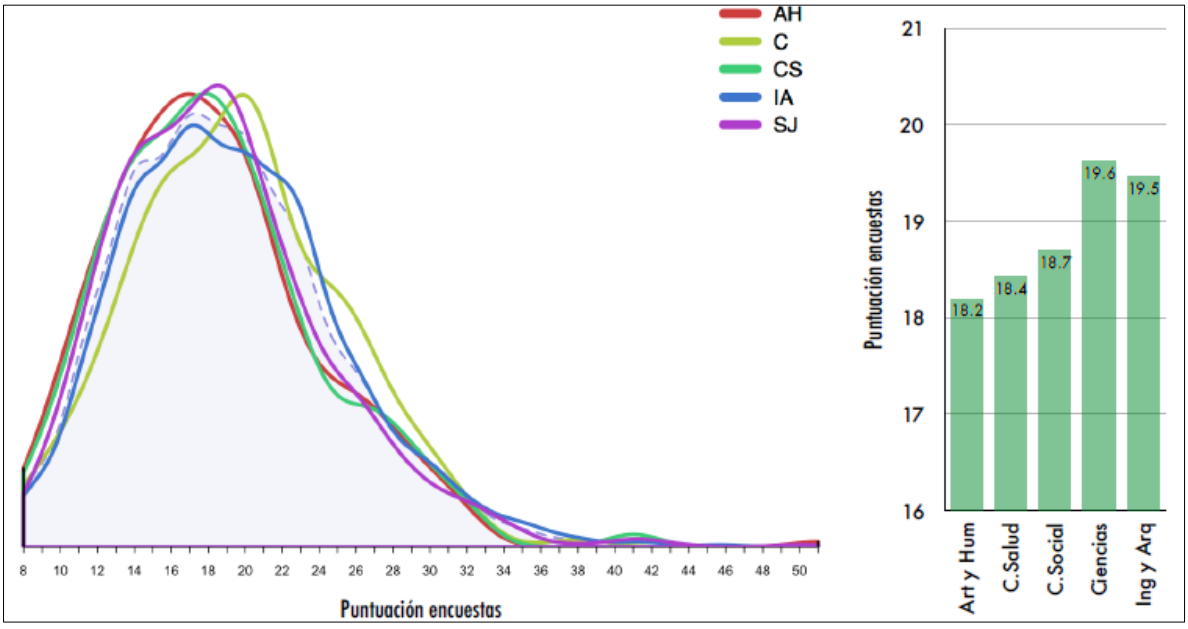


**Figura 7**

Factores Individuales: Localización Geográfica (2 preguntas)

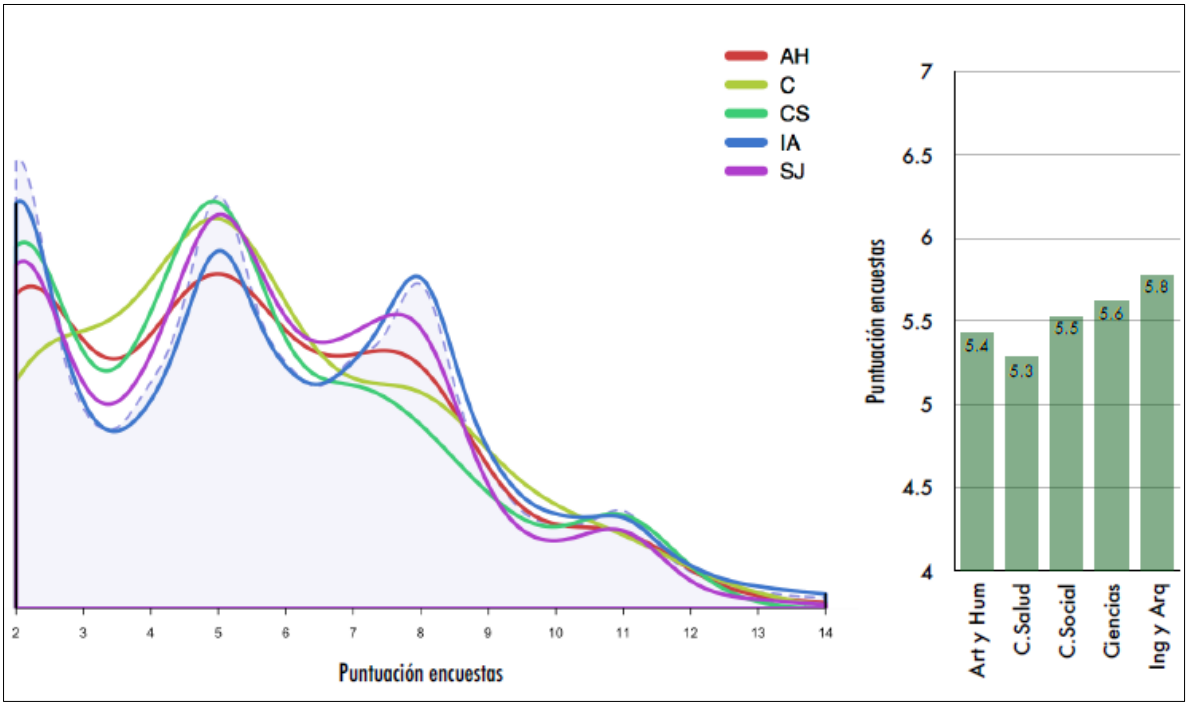


**Figura 8**

En el resto de factores individuales, influencia del entorno y localización geográfica las diferencias entre ramas se reducen.

En vista de los gráficos de las figuras 7 y 8, podemos decir que sí que existen diferencias entre ramas de estudio, principalmente en los factores relacionados con la consideración social de la titulación, la utilidad percibida (empleabilidad) y aspectos vocacionales.

Analizando estos resultados podemos interpretar la influencia de cada factor sobre el total de alumnos o por ramas, pero el resultado puede ser engañoso, no sabemos qué tipos distintos de alumnos existen ya que al hacer el sumatorio de todas las preguntas de cada factor perdemos la caracterización de cada alumno en particular. En dos alumnos con una suma de respuestas 8 podrían haber 4 combinaciones de respuestas distintas, 7-1, 6-2, 5-3 y 4-4.

Aquí es donde entra nuestro análisis para agrupar los alumnos por sus respuestas según la influencia de un factor frente a otro.

### 2.1.2. Estudios anteriores

En este punto vamos a presentar algunos estudios previos relacionados con la elección de estudios universitarios y los factores asociados a dicha elección.

Aunque hay diversos estudios al respecto, la mayoría de los autores han trabajado desde puntos de vista y contextos concretos. Buena parte de las investigaciones y artículos que hemos encontrado centran sus análisis en diferenciar los factores asociados a la elección según el género del estudiante.

Para nuestro trabajo, además de la Tesis de Pedro Pablo Soriano, hemos tomado como antecedentes 4 estudios realizados entre los años 2000 y 2012.

En el primer trabajo (Mosteiro García, Mª Josefa y Porto Castro, Ana M. Año 2000) [2] se estudian los motivos más influyentes en la elección de estudios de los estudiantes universitarios y se comprueba si éstos difieren según el género y el tipo de carrera elegida. El estudio concluye que los estudiantes aducen motivos similares a la hora de la elección de estudios universitarios y que las diferencias por género no son estadísticamente significativas. No obstante, analizando los datos por tipo de titulación los resultados sugieren que sí que existen diferentes perfiles entre alumnos de distintas titulaciones.

El trabajo concluye que los motivos más influyentes en el momento de la selección son: "Porque tengo aptitudes para ello" con un 76.2%, "Porque me gusta mucho" con un 74.4% y "Hay más posibilidades de encontrar trabajo" con un 52.4%. Como vemos, motivos relacionados con aspectos vocacionales y empleabilidad. Por otro lado, el estudio concluye que "Por tradición familiar" con un 80.9% y "Tiene más salidas para personas de mi sexo" con un 77.5% son los motivos menos influyentes.

En el segundo estudio tomado como antecedente (Sánchez García, Mª Fe. Año 2001) [3] se contemplan los motivos influyentes clasificados en aquellos con carácter personal, los de carácter profesional, y los denominados influencias de tipo sociofamiliar y se realiza un análisis por ramas de estudios. En el análisis general vemos de nuevo como aspectos relacionados con la vocación del estudiante y con la empleabilidad son los que más influyen en la elección. En el análisis por ramas observamos también que entre ramas de estudios la influencia o no de determinados factores varían siendo también los más significativos los relacionados con la vocación y la empleabilidad.

Como tercera referencia tomamos un estudio (González López, Ignacio. Año 2005) [4] donde se realiza un análisis de datos para constatar los motivos que llevan al alumnado a elegir estudios universitarios, en él se destaca la existencia prioritaria de preferencias personales y profesionales. El estudio concluye que los motivos principales que guían al alumno a elegir una titulación u otra son

“Me gusta” y “Por la profesión”, relacionados con aspectos vocacionales y “Oportunidades de empleo” relacionado con la empleabilidad.

Atendiendo al género de los estudiantes, no se encuentran variaciones significativas. Y en cuanto a la rama de especialización, en todas las ramas el motivo principal y único estadísticamente significativo es “Me gusta”.

En último y más reciente trabajo que hemos tomado como antecedente (Navarro Guzmán, C. y Casero Martínez, A. Año 2012) [5] se estudian los motivos de elección según el género de los alumnos. Los resultados muestran que las chicas eligen principalmente carreras de Humanidades, Ciencias Experimentales, Ciencias Sociales, Jurídicas y de la Salud y sus principales motivos de elección son porque les gusta, por vocación y para ayudar a otras personas, mientras que ellos se decantan por estudios técnicos para ganar un buen sueldo.

En los resultados por sexo y rama del estudio los motivos de elección más indicados por los encuestados no varían significativamente entre ramas de estudio y son los mismos que se ven en los resultados globales. Con esto, podemos decir que los motivos principales están relacionados en gran medida con aspectos relacionados con la vocación y empleabilidad y en menor grado con cuestiones como ayudar a otras personas, completar la formación de los estudiantes y tener un buen futuro.

Aparte de estos 4 estudios que tomamos como referencia del nuestro, hemos revisado otros estudios donde se analizan los motivos y objetivos de los estudiantes a la hora de elegir varias titulaciones concretas, enfrentando por ejemplo a los estudiantes de psicología, derecho y biología (Gámez, Elena y Marrero, Hipólito - 2003) [6] o estudian entornos determinados (Solsona Solé, Joan; Gómez Adillón, Mª Jesús y Saladrigues Solé, Ramón - 2016) [7] lo que reduce el análisis a determinados factores. En el caso de este último trabajo, se centran en factores como la proximidad geográfica, la reputación de la universidad o la facilidad de acceso.

También hay estudios (Valle, A.; González, R.; Cuevas, L.M - 1997) [8] que realizan el análisis con otro enfoque, desde el punto de vista de las metas de aprendizaje, logro o refuerzo social que mueven a los alumnos a elegir una titulación u otra y no hablan de factores propiamente dichos por lo que los hemos descartado como base para nuestro trabajo.

## 2.2. Análisis de datos

En este punto vamos a presentar los conceptos estadísticos y de minería de datos que vamos a utilizar en nuestro análisis.

### 2.2.1. Estadística descriptiva [9]

Esta rama de la estadística es la que se ocupa de convertir los datos en valores y gráficos fácilmente entendibles para nosotros y así poder interpretarlos de una manera rápida y eficaz. Describe los datos mediante gráficos y valores que los resuman.

A lo largo del análisis utilizaremos gráficos como el histograma, el boxplot o el diagrama de densidades, con los que estudiaremos las respuestas a las preguntas para ver su distribución y poder comparar preguntas teóricamente similares.

Una vez obtengamos los grupos, para compararlos utilizaremos un resumen numérico que nos permitirá describirlos. Los cálculos que utilizaremos para caracterizar cada grupo serán el valor central y la dispersión de cada variable sobre la que se realiza el análisis.

Además, para ver la relación lineal de las preguntas entre sí utilizaremos una matriz de correlaciones.

Presentamos cada uno de los conceptos mencionados:

***Resumen numérico: Mediana y rango intercuartílico***

Para elegir qué medidas utilizar para caracterizar nuestros datos es necesario estudiar qué tipos de datos tenemos. En nuestro caso los datos de las respuestas de la encuesta están en formato escala de 1 a 7, según el nivel de conformidad de los alumnos con la cuestión presentada. Más adelante presentaremos qué ventajas y desventajas tiene esta tipología de respuesta.

Atendiendo a la tipología de datos mencionada, y según varios artículos y estudios referenciados en la bibliografía [10] [11], para estas variables la medida de centralidad más idónea es la **mediana** y como medida de dispersión utilizaremos el **rango intercuartílico**.

Para las variables observables, como nota de acceso o nota de corte, variables cuantitativas, utilizaremos la **media** y la **desviación típica** como medidas de centralidad y dispersión.

A modo de ejemplo mostramos el resumen numérico resultado de aplicar la técnica de agrupamiento por las preguntas relacionadas con el factor vocacional y con el factor de empleabilidad sobre el total de observaciones añadiendo las

variables observables nota de acceso y nota de corte. En la cabecera de todas las tablas indicaremos la medida utilizada para cada variable.

**Tabla 1**

| Grupo | N. de observaciones | Mediana(iqr) Siempre he querido estudiar esta carrera | Mediana(iqr) Es más fácil encontrar trabajo | Media(dt) Nota Acceso | Media(dt) Nota Corte |
|---|---|---|---|---|---|
| **1** | 1487 | 5 (2) | 6 (1) | 9.48 (2.06) | 7.26 (2.06) |
| **2** | 865 | 2 (2) | 3 (3) | 7.74 (1.72) | 6.5 (1.57) |
| **3** | 2563 | 1 (2) | 6 (2) | 9.05 (1.99) | 6.89 (1.81) |

## *Boxplot*

Es un gráfico mediante el cual se visualiza la distribución de una variable numérica, suministra información sobre los valores mínimo y máximo, los cuartiles, y sobre la existencia de valores atípicos (outliers) y la simetría de la distribución.

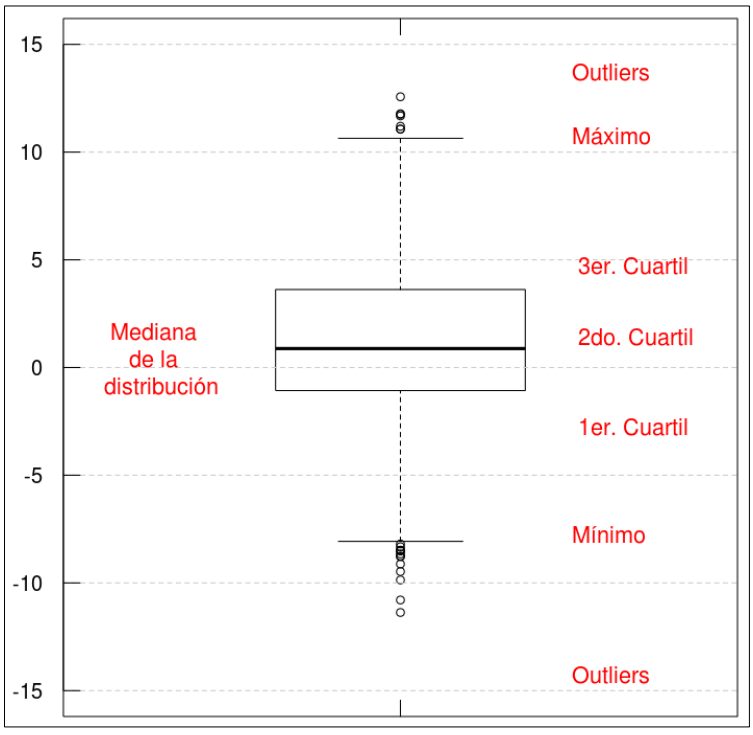


**Figura 9**

### *Histograma*

Un histograma resume una variable de un modo sencillo y eficaz utilizando las tablas de frecuencias. Es un diagrama de barras y la altura de cada barra es la frecuencia de cada variable. Una de las ventajas del histograma es que te permite entender qué es una función de densidad de probabilidad, que explicamos en el siguiente punto.

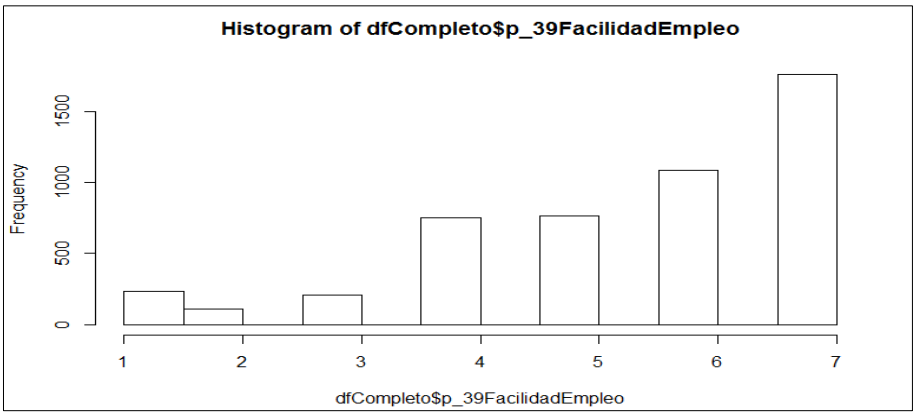


**Figura 10**

### *Gráficos de densidad*

Es un diagrama que representa una función de densidad que caracteriza el comportamiento probable de una población.

Una ventaja que los gráficos de densidad tienen sobre los histogramas es que son mejores en la determinación de la forma de la distribución, al no verse afectados por el número de contenedores o barras usadas en los histogramas. Por ejemplo, en un histograma con 4 barras no veríamos una curva lo suficientemente clara como en uno de 20 barras. En los gráficos de densidad no tenemos ese problema.

En el siguiente ejemplo vemos el diagrama de densidad de una de las preguntas.

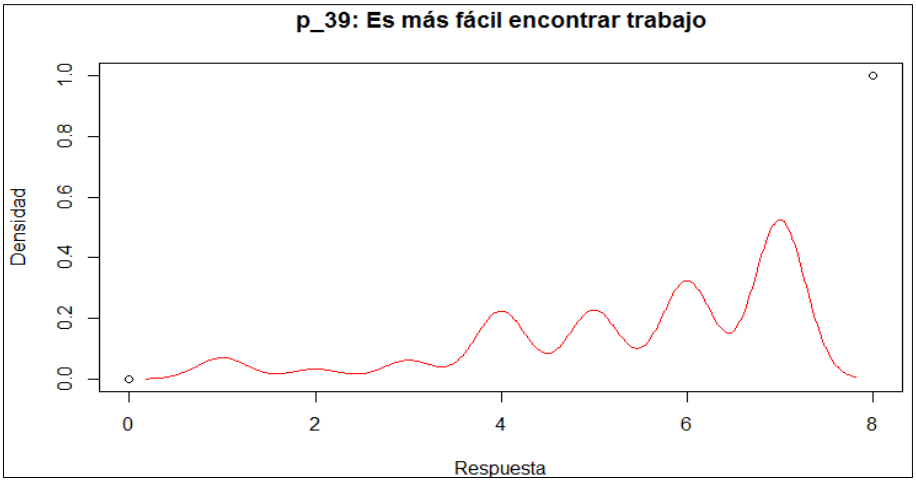


**Figura 11**

### *Matriz de correlación*

La correlación indica el grado de relación lineal entre variables. El valor de la correlación siempre está entre -1 y 1, negativo indica relación inversa. Una correlación de 0 o cercana a 0 indica que las variables no están relacionadas de forma lineal, pero no significa que no estén relacionadas de otra forma.

A modo de ejemplo se muestra la matriz de correlación entre dos preguntas.

**Tabla 2**

| | Siempre he querido estudiar esta carrera | Es más fácil encontrar trabajo |
|---|---|---|
| Siempre he querido estudiar esta carrera | | .08 |
| Es más fácil encontrar trabajo | .08 | |

### 2.2.2. Aprendizaje automático (Machine Learning)

El aprendizaje automático [12] o Machine Learning un método científico que nos permite usar toda la capacidad computacional que tenemos en nuestros ordenadores y dispositivos para extraer patrones y relaciones que hay en los datos.

Los ingredientes imprescindibles para cualquier solución de aprendizaje automático son los datos y los modelos. A continuación presentamos los distintos modelos y algoritmos que existen entre los cuales el que hemos elegido para nuestro análisis.

#### *Aprendizaje supervisado*

Resuelve problemas de aprendizaje automático donde la máquina aprende de un conjunto de casos o instancias previamente clasificados por un experto, es decir entrena con datos preparados, bajo supervisión.

Los problemas de aprendizaje supervisado son básicamente la clasificación y la regresión, dependiendo de si el campo objetivo es categórico o numérico, respectivamente.

Un ejemplo de modelo de aprendizaje supervisado son los árboles de decisión. Los árboles de decisión son un tipo de modelo predictivo donde se utiliza un grafo con estructura de árbol para la clasificación de los datos. Cada nodo del árbol simboliza una pregunta y cada rama corresponde a una respuesta concreta a dicha pregunta. Los nodos terminales u hojas son aquellos donde el modelo ya ha clasificado los datos y el camino desde la raíz a la hora define las reglas de clasificación.

Los modelos de aprendizaje supervisado se usan para predecir un campo objetivo, y aprenden las reglas para predecirlo a partir de los casos o instancias usadas para el entrenamiento. La forma de evaluar un modelo es comparar los valores reales de las instancias donde se conoce el campo objetivo con los valores que predice el modelo. Un factor importantísimo es no usar nunca en el proceso de evaluación instancias usadas para el entrenamiento.

#### *Aprendizaje No supervisado*

Resuelve problemas de aprendizaje automático sin que se requiera ningún etiquetado o clasificación previa de los casos o instancias. Se basa en los datos y su objetivo es determinar relaciones de similitud, diferencia o asociación.

En aprendizaje no supervisado existen tres tipos de modelos:

- Clustering: Buscan instancias similares entre sí y distintas a las demás para formar agrupaciones de los datos llamadas clusters. Estos modelos nos permiten predecir a qué clúster corresponde un nuevo elemento.

- Detector de anomalías: Buscan instancias distintas de la tónica general del resto de datos.

- Buscadores de asociaciones: Buscan relaciones existentes entre diversos valores de los campos proporcionados. Las reglas de asociación son de la forma: cuando este campo tiene este valor entonces, en general, aquel campo tiene aquel valor.

El proceso hasta llegar a la solución final puede que conlleve el uso de más de un tipo de modelo. En nuestro caso, como hemos comentado, vamos a utilizar técnicas de clustering para buscar agrupaciones en nuestros datos.

### *Clustering*

El objetivo del clustering es estructurar los datos en grupos según su similitud. Cada grupo contendrá instancias de los datos u observaciones consideradas similares según los atributos sobre los que se realiza la agrupación. A su vez, los datos agrupados en un clúster tienen que ser claramente distintos a los de los demás clusters. El clustering busca maximizar la similitud intra-grupo y la diferencia inter-grupos.

Para determinar la similaridad de los datos se usa el concepto de distancia. En nuestro caso, la herramienta utilizada para realizar el análisis ofrece 4 métodos de cálculo de distancia, los examinaremos en el siguiente punto.

Según si conocemos el número de grupos o no podemos diferenciar entre dos tipos de clustering:

- Clustering jerárquico.
- Clustering no jerárquico.

### *Clustering Jerárquico*

En el clustering jerárquico no se conoce a priori el número de grupos buscados.

Los algoritmos de clustering jerárquico pueden ser de dos tipos:

- Aglomerativos: parten de tantos grupos como observaciones y van fusionando los grupos más similares formando clusters más grandes hasta llegar a formar un único clúster.

- Divisivos: Se parte de un solo grupo y se va particionando en distintos clústeres.

Normalmente estos métodos jerárquicos suelen ser ascendentes, es decir que sucesivamente van fusionando grupos desde el elemento individual hacia arriba.

El gráfico que resume el proceso de composición de los clústeres se denomina dendrograma.

Este gráfico ayuda a elegir el número de clusters óptimo que podrían representar mejor la estructura de los datos en base a su medida de similitud. "Cortando" el dendrograma por donde observamos un salto importante en la longitud de las líneas que definen los grupos podremos identificar el número de clusters.

A continuación en la figura 12 presentamos el dendrograma obtenido de aplicar clustering jerárquico a nuestros datos:

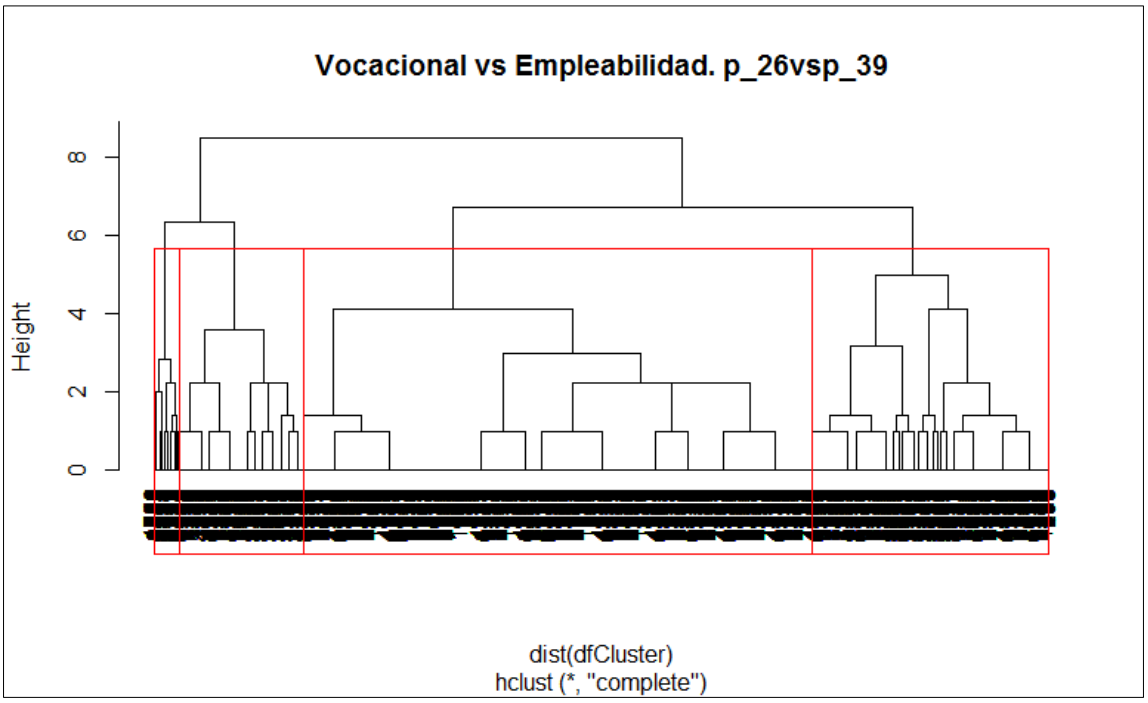


**Figura 12**

Un concepto importante al aplicar algoritmos de clustering es el Centroide, es el representante de cada grupo, los centroides son las medias de cada variable de cada clúster y ayudan a interpretar cada clúster.

Aplicar un algoritmo de clustering jerárquico puede ser el paso previo al clustering no jerárquico para poder determinar el número de clusters.

### *Clustering no jerárquico*

A diferencia del clustering jerárquico, conocemos el número de grupos a priori y las observaciones son asignadas a cada uno de los grupos.

El algoritmo más conocido de clustering no jerárquico es k-means. Este algoritmo lleva a cabo un proceso iterativo para encontrar k grupos que minimicen la suma de distancias de sus centroides.

Los pasos que se siguen en un algoritmo de clustering jerárquico son los siguientes:

1. Determinar los k centroides iniciales.
2. Asignar cada observación al grupo más cercano.
3. Calcular nuevos centroides.
4. Repetir el algoritmo hasta que no hayan cambios, entonces podremos decir que la solución es estable.

Como podemos intuir, este algoritmo es muy sensible al número inicial de centroides.

### *¿Clustering Jerárquico o No jerárquico?*

Elegir entre qué tipo de análisis utilizar no tiene que ser una decisión cerrada ya que ambas técnicas se complementan entre sí.

Cuando sepamos el número de grupos en los que se unen las observaciones aplicaremos análisis no jerárquico. Sin embargo, cuando no tenemos información previa, una de las opciones para identificar el número óptimo de grupos es realizar un análisis jerárquico.

Como hemos visto, determinar el número de clusters es una de las tareas más sensibles en el análisis clúster y otra solución para identificar el número de grupos es el método Elbow [13]. Este método implica estudiar un conjunto de posibles números de clusters en función de cómo ese número minimiza la similaridad entre clusters, para ello se representan gráficamente el vector de la suma de cuadrados entre clusters para diferentes números de clusters.

En el siguiente punto aplicaremos esta solución para ver si coincide con el resultado de aplicar clustering jerárquico.

### 2.2.3. Herramienta de análisis: R + RStudio

El software estadístico que hemos elegido para realizar el análisis es R + RStudio [14]. Utilizaremos el motor R con la interfaz de usuario RStudio, que nos permite sacar el máximo partido a R.

***¿Qué es R?***

R [14] es un lenguaje y entorno de programación para el análisis estadístico y gráfico. Son un conjunto de herramientas que pueden extenderse fácilmente mediante paquetes, librerías o creando nuestras propias funciones.

***Motivos para elegir R***

Algunos de los motivos por los que elegimos R son los siguientes:

- Gran potencia para procesamiento y manipulación de datos.
- Gratuito y de código abierto.
- Proyecto colaborativo.
- Gran cantidad de documentación.
- Consume pocos recursos.
- Disponible para todos los sistemas operativos.
- Lenguaje orientado a objetos.

R es el motor que hay detrás de la mayoría de software estadístico del mercado, como por ejemplo SPSS, STATA o SAS.

Como hemos comentado, uno de los puntos fuertes de R es su gran comunidad de usuarios, con toda la documentación, desarrollo, ejemplos y soporte que eso significa. Los usuarios pueden crear sus propios paquetes que se instalan como extensiones del programa base. Existen infinidad de paquetes para casi todo lo que queramos realizar: paquetes de importación de datos, de conversión, específicos para análisis de encuestas, gráficos, informes, etc.

### *Análisis clúster con R*

Como comentamos en la introducción, nuestro trabajo se centra en aplicar las técnicas de clustering presentadas en el punto anterior sobre los datos que disponemos.

En R existen muchos paquetes dedicados al análisis clúster, para nuestro análisis utilizaremos los paquetes ***stats [15]*** y ***cluster [16]***.

Para el cálculo de la matriz de distancias utilizaremos la función ***daisy*** del paquete ***cluster.***

Dentro del paquete ***stats***, las funciones más importantes que usaremos son las siguientes:

- ***hclust*** - Clustering jerárquico. Dendrograma.
- ***kmeans*** - Clustering no jerárquico.

Estas tres funciones soportan varios métodos de cálculo distintos cada una, estudiaremos cada uno de los métodos para poder elegir el que, según nuestro criterio, nos ofrezca mejores resultados sobre nuestros datos.

Como explicamos en el punto anterior, el algoritmo de clustering agrupa las observaciones en función de su similitud y esta similitud la medimos con un concepto llamado distancia. El paquete ***cluster*** de R dispone de la función ***daisy*** para este propósito. Esta función soporta 3 métodos de cálculo de distancia: manhattan, euclídea y gower.

Distancia euclídea: Es la más utilizada, es la distancia entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras. Su definición coincide con el concepto más común de distancia, la que podríamos medir con una regla.

Método Manhattan: Distancia absoluta entre dos puntos. Utilizando un símil con un sistema de gps, la distancia entre dos puntos no es la recta que los une, sino el mínimo número de calles que se deben recorrer.

Distancia gower: Útil cuando las variables usadas no son numéricas (nominales, binarias, ordinales o incluso combinaciones de ellas).

Atendiendo a la tipología de nuestros datos, usaremos el método gower. Cuando expliquemos la tipología de las respuestas del cuestionario veremos las razones.

La función ***hclust*** que hemos usado para dibujar el dendrograma también soporta varios métodos de cálculo, en este caso métodos de fusión de grupos (criterio para determinar la distancia entre dos grupos).

Complete: La distancia entre dos clusters se mide atendiendo a sus elementos más dispares, es decir, la similitud viene dada por la máxima distancia entre sus componentes.

Single: La distancia entre dos clusters viene dada por la mínima distancia entre sus componentes.

Average: La similitud de dos clusters se calcula como la media aritmética entre la distancia de las componentes de dichos clusters.

Centroid: La semejanza entre dos clusters viene dada por la semejanza de sus centroides, es decir, los vectores de medias de las variables medidas sobre los individuos del clúster.

Ward: En cada etapa se unen los dos clusters para los cuales se tenga menor incremento del valor total de la suma de los cuadrados de las diferencias, dentro de cada clúster, de cada individuo al centroide del clúster.

En este caso, nosotros trabajaremos con Complete, determinaremos la distancia entre dos clusters como la distancia de sus elementos más dispares, el valor por defecto de ***hclust***.

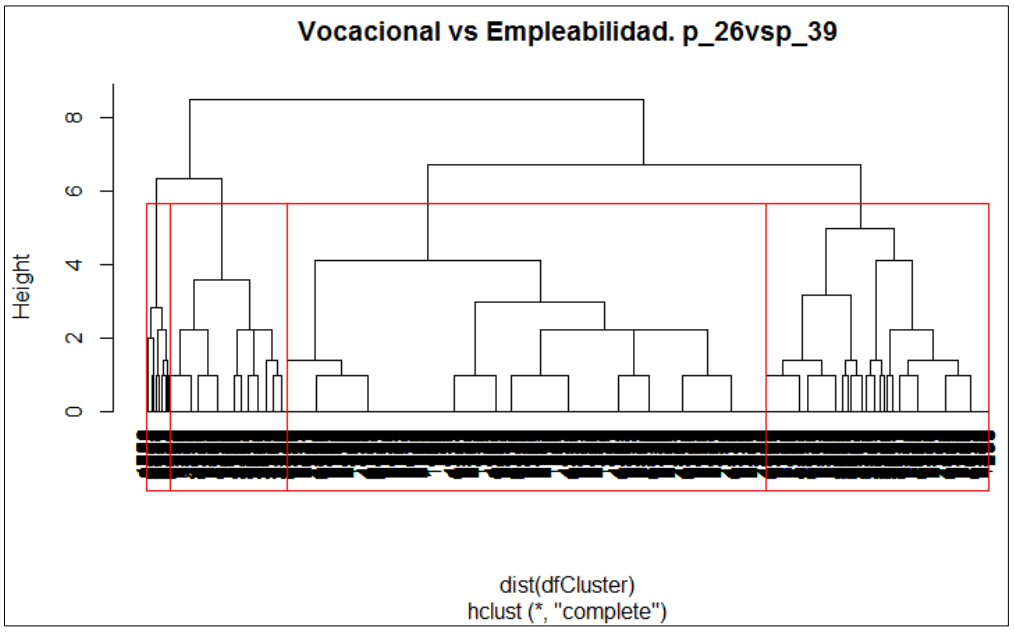


**Figura 13**

Según vemos en el dendrograma resultante, es probable que el número óptimo de clusters sea 4.

A continuación contrastamos ese número de clusters con el método Elbow [13] mencionado anteriormente.

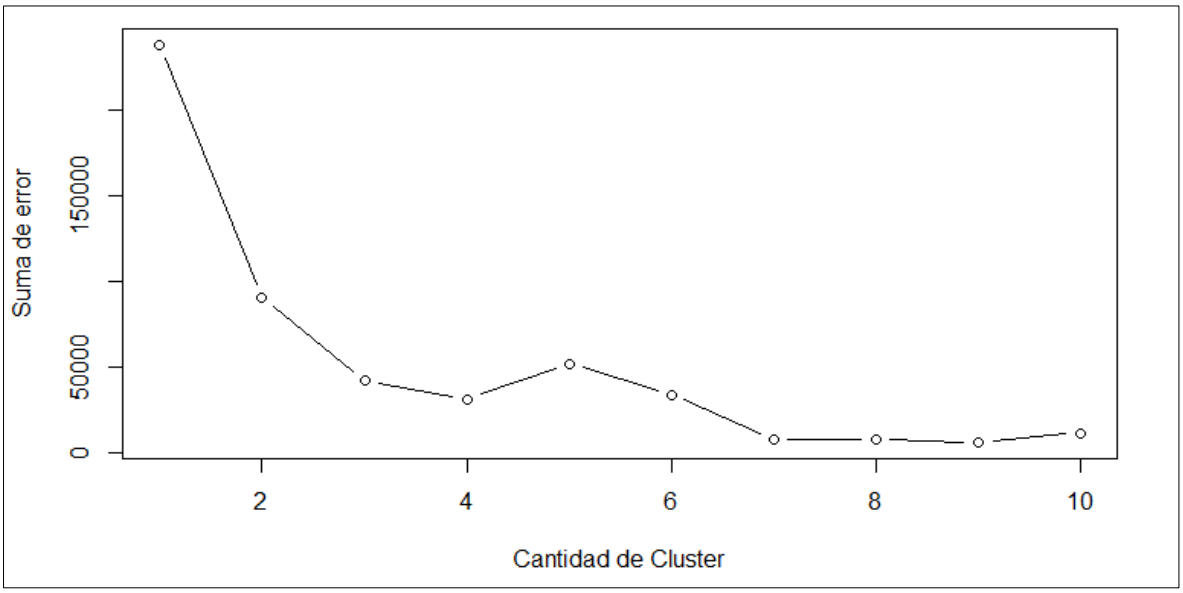


**Figura 14**

En el gráfico de la figura 14 vemos que a partir del número 4 la diferencia de suma de error es insustancial, lo cual concuerda con lo que observamos en el dendrograma resultante del algoritmo de clustering jerárquico.

Una vez ya tenemos identificados el número óptimo de clusters, pasamos al algoritmo de clustering no jerárquico, kmeans, para el cual R incluye 4 variaciones:

- Lloyd
- Forgy
- MacQueen
- Hartigan-Wong

En el apartado de análisis compararemos esas 4 variaciones y elegiremos la que mejores resultados sobre nuestros datos nos proporcione.

# 3. Fuentes de datos

Como hemos ido comentando en los puntos anteriores, nuestro estudio se basa en un trabajo previo en el que se desarrolla un cuestionario sobre factores que influyen a los estudiantes en el momento de la elección de estudios universitarios en el SUPE.

## 3.1. Técnica aplicada: Encuestas

El cuestionario desarrollado consta de 34 preguntas de respuesta ordinal Likert 1 a 7 (1= máximo desacuerdo a 7 = máximo acuerdo).

Los factores que estudia la encuesta son los siguientes:

- Factores individuales, vinculados a aspectos relacionados con el estudiante y sus intereses personales, algunos de carácter académico y otros claramente no.

- Factores sociales, relacionados con aspectos que tienen que ver con la percepción social de la titulación, su empleabilidad, la percepción que se tiene del prestigio de la Universidad, del título, o de la profesión a la que da acceso.

- Datos de contexto, relativos al valor de la nota de corte de una titulación, a la oferta de plazas o a la demanda que se ha producido en una determinada titulación y universidad en los años inmediatamente anteriores al de la toma de la decisión.

Los datos de contexto fueron proporcionados por las universidades colaboradoras.

Las 34 preguntas sobre los factores se distribuyeron de la siguiente forma:

- Datos generales - 9 preguntas
- Preguntas referentes a factores sociales
    - Consideración de la Universidad - 5 preguntas
    - Utilidad percibida - 3 preguntas
    - Consideración social - 3 preguntas
- Preguntas referentes a factores individuales
    - Aspectos vocacionales - 4 preguntas
    - Influencia del entorno - 8 preguntas
    - Localización geográfica - 2 preguntas

Una vez pasada la encuesta se recogieron un total de 5.111 encuestas válidas. En la encuesta participaron alumnos de 31 universidades públicas y de 116 titulaciones distintas lo que permite tener información de prácticamente todo el sistema universitario público español y se realizó principalmente por vía telemática (75%) entre Octubre de 2010 y Abril de 2013.

El número de participantes por rama de estudios son los siguientes:

- Ingeniería y Arquitectura - 3239
- Artes y Humanidades - 319
- Ciencias - 211
- Ciencias de la Salud - 545
- Ciencias Sociales y Jurídicas - 797

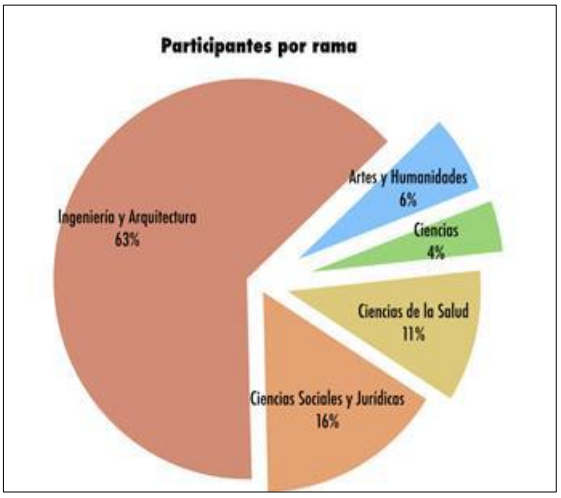


**Figura 15**

Y la distribución por género:

- Hombres - 62%
- Mujeres - 38%

Con la siguiente proporción por rama:

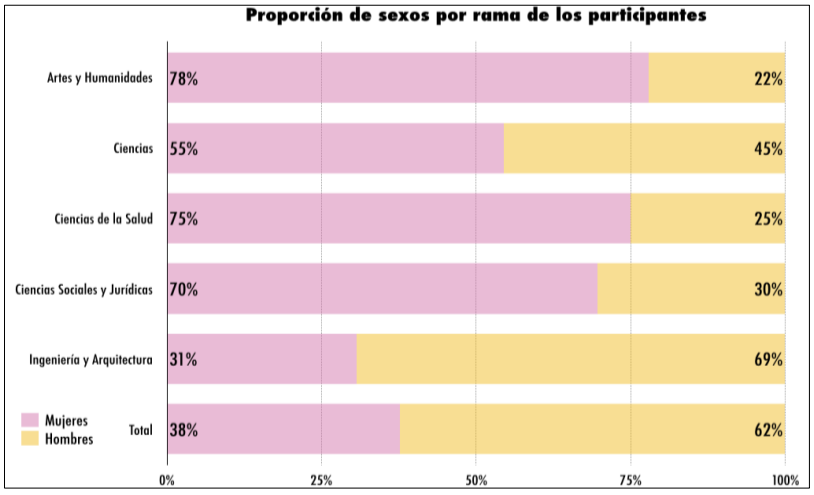


**Figura 16**

## 3.2. Datos

Los resultados de las encuestas se nos proporcionaron en un Excel, en los siguientes puntos detallamos la información que contiene.

### *Datos de acceso universitario*

Estos datos fueron proporcionados por las universidades colaboradoras y proporcionan información sobre la titulación en la que se matricularon los alumnos y datos como la nota de corte, plazas ofertadas, solicitudes de ingreso y número de matrículas en el año anterior.

Con estos datos tenemos la información necesaria sobre la oferta/demanda de la titulación en la que se matricularon los alumnos así como las condiciones de ingreso relacionadas con la nota del alumno y la de corte de la titulación.

**Tabla 3**

| | **Pregunta** | **Tipo respuesta** |
|---|---|---|
| especialidad_ITI | Especialidad ITI | 1 / 2 / 3 / 4 / 5 |
| UNI_CEN_TIT | Código de titulación | Literal numérico |
| Universidad | Universidad | Literal |
| Centro | Centro | Literal |
| Titulación | Titulación | Literal |
| Grado | Grado | Literal |
| Rama | Rama | Literal |
| Año | Año encuesta | YYYY |
| Oferta_plazas | Plazas ofertadas | Entero |
| Nota_corte | Nota de corte | Decimal |
| Solicitudes | Número de solicitudes de ingreso | Entero |
| matriculados_curso_anterior | Matriculados curso anterior | Entero |
| dif_notas | Diferencia entre nota de acceso y nota de corte | Entero |

### *Preguntas*

Derivado de los factores influyentes mencionados anteriormente, se definieron una serie de preguntas para cada uno de los factores, además de 6 preguntas de datos generales con las que se pretendía poner en contexto la situación del alumno en el momento de la elección.

Las preguntas se organizaron en 8 secciones, que se corresponden con los datos generales y de percepción de la oferta y demanda y con los 6 factores influyentes indicados (3 factores individuales y 3 factores sociales).

A continuación se detallan las preguntas por secciones del cuestionario:

## Sección Datos Generales

**Tabla 4**

| | **Pregunta** | **Tipo de respuesta** |
|---|---|---|
| p_3 | Sexo | F o M |
| p_5 | ¿Qué rama de Bachillerato o qué otros estudios cursaste?<br>•Artes<br>•Ciencias y Tecnología<br>•Humanidades y Ciencias Sociales<br>•Ciclos Formativos de Grado Superior<br>•Otros estudios previos | A1 / A2 / A3 / A4 / A5 |
| p_4 | ¿Cuál fue tu nota de acceso a la Universidad? | Decimal |
| curso_ing | ¿En qué año ingresaste en esta titulación? | YYYY |
| p_7 | La Universidad en la que estudio fue mi primera opción | 0 / 1 |
| p_6 | La titulación que curso fue mi primera opción en la preinscripción | 0 / 1 |

## Sección Percepción de la oferta/demanda

**Tabla 5**

| | **Pregunta** | **Tipo de respuesta** |
|---|---|---|
| p_08 | ¿Ha influido la nota de corte en la elección? | Likert 1 a 7 |
| p_10 | Sabía cuántas plazas se ofertaban | Likert 1 a 7 |
| p_11 | Por mi nota no pude elegir las titulaciones que quería | Likert 1 a 7 |
| p_12 | Elegí entre las titulaciones en que sabía que podía ser admitido | Likert 1 a 7 |

## Sección Localización geográfica

**Tabla 6**

| | **Pregunta** | **Tipo de respuesta** |
|---|---|---|
| p_13 | La proximidad del centro a mi domicilio fue un factor determinante | Likert 1 a 7 |
| p_14 | Prefería estudiar en otro lugar (ciudad, comunidad,...) | Likert 1 a 7 |

## Sección Entorno escolar

**Tabla 7**

| | **Pregunta** | **Tipo de respuesta** |
|---|---|---|
| p_15 | Mis compañeros influyeron en mi decisión | Likert 1 a 7 |
| p_16 | El orientador me ayudó a tomar la decisión | Likert 1 a 7 |
| p_17 | Hubo algunos profesores que influyeron decididamente | Likert 1 a 7 |
| p_18 | La página Web de la Universidad me proporcionó la información necesaria | Likert 1 a 7 |
| p_19 | La visita al centro/Universidad o la asistencia a las jornadas de acogida fueron determinantes | Likert 1 a 7 |

## Sección Entorno próximo

**Tabla 8**

| | **Pregunta** | **Tipo de respuesta** |
|---|---|---|
| p_20 | Mis padres/hermanos habían estudiado esa carrera | Likert 1 a 7 |
| p_21 | Mis padres/hermanos habían estudiado en esa Universidad | Likert 1 a 7 |
| p_22 | Amigos que estaban o habían estudiado aquí influyeron positivamente | Likert 1 a 7 |

## Sección Aspectos vocacionales

**Tabla 9**

| | **Pregunta** | **Tipo de respuesta** |
|---|---|---|
| p_23 | Siempre he querido estudiar esta carrera | Likert 1 a 7 |
| p_24 | Fue una decisión de última hora | Likert 1 a 7 |
| p_25 | Creo que mis habilidades personales son adecuadas al título | Likert 1 a 7 |
| p_26 | Siempre he sacado buenas notas en las asignaturas de bachillerato relacionadas con la titulación | Likert 1 a 7 |

## Sección Consideración social

**Tabla 10**

| | **Pregunta** | **Tipo de respuesta** |
|---|---|---|
| p_29 | La calidad y el prestigio de la Universidad fueron determinantes en mi elección | Likert 1 a 7 |
| p_30 | Le di más importancia al título que a la universidad | Likert 1 a 7 |
| p_31 | El hecho de ser titulado por una Universidad u otra da mayores posibilidades laborales | Likert 1 a 7 |
| p_32 | El título está prestigiado socialmente | Likert 1 a 7 |
| p_33 | El título es reconocido internacionalmente | Likert 1 a 7 |
| p_34 | Da acceso a una profesión reconocida | Likert 1 a 7 |

## Sección Empleabilidad

**Tabla 11**

| | **Pregunta** | **Tipo de respuesta** |
|---|---|---|
| p_36 | He elegido el título porque me gusta sin preocuparme las salidas profesionales | Likert 1 a 7 |
| p_37 | Creo que los salarios que se consiguen en esta titulación son mejores que en otras | Likert 1 a 7 |
| p_38 | Creo que existe demanda de trabajo en el sector | Likert 1 a 7 |
| p_39 | Es más fácil encontrar trabajo | Likert 1 a 7 |

## 3.3. Tipología de respuestas

A excepción de los datos generales y de contexto, la tipología de las respuestas del cuestionario es formato escala Likert 1 a 7 (1= máximo desacuerdo y 7 = máximo acuerdo).

La escala de Likert [17] es una herramienta de medición que, a diferencia de preguntas dicotómicas con respuesta sí/no, nos permite medir actitudes y conocer el grado de conformidad del encuestado con cualquier afirmación que le propongamos.

Los datos recogidos por este tipo de preguntas se pueden agrupar en cuatro niveles de medición:

- Datos nominales: El nivel de medición más débil que representa categorías sin representación numérica.
- Datos ordinales: Datos en los que es posible ordenar o clasificar las respuestas, pero no es posible medir la distancia.
- Datos de intervalo: En general, datos enteros en los que se pueden realizar mediciones de pedidos y distancias.
- Datos de relación: datos en los que es posible el ordenamiento significativo, distancia, decimales y fracciones entre variables.

En nuestro caso, nuestros datos son ordinales, por lo que no podemos determinar la distancia entre dos respuestas.

Como hemos comentado anteriormente este es el motivo por el cual para analizar estos datos con estadística descriptiva utilizamos la mediana en lugar de la media. Y para calcular la matriz de distancias usamos funciones específicamente desarrolladas para este tipo de datos.

Para facilitar la interpretación, en el momento en el que analicemos los resultados del análisis, vamos a establecer 3 grados de influencia.

- Poca influencia: Valores de 1 a 3.
- Influencia parcial: 4.
- Mucha influencia: Valores 5 a 7.

# 4. Análisis

Llegados a este punto, vamos a comenzar el análisis con R de nuestros datos. Para ello hemos definido una serie de pasos con los cuales esperamos llegar a obtener conclusiones sobre los distintos perfiles de alumnos que existen en nuestras universidades.

Los pasos que seguiremos son los siguientes:

1. Preprocesado de datos: Carga de los datos en R, preparación, limpieza y organización de los datos para facilitar el análisis y su interpretación.
2. Selección de variables: Analizaremos las variables significativas y factores obtenidos del TFM anteriormente mencionado.
3. Exploración de datos: Resumen estadístico de cada variable del análisis.
4. Clustering con el total de variables significativas.
5. Clustering por dos factores o variables significativas.
6. Interpretación de resultados.

## 4.1. Prepocesado de datos

En este primer paso, hemos trabajado para cargar y preparar los datos de manera que facilitemos su manejo e interpretación.

### 4.1.1 Preparar los datos en CSV e importarlos en R.

Como hemos comentado en la parte de fuentes de datos, los datos se nos presentaron en un fichero Excel con varias hojas entre las cuales una de ellas contenía las respuestas de 5113 encuestas tal y como se ilustra en la figura 17.

| S | T | U | V | W | X | Y | Z | AA |
|---|---|---|---|---|---|---|---|---|
| p_10 | p_11 | año | oferta_plazas | nota_corte | solicitudes_1 | matriculados_curso_anterior | p_12 | p_13 |
| 0 | 0 | 2009 | 80 | 5 | 92 | 85 | 3 | 4 |
| 1 | 0 | 2011 | 110 | 10,238 | 41 | 110 | 1 | 1 |
| 0 | 1 | 2012 | 110 | 9,733 | 45 | 108 | 4 | 5 |
| 0 | 1 | 2011 | 60 | 10,255 | 54 | 64 | 4 | 5 |
| 0 | 0 | 2010 | 110 | 9,505 | 41 | 110 | 5 | 2 |

**Figura 17**

Con esto, guardamos esa hoja en formato csv usando como separador de campos “**;**”.

En R hay multitud de funciones para la importación de datos, nosotros usamos ***read.csv*** del paquete ***utils*** y cargamos las encuestas en un objeto tipo ***data.frame***. Este objeto será el que utilizaremos para extraer todos los subconjuntos con los que trabajaremos en el análisis, ya sean de filas o de columnas.

Para facilitar el manejo y la identificación de las preguntas, preparamos un diccionario con los enunciados de las preguntas reducidos. Nuestro data.frame queda de la siguiente manera:

| oferta_plazas | nota_corte | p_12EligioDondePodiaSerAdmitido | p_13ProximidadDeterminante |
|---|---|---|---|
| 255 | 11.959 | 2 | 2 |
| 255 | 11.959 | 7 | 1 |
| 255 | 11.959 | 7 | 1 |

**Figura 18**

Podemos observar que el nombre de las columnas incluye el enunciado reducido y tenemos nuestros datos perfectamente cargados y etiquetados en R.

### 4.1.2 Limpieza de datos

En este punto descartamos las observaciones que no nos sirven y seleccionamos las variables que nos interesan descartando el resto. Aquí es donde, aunque en nuestro caso no es necesario, homogeneizaríamos los datos en cuanto a formato.

En un primer paso, seleccionaremos las variables (columnas) que nos interesan. Para nuestro análisis nos quedamos con las siguientes:

- Variables de preguntas.
- Rama
- Universidad
- Grado
- Titulación
- Oferta de plazas
- Nota de corte
- Nota de acceso

Una vez tenemos nuestro data.frame con las variables que pretendemos utilizar en el análisis, descartamos las observaciones en las que alguna de las variables tenga un valor nulo. Para ello usamos la función ***na.omit*** del paquete ***stats***. Después de esto nos quedamos con 4915 observaciones válidas.

### 4.1.3 Preparar subconjuntos de datos por Ramas

Como hemos comentado anteriormente, parte de nuestro análisis pretende buscar diferentes perfiles entre las distintas ramas de estudio. Para esto, se preparan 5 subconjuntos sobre el total de los datos con las observaciones de cada rama, quedando de la siguiente manera:

Ingeniería y Arquitectura → 3319 observaciones.
Artes y Humanidades → 311 observaciones.
Ciencias → 205 observaciones
Ciencias de la Salud → 522 observaciones
Ciencias Sociales y Jurídicas → 758 observaciones

## 4.2. Selección de variables

Del TFM de Regina Sandra Mollaplaza extraemos los factores y preguntas significativas derivadas del análisis factorial que realiza. En este punto vamos a estudiar esos factores y la correlación entre ellos.

### 4.2.1 PCA y Análisis Factorial: Preguntas significativas

El análisis realizado determina la existencia de 8 factores. El valor de cada uno de los factores viene dado por la combinación de varias preguntas. En la tabla 12 se muestran los factores y preguntas significativas asociadas.

**Tabla 12**

| | | |
|---|---|---|
| F1 | Prestigio social de la titulación. | El título está prestigiado socialmente |
| | | El título es reconocido internacionalmente |
| F2 | Empleabilidad | Creo que existe demanda de trabajo en el sector |
| | | Es más fácil encontrar trabajo |
| F3 | Aspectos vocacionales | Siempre he querido estudiar esta carrera |
| | | Fue una decisión de última hora |
| F4 | Prestigio social de la Universidad | La calidad y el prestigio de la Universidad fueron determinantes en mi elección |
| F5 | Entorno académico | El orientador me ayudó a tomar la decisión |
| | | Hubo algunos profesores que influyeron decididamente |
| F6 | Adaptación personal a la titulación | Siempre he sacado buenas notas en las asignaturas de bachillerato relacionadas con la titulación |
| F7 | Entorno familiar | Mis padres/hermanos habían estudiado en esa Universidad |
| F8 | Fuentes de información | La visita al centro/Universidad o la asistencia a las jornadas de acogida fueron determinantes |

Llama la atención las preguntas significativas del factor F3, aspectos vocacionales, ya que son totalmente opuestas. En el siguiente punto estudiaremos este caso concreto para ver porqué se seleccionan estas preguntas como significativas.

### 4.2.2 Correlación

Tal y como explicamos en el punto de estadística descriptiva, la correlación indica el grado de relación lineal entre variables.

Un aspecto importante para aplicar clustering con éxito es que las variables utilizadas para agrupar estén muy poco o nada correlacionadas y así evitar información redundante.

Vamos a estudiar la correlación entre nuestras preguntas significativas.

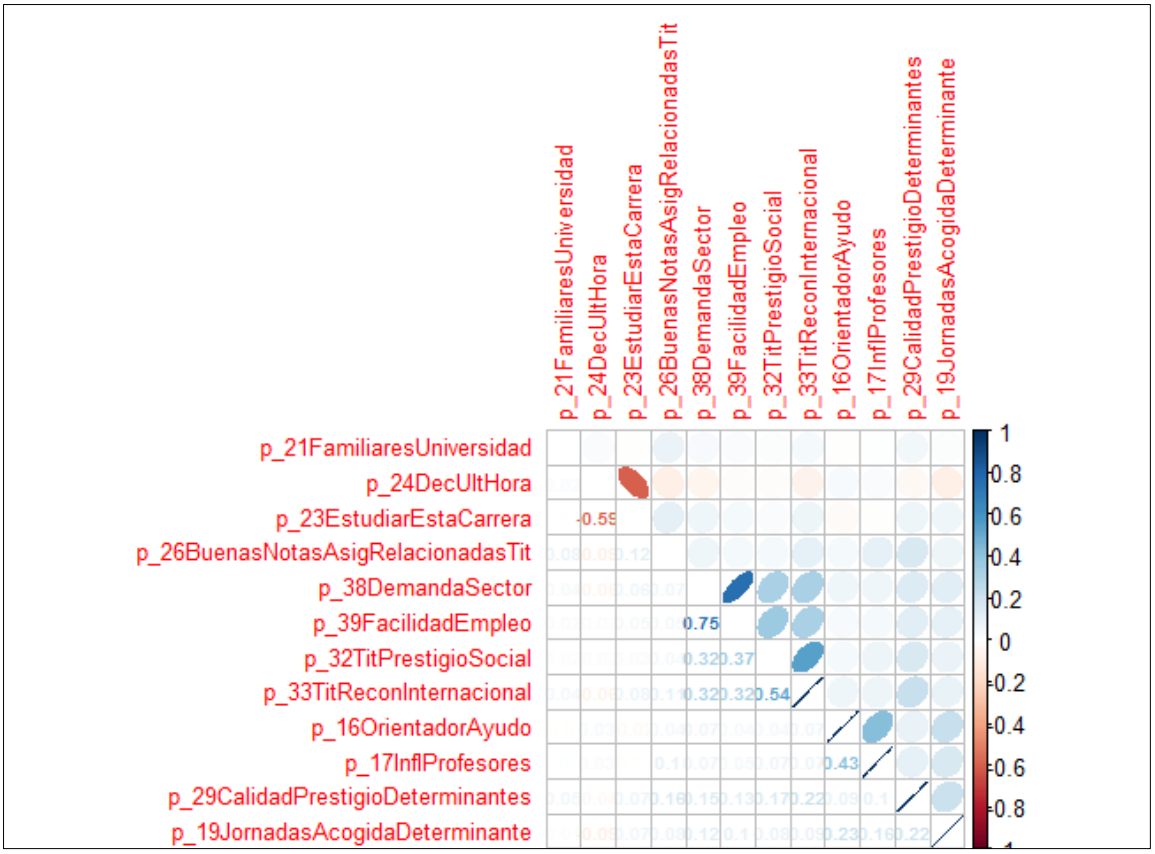


**Figura 19**

Tal y como hemos adelantado, en el gráfico de la figura 19 podemos ver cómo la pregunta p_23 (Siempre he querido estudiar esta carrera) tiene una fuerte correlación inversa con la pregunta p_24 (Fue una decisión de última hora). Es decir, contra más puntuación tengamos en la p_23, menor puntuación tendrá la p_24 y viceversa.

También vemos las siguientes correlaciones positivas:

- p_38 y p_39, ambas del factor relacionado con la empleabilidad.
- p_38, p_39, p_32 y p_33. Es decir los factores F1 y F2.
- p_16 y p17, ambas del factor relacionado con la influencia del entorno académico.

Vistos estos resultados, podemos descartar aplicar clustering por las preguntas significativas de F1 y F2. Con el resto de factores no tendremos problemas.

### 4.2.3 Exploración de los datos

En este punto realizamos un estudio descriptivo de las preguntas significativas para obtener un resumen global de las preguntas significativas.

Para ello preparamos la tabla 13 con los cálculos de cada pregunta y traducimos los datos a histogramas y gráficos de densidad para interpretarlos de forma rápida y eficaz.

**Tabla 13**

| Pregunta | Mediana (IQR) |
|---|---|
| F1 - El título está prestigiado socialmente | 5 (2) |
| F1 - El título es reconocido internacionalmente | 5 (2) |
| F2 - Creo que existe demanda de trabajo en el sector | 4 (3) |
| F2 - Es más fácil encontrar trabajo | 4 (3) |
| F3 - Siempre he querido estudiar esta carrera | 4 (3) |
| F3 - Fue una decisión de última hora | 2 (3) |
| F4 - La calidad y el prestigio de la Universidad fueron determinantes en mi elección | 4 (3) |
| F5 - El orientador me ayudó a tomar la decisión | 1 (1) |
| F5 - Hubo algunos profesores que influyeron decididamente | 1 (3) |
| F6 - Siempre he sacado buenas notas en las asignaturas de bachillerato relacionadas con la titulación | 6 ( 3) |
| F7 - Mis padres/hermanos habían estudiado en esa Universidad | 1 (0) |
| F8 - La visita al centro/Universidad o la asistencia a las jornadas de acogida fueron determinantes | 2 (3) |

Con estos resultados vamos a seleccionar qué factores queremos utilizar para nuestro análisis clúster basándonos en dos criterios:

- Los factores cuyos valores tengan una dispersión que nos permita obtener agrupaciones con valores distintos.
- Factores cuyo enfrentamiento 1 vs 1 nos pueda ofrecer información más interesante.

Descartamos los factores F5, F7 y F8 porque presentan valores centrales muy bajos y, en el caso de F5 y F7, con muy poca dispersión.

Por último, el factor 6 sí que presenta valores centrales y de dispersión altos, pero creemos que la interpretación de la pregunta se puede incluir también en el factor vocacional. Por este motivo también lo excluimos de nuestro análisis.

En base a estos criterios, seleccionamos los siguientes factores y preguntas:

- Factor 1: Preguntas p_32 y p_33

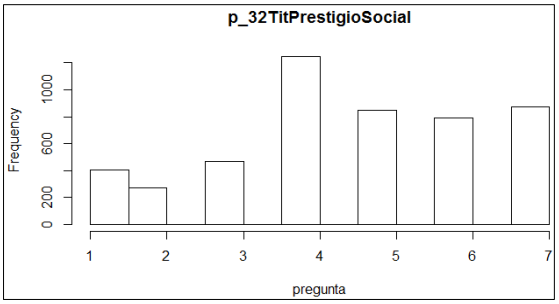


**Figura 20**

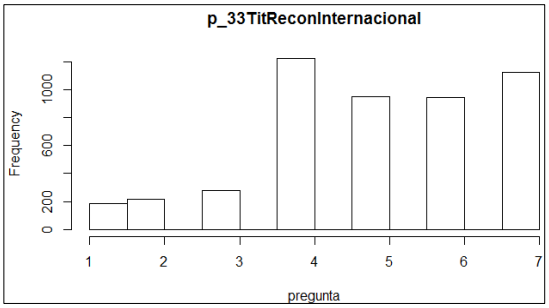


**Figura 21**

- Factor 2: Preguntas p_38 y p_39

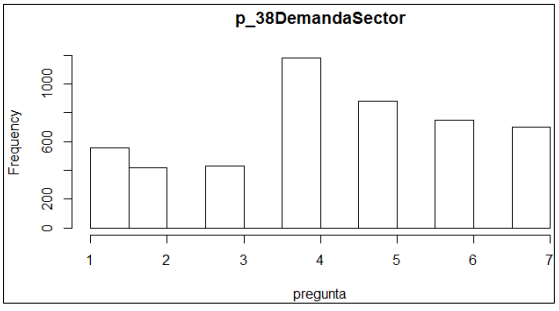


**Figura 22**

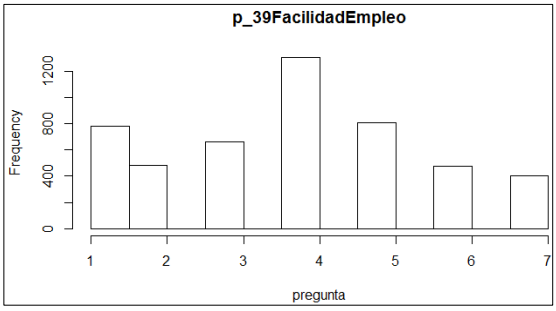


**Figura 23**

- Factor 3: Preguntas p_23 y p_24

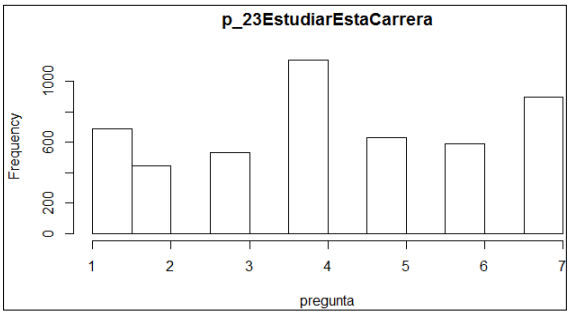


**Figura 24**

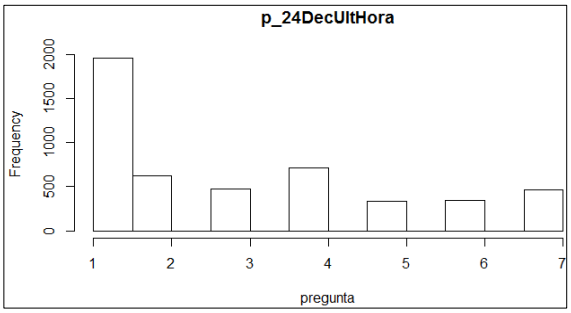


**Figura 25**

- Factor 4: Pregunta p_29

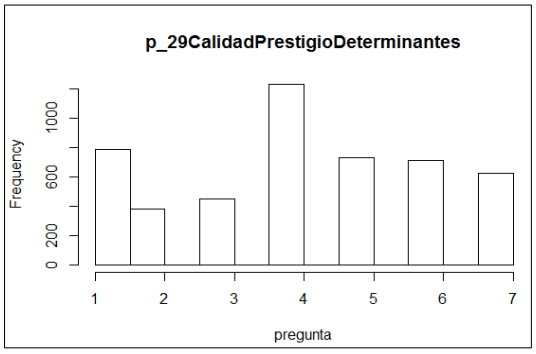


**Figura 26**

## 4.3. Clustering

Como hemos comentado en el apartado de R, la función que vamos a utilizar para hacer el clustering jerárquico (kmeans) ofrece 4 algoritmos de cálculo distintos. Vamos a compararlos para ver cuál ofrece mejores resultados.

Las diferencias entre los 4 algoritmos se centran en la forma de elegir los centroides iniciales y la forma para asignar cada individuo al nuevo clúster.

Para compararlas usaremos la distancia intracluster [18], es decir, la suma de las distancias entre los centroides. Elegiremos el algoritmo que tenga mayor distancia ya que eso significa que será la mejor separación de grupos.

1. Ejecutamos kmeans 10 veces para cada algoritmo y guardamos la distancia intracluster de cada iteración.
2. Calculamos la media de distancia intracluster para cada algoritmo.

Al realizar el proceso en R observamos que los resultados son iguales para los 4 algoritmos, la media de distancia intracluster es de 19529. Por lo tanto, realizaremos el análisis con el algoritmo por defecto, ***Hartigan-Wong***.

En la figura 27 se pueden ver la distribución de los clusters obtenidos sobre el total de observaciones agrupando por preguntas relacionadas con empleabilidad y con vocación.

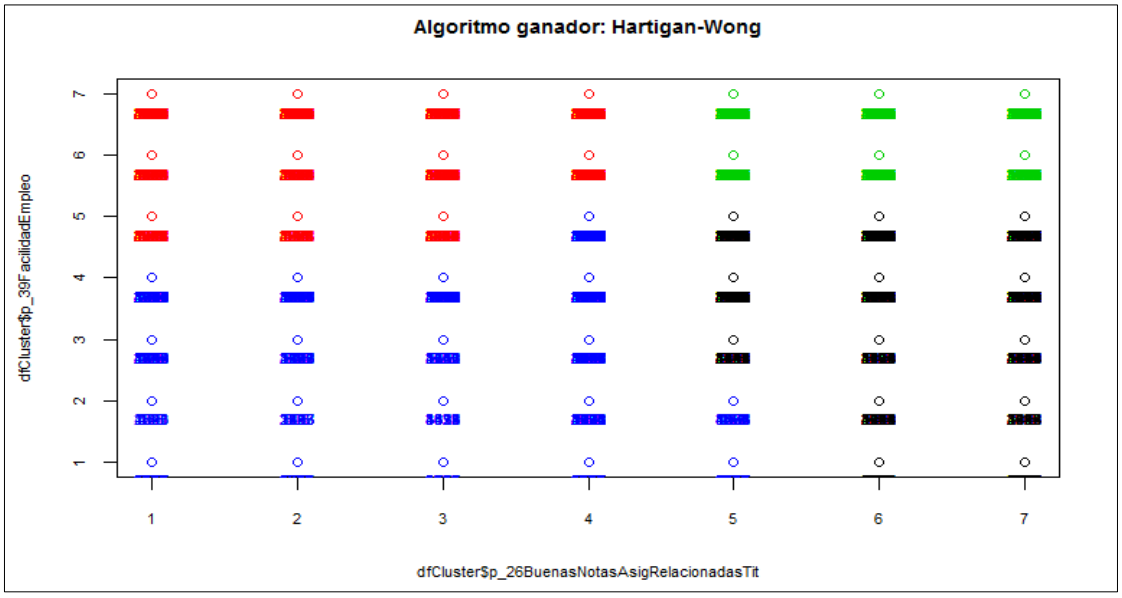


**Figura 27**

### 4.3.1 Clustering con todas las preguntas significativas

Antes de empezar el análisis enfrentando los factores 1 vs 1 vamos a agrupar nuestras observaciones en base a todas las preguntas significativas.

Determinamos el número idóneo de clusters mediante el método Elbow.

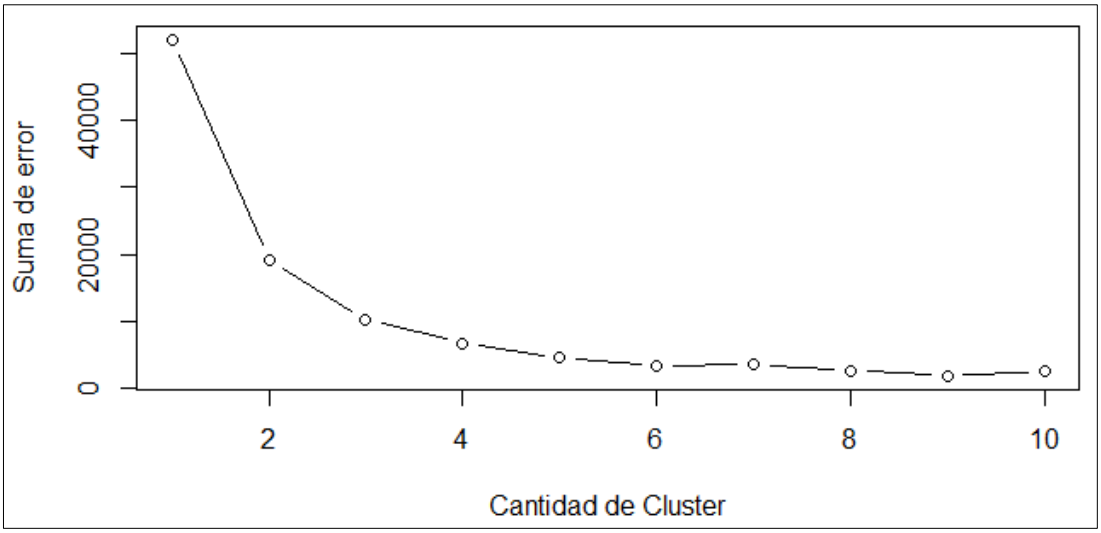


**Figura 28**

Agrupamos con k = 4 y obtenemos los siguientes grupos caracterizados en la tabla 14:

**Tabla 14**

| Grupo | n | F1 : Prestigio social de la titulación | | F2: Empleabilidad | | F3: Vocación | | F4: Calidad y prestigio de la universidad |
|---|---|---|---|---|---|---|---|---|
| | | p_32 Median a (iqr) | p_33 Median a (iqr) | p_38 Median a (iqr) | p_39 Median a (iqr) | p_23 Median a (iqr) | p_24 Median a (iqr) | p_29 Mediana (iqr) |
| 1 | 2117 | 4(1) | 5(2) | 4(1) | 4(2) | 4(3) | 2(3) | 4(2) |
| 2 | 755 | 6(3) | 6(2) | 6(3) | 5(2) | 2(3) | 6(3) | 5(3) |
| 3 | 995 | 6(2) | 6(2) | 6(2) | 5(2) | 6(3) | 1(1) | 6(3) |
| 4 | 1048 | 3(3) | 4(3) | 2(2) | 1(1) | 4(4) | 4(4) | 3(3) |

El grupo 1, con 2117 observaciones, está formado por alumnos con valores medios en los 4 factores.

Los grupos 2 y 3 en cambio tienen valores altos en los factores relacionados con el prestigio de la titulación, la empleabilidad y con el prestigio de la universidad, únicamente difieren en el factor relacionado con aspectos vocacionales. Los alumnos del grupo 2 no se ven influenciados por aspectos vocacionales mientras los alumnos del grupo 3 se ven muy influenciados por estos aspectos.

En el grupo 4 se agrupan alumnos con valores medios en los factores relacionados con el prestigio de la titulación y universidad, y con factores relacionados con aspectos vocacionales. En cambio tienen valores muy bajos en el factor relacionado con la empleabilidad.

En la tabla 15 mostramos el resumen de las variables observables por cada clúster.

**Tabla 15**

| Grupos | n | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|
| 1 | 2117 | F 849 M 1200 N 68 | 8.95(2.03) | 7(1.92) | 1.95(2.18) |
| 2 | 755 | F 302 M 437 N 16 | 9.14(2.09) | 7.13(2.02) | 2.01(2.11) |
| 3 | 995 | F 321 M 640 N 34 | 9.44(2.09) | 7.18(2.02) | 2.26(2.26) |
| 4 | 1048 | F 563 M 429 N 56 | 8.45(2.05) | 6.65(1.64) | 1.8(2.34) |

Observamos que los alumnos del grupo 3 (puntuaciones altas en los 4 factores) son lo que tienen mayor nota de acceso y entran a la titulación superando la nota de corte en una media de 2.26 puntos.

### 4.3.2 Clustering por factores 1 vs 1

En el punto anterior hemos seleccionado 4 factores sobre los cuales vamos a realizar nuestro análisis.

- Factor 1: Prestigio social de la titulación
- Factor 2: Empleabilidad
- Factor 3: Aspectos vocacionales
- Factor 4: Calidad y prestigio social de la universidad

Como vimos en el estudio de la correlación, los factores 1 y 2 tienen están muy relacionados linealmente y la información que nos proporcionarán será en gran medida redundante, por lo que descartamos hacer un análisis 1 vs 1 entre esos factores.

Con esto, hacemos los siguientes análisis sobre el total de observaciones:

- Prestigio social de la titulación vs Calidad y prestigio de la universidad
- Empleabilidad vs Vocación
- Calidad y prestigio de la universidad vs Vocación
- Prestigio social de la titulación vs Vocación

Una vez tengamos los grupos, para los dos primeros profundizaremos en el estudio por ramas.

## 4.4. Resultados

A continuación se detallan los resultados del análisis clúster por los factores indicados utilizando las preguntas significativas de cada uno.

Antes de aplicar clustering no jerárquico determinaremos el número idóneo de clúster mediante el método Elbow y obtendremos el dendrograma resultante de aplicar clustering jerárquico para corroborar el resultado.

Para cada análisis mostramos una tabla con la caracterización de cada grupo resultante con respecto a las variables utilizadas para agrupar y las variables observables Sexo, Nota de Acceso, Nota de Corte y Diferencia entre nota de acceso y nota de corte.

Como hemos comentado, aplicaremos clustering sobre el total de observaciones y posteriormente, para los análisis “Prestigio social de la titulación vs Calidad y prestigio de la Universidad” y “Empleabilidad vs Vocación”, repetiremos el análisis por ramas.

### *Prestigio social de la titulación vs Calidad y prestigio de la universidad*

Realizamos el análisis agrupando por las preguntas relacionadas a los factores.

**Tabla 16**

| Factor | Pregunta |
|---|---|
| 1 | p_32 - El título está prestigiado socialmente |
| | p_33 - El título es reconocido internacionalmente |
| 4 | p_29 - La calidad y el prestigio de la Universidad fueron determinantes en mi elección |

Determinamos el número idóneo de clusters:

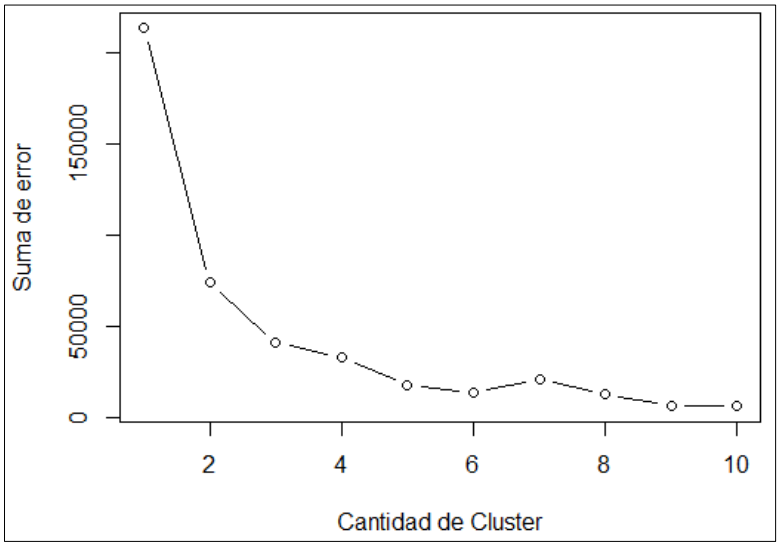


**Figura 29**

En la figura 29 vemos cómo a partir de 5 clusters la reducción de la suma de error no es significativa. En el siguiente dendrograma mostramos los grupos definidos.

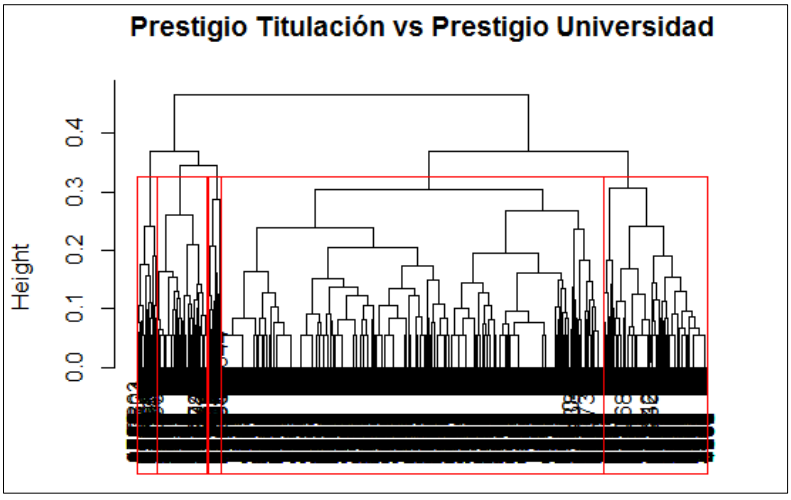


**Figura 30**

Agrupamos con k = 5 y caracterizamos los clusters:

**Tabla 17**

| Grupos | n | Prestigio social de la titulación | | Calidad y prestigio de la univers. | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|
| | | p_32 Mediana (iqr) | p_33 Mediana (iqr) | p_29 Mediana (iqr) | | | | |
| 1 | 1606 | 4(1) | 5(1) | 5(2) | F 695 M 955 N 56 | 9.07(2.09) | 7.09(1.88) | 1.99(2.2) |
| 2 | 834 | 3(1) | 4(0) | 3(3) | F 363 M 448 N 23 | 8.68(1.88) | 6.7(1.7) | 1.98(2.15) |
| 3 | 1319 | 6(1) | 7(1) | 6(2) | F 483 M 788 N 48 | 9.48(2.09) | 7.38(2.1) | 2.11(2.28) |
| 4 | 511 | 1(1) | 2(2) | 2(3) | F 267 M 220 N 24 | 8.37(2.06) | 6.55(1.57) | 1.82(2.27) |
| 5 | 645 | 6(2) | 6(2) | 1(1) | F 227 M 395 N 23 | 8.51(2.04) | 6.6(1.91) | 1.91(2.23) |

En la tabla 17 podemos ver como los grupos 1 y 2 están compuestos por alumnos para los que el prestigio de la titulación y de la universidad es importante en cierto grado, más importante para los del grupo 1 que para los del 2.

En el grupo 3 se agrupan los alumnos que dan mucha importancia al prestigio a la hora de seleccionar titulación y universidad, esto alumnos tienen una nota de acceso sensiblemente más alta que el resto y acceden a la titulación con más diferencia de nota con respecto a la nota de corte.

El grupo 4 está compuesto por alumnos que no dan nada de importancia al prestigio de la titulación y de la universidad. En la tabla vemos que estos alumnos tienen una nota de acceso menor que el resto.

Y por último, en el grupo 5 tenemos a alumnos que dan mucha importancia al prestigio de la titulación, por contra, el prestigio del a universidad no influye en nada en su decisión.

Resumiendo y quedándonos con los perfiles más significativos, según los resultados podemos afirmar que existen 3 de perfiles claramente marcados.

Un perfil en el que los alumnos afirman que el prestigio de la titulación y de la universidad es un aspecto muy influyente en el momento de la elección.

Un perfil totalmente opuesto al anterior donde los alumnos niegan influencia alguna de estos factores en el momento de la elección.

Y un último perfil donde los alumnos afirman dar más valor al prestigio de la titulación por encima del de la universidad, que no lo consideran nada importante.

Para poder compararlos con los perfiles encontrados por ramas, llamaremos a estos perfiles de la siguiente manera:

- P1: Prestigio titulación y prestigio Universidad muy importantes.
- P2: Prestigio titulación y prestigio Universidad poco importantes.
- P3: Prestigio titulación muy importante, prestigio Universidad nada importante.

Continuamos el trabajo con estos factores realizando el análisis por ramas para ver si encontramos estos perfiles a nivel de rama.

## Rama Ingeniería y Arquitectura

Número de grupos determinado por el método Elbow y dendrograma: 5.

**Tabla 18**

| Grupo | n | Prestigio social de la titulación | | Calidad y prestigio de la universidad | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|
| | | p_32 Mediana (iqr) | p_33 Mediana (iqr) | p_29 Mediana (iqr) | | | | |
| 1 | 369 | 2(2) | 3(3) | 3(3) | F 123 M 231 N 15 | 8.42(1.95) | 6.54(1.53) | 1.88(2.04) |
| 2 | 872 | 6(1) | 6(2) | 5(1) | F 231 M 616 N 25 | 9.16(2.07) | 7.22(2.04) | 1.95(2.1) |
| 3 | 417 | 7(1) | 7(1) | 7(1) | F 118 M 287 N 12 | 9.33(2.02) | 7.57(1.97) | 1.76(1.87) |
| 4 | 581 | 5(2) | 5(2) | 1(1) | F 158 M 412 N 11 | 8.32(1.99) | 6.35(1.66) | 1.97(2.09) |
| 5 | 880 | 4(1) | 4(1) | 4(1) | F 295 M 556 N 29 | 8.82(2.04) | 6.97(1.79) | 1.85(1.9) |

En Ingeniería y Arquitectura encontramos los mismos perfiles que en el análisis sobre la totalidad de alumnos a diferencia del perfil P2 (Prestigio titulación y prestigio Universidad poco importantes) donde encontramos valores ligeramente superiores.

### Rama Artes y Humanidades

Número de grupos determinado por el método Elbow y dendrograma: 4.

**Tabla 19**

| Grupos | n | Prestigio social de la titulación | | Calidad y prestigio de la universidad | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|
| | | p_32 Mediana (iqr) | p_33 Mediana (iqr) | p_29 Mediana (iqr) | | | | |
| 1 | 31 | 7(2) | 6(3) | 1(0.5) | F21 M10 | 8.98(1.59) | 6.08(1.37) | 2.9(1.9) |
| 2 | 83 | 5(3) | 7(1) | 6(2) | F52 M31 | 9.37(2.13) | 6.12(1.55) | 3.2(2.83) |
| 3 | 115 | 3(1) | 4(1) | 4(1) | F85 M30 | 9.1(1.77) | 6.51(1.49) | 2.6(2.3) |
| 4 | 82 | 1(1) | 2(1) | 2(3) | F62 M20 | 8.62(1.98) | 6.66(1.49) | 1.95(2.01) |

En Artes y Humanidades también encontramos los mismos perfiles que en el análisis general.

### Rama Ciencias

Número de grupos determinado por el método Elbow y dendrograma: 5.

**Tabla 20**

| Grupos | n | Prestigio social de la titulación | | Calidad y prestigio de la universidad | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|
| | | p_32 Mediana (iqr) | p_33 Mediana (iqr) | p_29 Mediana (iqr) | | | | |
| Grupo 1 | 50 | 4(1.75) | 4.5(1) | 4(1) | F30 M20 | 9.91(2.07) | 8.67(2.13) | 1.24(2.2) |
| Grupo 2 | 40 | 7(1) | 7(0) | 7(2) | F18 M22 | 11.14(1.8) | 9.71(2) | 1.43(1.72) |
| Grupo 3 | 55 | 5(1) | 6(1) | 5(1) | F36 M19 | 11.37(1.5) | 9.71(2.2) | 1.66(2.084) |
| Grupo 4 | 34 | 3(1) | 4(1) | 6(1) | F17 M17 | 10.53(1.6) | 8.91(1.87) | 1.62(1.74) |
| Grupo 5 | 26 | 2(1) | 2(2) | 2(3) | F11 M15 | 9.04(2.20) | 6.93(1.64) | 2.10(2.23) |

En Ciencias volvemos a encontrar los perfiles P1 (Prestigio titulación y prestigio Universidad muy importantes) y P2 (Prestigio titulación y prestigio Universidad poco importantes) pero hay una diferencia clara con respecto a P3.

No encontramos ningún perfil con esos valores, en cambio sí que encontramos un perfil opuesto donde los alumnos indican que el prestigio de la universidad ha sido determinante en su elección mientras que el prestigio de la titulación les influyó menos.

De aquí en adelante, a este nuevo perfil vamos a llamarlo de la siguiente manera:

- P4: Prestigio Universidad muy importante, prestigio titulación poco importante.

### Rama Ciencias de la Salud

Número de grupos determinado por el método Elbow y dendrograma: 5.

**Tabla 21**

| Grupos | n | Prestigio social de la titulación | | Calidad y prestigio de la universidad | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|
| | | p_32 Mediana (iqr) | p_33 Mediana (iqr) | p_29 Mediana (iqr) | | | | |
| 1 | 74 | 6(2) | 6(2) | 1(1) | F41 M20 N 13 | 8.67(1.90) | 7.87(1.98) | 0.80(2.67) |
| 2 | 91 | 7(1) | 7(0) | 6(2) | F38 M35 N18 | 9.82(2.10) | 7.62(1.95) | 2.20(2.65) |
| 3 | 80 | 1(1) | 2.5(2) | 2.5(3) | F42 M24 N14 | 8.49(2.24) | 7.61(1.90) | 0.87(3.08) |
| 4 | 144 | 4(1) | 6(1.25) | 5(1) | F80 M45 N19 | 9.37(2.31) | 7.90(2.05) | 1.47(3.09) |
| 5 | 133 | 4(1) | 4(1) | 4(1) | F73 M42 N18 | 8.95(1.86) | 7.54(1.81) | 1.41(2.63) |

En Ciencias de la Salud encontramos los mismos perfiles que en el análisis general.

## Rama Ciencias Sociales y Jurídicas

Número de grupos determinado por el método Elbow y dendrograma: 4.

**Tabla 22**

| Grupos | n | Prestigio social de la titulación | | Calidad y prestigio de la universidad | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|
| | | p_32 Mediana (iqr) | p_33 Mediana (iqr) | p_29 Mediana (iqr) | | | | |
| 1 | 111 | 5(3) | 5(2) | 1(1) | F82 M29 | 8.07(1.91) | 5.98(1.33) | 2.09(2.26) |
| 2 | 121 | 1(1) | 2(2) | 4(3) | F85 M36 | 8.46(2.05) | 6.12(1.30) | 2.33(2.29) |
| 3 | 230 | 6(2) | 7(1) | 6(2) | F143 M87 | 9.27(2.15) | 6.32(1.54) | 2.95(2.57) |
| 4 | 296 | 4(1) | 5(1) | 4(1) | F194 M102 | 9.21(1.88) | 6.43(1.54) | 2.78(2.16) |

En Ciencias Sociales y Jurídicas volvemos a encontrar los perfiles P1 (Prestigio titulación y prestigio Universidad muy importantes) y P2 (Prestigio titulación y prestigio Universidad poco importantes) pero no encontramos ningún perfil similar a P3 (Prestigio titulación muy importante, prestigio Universidad nada importante).

En cambio, como en la rama de Ciencias, volvemos a encontrar un perfil donde los alumnos indican que el prestigio de la universidad ha influido en su elección mientras que no tuvieron nada en cuenta el prestigio de la titulación. Este perfil sería como P4 (Prestigio Universidad muy importante, prestigio titulación nada importante) pero en menor grado.

### *Empleabilidad vs Vocación*

Realizamos el análisis agrupando por las preguntas relacionadas a los dos factores indicados.

**Tabla 23**

| Factor | Pregunta |
|---|---|
| 2 | p_38 - Creo que existe demanda de trabajo en el sector |
| | p_39 - Es más fácil encontrar trabajo |
| 3 | p_23 - Siempre he querido estudiar esta carrera |
| | p_24 - Fue una decisión de última hora |

Determinamos el número idóneo de clusters:

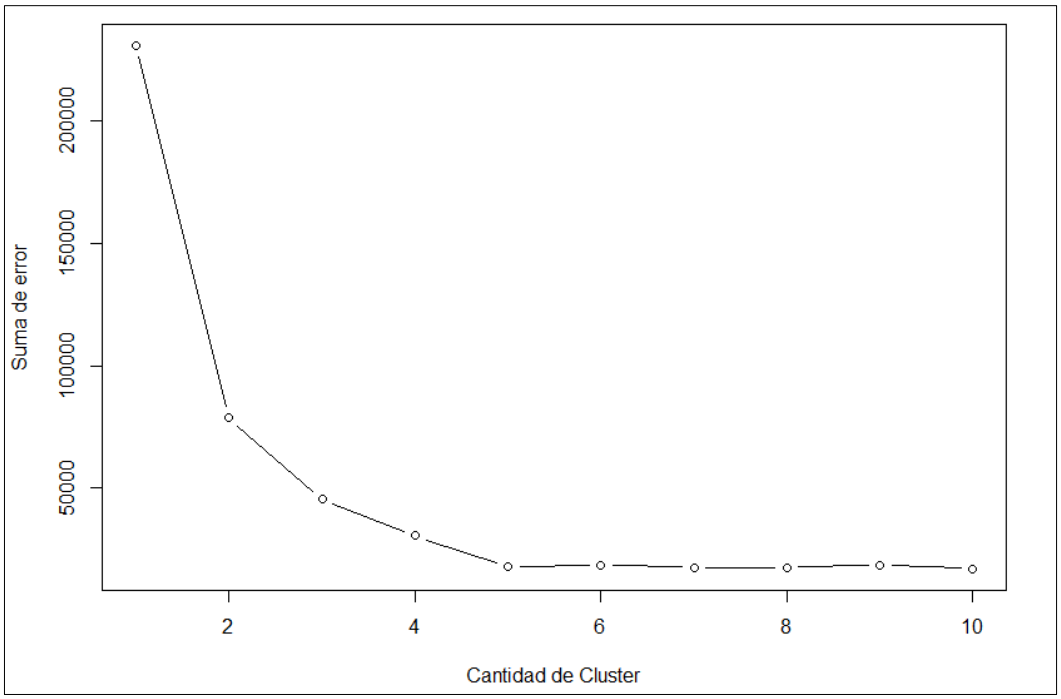


**Figura 31**

En la gráfica vemos cómo a partir de 5 clusters la reducción de la suma de error no es significativa. En el siguiente dendrograma mostramos los grupos definidos.

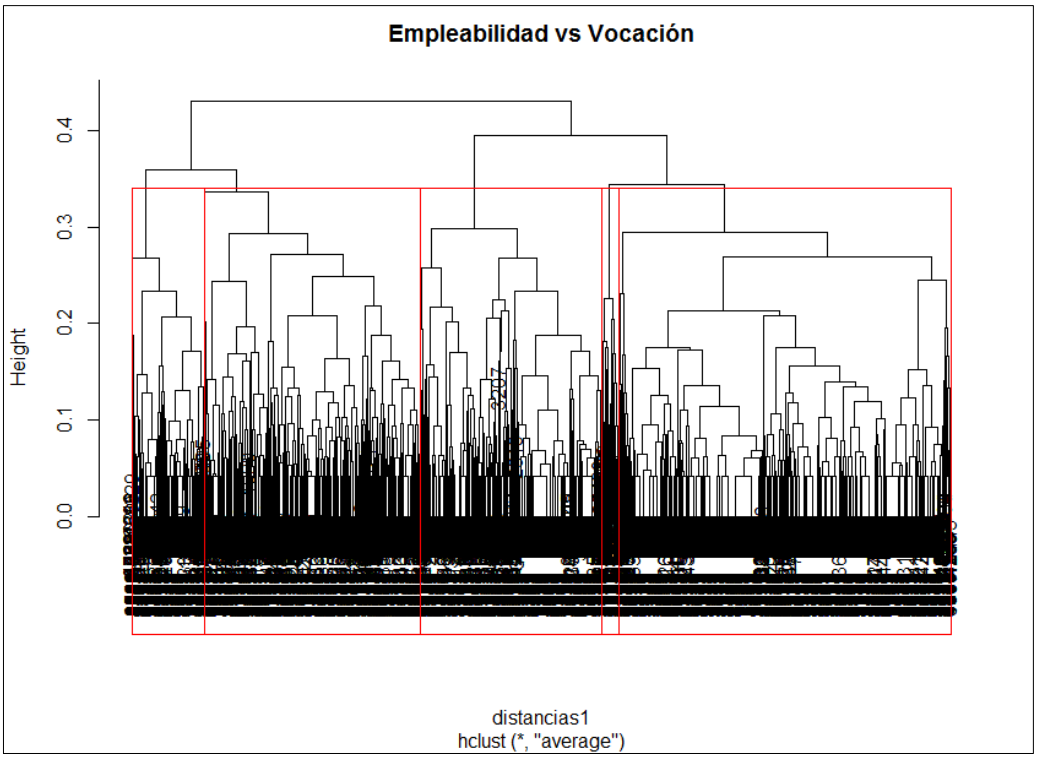


**Figura 32**

Agrupamos con k = 5 y caracterizamos los clusters:

**Tabla 24**

| Grupo | n | Empleabilidad | | Vocación | | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|---|
| | | p_38 Mediana (iqr) | p_39 Mediana (iqr) | p_23 Mediana (iqr) | p_24 Mediana (iqr) | | | | |
| 1 | 1015 | 3(3) | 2(2) | 6(2) | 1(0) | F 471<br>M 487<br>N 57 | 8.75(2.06) | 7.01(1.84) | 1.75(2.25) |
| 2 | 661 | 2(2) | 1(1) | 2(3) | 5(3) | F 375<br>M 255<br>N 31 | 8.67(2.07) | 6.76(1.75) | 1.91(2.24) |
| 3 | 839 | 5(3) | 5(2) | 2(2) | 6(2) | F 335<br>M 489<br>N 15 | 9.17(2.07) | 7.15(2.06) | 2.02(2.07) |
| 4 | 1402 | 4(1) | 4(1) | 4(1) | 3(2) | F 556<br>M 802<br>N 44 | 9.07(2.05) | 6.98(1.9) | 2.09(2.24) |
| 5 | 998 | 6(2) | 6(1.75) | 6(2) | 1(0) | F 298<br>M 673<br>N 27 | 9.08(2.13) | 6.96(1.96) | 2.12(2.27) |

En este caso, los 5 grupos difieren entre sí. En el grupo 1 encontramos a los alumnos para los que los aspectos vocacionales han determinado su elección de titulación mientras que la empleabilidad no les ha influido.

Los alumnos que forman el grupo 2 que no dan prácticamente importancia a aspectos relacionados con la empleabilidad y la vocación, e indican que tomaron la decisión a última hora. Analizando las variables observables, vemos que estos alumnos son los que menor nota de acceso tienen y que la nota media de corte de sus titulaciones es también la menor.

En el grupo 3 nos encontramos a alumnos para los cuales el factor empleabilidad es muy influyente mientras que aspectos vocaciones no les ha influido y la decisión de qué titulación cursar fue tomada a última hora. Estos alumnos son los que tienen una nota de acceso mayor a las de los otros grupos y acceden también a titulaciones con mayor nota de corte.

En el grupo 4, los alumnos tienen valores medios en ambos factores. Indican una influencia parcial de aspectos relacionados con la empleabilidad y vocacionales.

Por último, en el grupo 5 tenemos a alumnos que indican que en la elección les han influido mucho ambos factores, estos alumnos acceden con una nota alta a titulaciones con una nota de corte media (6.96).

Resumiendo y quedándonos con los perfiles más significativos, según los resultados podemos afirmar que existen 4 de perfiles claramente marcados.

Un perfil en el que los alumnos afirman que aspectos vocacionales han sido los que les han llevado a hacer esa elección sin tener en cuenta aspectos relacionados con la empleabilidad.

Un perfil opuesto al anterior donde los alumnos indican que en la elección les influyó mucho la empleabilidad y sin tener en cuenta aspectos vocacionales.

Un perfil donde ni la vocación ni la empleabilidad influyeron a los alumnos en el momento de la elección.

Y un último perfil donde ambos factores influyeron a los alumnos a la hora de tomar decisión.

Para poder compararlos con los perfiles encontrados por ramas, llamaremos a estos perfiles de la siguiente manera:

- P1: Aspectos vocacionales importantes, empleabilidad poco importante.
- P2: Empleabilidad importante, aspectos vocacionales poco importantes.
- P3: Empleabilidad y aspectos vocacionales poco importantes.
- P4: Empleabilidad y aspectos vocacionales muy importantes.

Continuamos el trabajo con estos factores realizando el análisis por ramas para ver si encontramos estos perfiles a nivel de rama.

### Rama Ingeniería y Arquitectura

Número de grupos determinado por el método Elbow y dendrograma: 5.

**Tabla 25**

| Grupo | n | Empleabilidad | | Vocación | | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|---|
| | | p_38 Mediana (iqr) | p_39 Mediana (iqr) | p_23 Mediana (iqr) | p_24 Mediana (iqr) | | | | |
| 1 | 796 | 6(2) | 5(2) | 6(2) | 1(0) | F158 M622 N 16 | 8.82(2.03) | 6.92(1.89) | 1.90(2.04) |
| 2 | 373 | 2(2) | 2(1) | 6(2) | 1(0) | F133 M213 N 27 | 8.61(2.06) | 6.94(1.79) | 1.67(1.94) |
| 3 | 417 | 3(2) | 2(2) | 2(2) | 5(3) | F177 M222 N 18 | 8.75(2.05) | 6.78(1.76) | 1.97(2.02) |
| 4 | 669 | 6(2) | 5(2) | 2(3) | 5(2) | F222 M436 N 11 | 9.07(2.10) | 7.14(2.04) | 1.93(2.02) |
| 5 | 864 | 4(1) | 4(2) | 4(1) | 2(2) | F235 M609 N 20 | 8.83(2.05) | 6.93(1.83) | 1.90(1.99) |

En Ingeniería y Arquitectura encontramos los mismos perfiles que en el análisis general.

## Rama Artes y Humanidades

Número de grupos determinado por el método Elbow y dendrograma: 4.

**Tabla 26**

| Grupo | n | Empleabilidad | | Vocación | | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|---|
| | | p_38 Mediana (iqr) | p_39 Mediana (iqr) | p_23 Mediana (iqr) | p_24 Mediana (iqr) | | | | |
| 1 | 45 | 4(2) | 4(2) | 2(3) | 6(3) | F27 M18 | 9.28(1.87) | 5.77(1.37) | 3.51(2.02) |
| 2 | 85 | 4(1) | 4(2) | 6(3) | 1(1) | F57 M28 | 9.52(1.86) | 6.18(1.42) | 3.33(2.41) |
| 3 | 113 | 2(2) | 1(1) | 6(2) | 1(1) | F86 M27 | 8.75(1.95) | 6.80(1.52) | 1.96(2.37) |
| 4 | 68 | 1(1) | 1(1) | 3(3) | 4.5(2.25) | F50 M18 | 8.73(1.88) | 6.45(1.50) | 2.29(2.24) |

En Artes y Humanidades volvemos a encontrar los perfiles P1 (Aspectos vocacionales importantes, empleabilidad poco importante) y P3 (Empleabilidad y aspectos vocacionales poco importantes) pero hay diferencias con respecto a P2 y P4.

Encontramos un perfil similar a P2 (Empleabilidad importante, aspectos vocacionales poco importantes) pero con valores significativamente más bajos en la influencia de la empleabilidad.

En cuanto a P4 (Empleabilidad y aspectos vocacionales muy importantes), los valores referentes al factor de empleabilidad también son más bajos que en el análisis general.

## Rama Ciencias

Número de grupos determinado por el método Elbow y dendrograma: 4.

**Tabla 27**

| Grupo | n | Empleabilidad | | Vocación | | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|---|
| | | p_38 Mediana (iqr) | p_39 Mediana (iqr) | p_23 Mediana (iqr) | p_24 Mediana (iqr) | | | | |
| 1 | 34 | 3(2) | 2(1) | 1(1) | 6(3) | F22 M12 | 10.24(2.11) | 8.41(2.22) | 1.84(2.06) |
| 2 | 49 | 4(2) | 3(1) | 6(3) | 1(1) | F26 M23 | 10.55(2.02) | 8.69(2.46) | 1.85(2.18) |
| 3 | 50 | 6(2) | 5(1.75) | 5(2) | 1(1) | F24 M26 | 10.70(2.03) | 9.24(2.07) | 1.46(2.00) |
| 4 | 72 | 5(2) | 4(1) | 3(3) | 5(2) | F40 M32 | 10.55(1.86) | 9.25(2.08) | 1.30(1.87) |

En Ciencias encontramos de manera clara los perfiles P3 (Empleabilidad y aspectos vocacionales poco importantes) y P4 (Empleabilidad y aspectos vocacionales muy importantes).

Para P1 (Aspectos vocacionales importantes, empleabilidad poco importante) tenemos valores más altos en aspectos relacionados con la empleabilidad, lo que nos deja un perfil con cierta influencia de la empleabilidad y mucha influencia de aspectos vocacionales.

Para P2 (Empleabilidad importante, aspectos vocacionales poco importantes) tenemos valores más suaves en ambos factores. Este perfil de alumnos considera bastante influyente la empleabilidad y aspectos vocacionales influyen poco sin llegar a ser extremos.

### Rama Ciencias de la Salud

Número de grupos determinado por el método Elbow y dendrograma: 4.

**Tabla 28**

| Grupo | n | Empleabilidad | | Vocación | | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|---|
| | | p_38 Mediana (iqr) | p_39 Mediana (iqr) | p_23 Mediana (iqr) | p_24 Mediana (iqr) | | | | |
| 1 | 134 | 2(2) | 1(1) | 6(3) | 1(1) | F69 M46 N19 | 8.64(2.35) | 7.92(1.99) | 0.72(2.91) |
| 2 | 77 | 1(1) | 1(1) | 2(2) | 6(3) | F40 M24 N13 | 8.90(2.17) | 7.41(1.78) | 1.49(2.91) |
| 3 | 109 | 4(2) | 4(1) | 2(2) | 4(3) | F58 M37 N14 | 8.96(1.87) | 7.81(1.97) | 1.14(2.77) |
| 4 | 202 | 5(2) | 4(1) | 5(3) | 1(1) | F107 M59 N36 | 9.58(2.03) | 7.63(1.95) | 1.95(2.78) |

En Ciencias de la Salud volvemos a encontrar los perfiles P1 (Aspectos vocacionales importantes, empleabilidad poco importante) y P3 (Empleabilidad y aspectos vocacionales poco importantes) pero hay diferencias con respecto a P2 y P4.

Encontramos un perfil similar a P2 (Empleabilidad importante, aspectos vocacionales poco importantes) pero con valores significativamente más bajos en el factor empleabilidad.

En cuanto a P4 (Empleabilidad y aspectos vocacionales muy importantes), los valores en ambos son más bajos que en el análisis general.

### Rama Ciencias Sociales y Jurídicas

Número de grupos determinado por el método Elbow y dendrograma: 4.

**Tabla 29**

| Grupo | n | Empleabilidad | | Vocación | | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|---|
| | | p_38 Mediana (iqr) | p_39 Mediana (iqr) | p_23 Mediana (iqr) | p_24 Mediana (iqr) | | | | |
| 1 | 127 | 2(3) | 1(2) | 1(1) | 6(2) | F97 M30 | 8.68(2.09) | 6.08(1.31) | 2.60(2.41) |
| 2 | 201 | 4(1) | 4(1) | 3(2) | 3(2) | F134 M67 | 9.16(2.02) | 6.37(1.56) | 2.79(2.38) |
| 3 | 80 | 7(0.25) | 7(1) | 4(3) | 3(3.25) | F42 M38 | 8.89(1.93) | 6.45(1.64) | 2.44(2.06) |
| 4 | 199 | 5(2) | 4(1) | 5(2) | 1(0) | F119 M80 | 9.41(1.95) | 6.49(1.53) | 2.93(2.32) |
| 5 | 151 | 1(1) | 1(1) | 5(3) | 1(2) | F112 M39 | 8.29(2.05) | 5.98(1.30) | 2.31(2.38) |

En Ciencias Sociales y Jurídicas volvemos a encontrar los perfiles P1 (Aspectos vocacionales importantes, empleabilidad poco importante) y P3 (Empleabilidad y aspectos vocacionales poco importantes) pero hay diferencias con respecto a P2 y P4.

Encontramos un perfil similar a P2 (Empleabilidad importante, aspectos vocacionales poco importantes) pero con valores significativamente más bajos en el factor empleabilidad.

En cuanto a P4 (Empleabilidad y aspectos vocacionales muy importantes), los valores del factor vocacional son significativamente inferiores, indicando el factor como parcialmente influyente.

### *Calidad y prestigio de la universidad vs Vocación*

Realizamos el análisis agrupando por las preguntas de los dos factores.

**Tabla 30**

| Factor | Pregunta |
|---|---|
| 3 | p_23 - Siempre he querido estudiar esta carrera |
| | p_24 - Fue una decisión de última hora |
| 4 | p_29 - La calidad y el prestigio de la Universidad fueron determinantes en mi elección |

Determinamos el número idóneo de clusters:

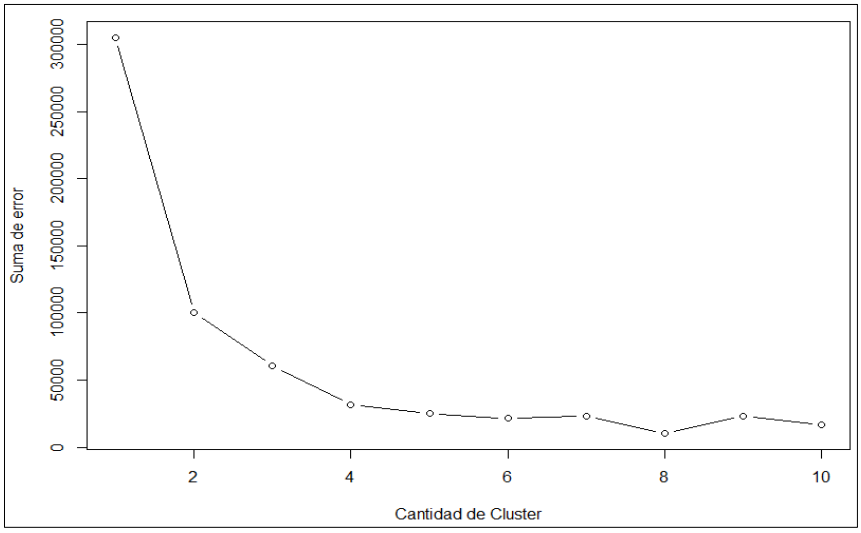


**Figura 33**

En la gráfica vemos cómo a partir de 4 clusters la reducción de la suma de error no es significativa. Mostramos los grupos definidos en un dendrograma.

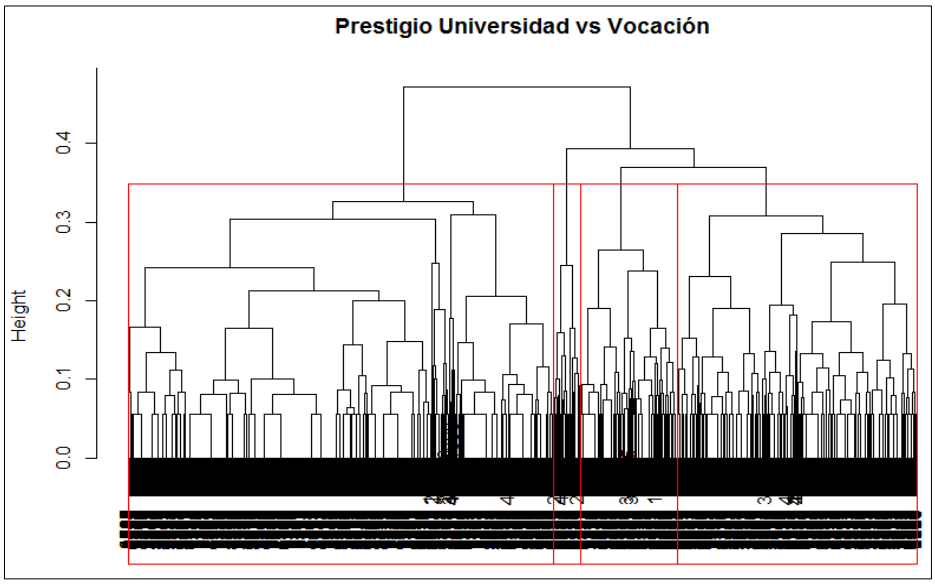


**Figura 34**

Agrupamos con k = 4 y caracterizamos los clusters:

**Tabla 31**

| Grupo | n | Calidad y prestigio de la universidad | Vocación | | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|
| | | p_29 Mediana (iqr) | p_23 Mediana (iqr) | p_24 Mediana (iqr) | | | | |
| 1 | 861 | 2(2) | 5(3) | 1(1) | F 340 M 489 N 32 | 8.44(1.95) | 6.57(1.76) | 1.87(2.27) |
| 2 | 1051 | 2(3) | 2(3) | 4(2) | F 461 M 560 N 30 | 8.66(2.07) | 6.67(1.82) | 1.99(2.22) |
| 3 | 2054 | 5(2) | 6(3) | 1(1) | F 795 M1176 N 83 | 9.19(2.09) | 7.14(1.91) | 2.05(2.24) |
| 4 | 949 | 6(1) | 2(3) | 6(3) | F 439 M 481 N 29 | 9.32(2.04) | 7.35(2.02) | 1.97(2.15) |

En este caso también vemos que los 4 grupos están claramente diferenciados.

En el grupo 1 se agrupan los alumnos que priman aspectos vocacionales frente a la calidad y prestigio de la universidad, estos alumnos acceden con notas más bajas a titulaciones con notas de corte también más bajas que en el resto de grupos.

En el grupo 2 tenemos a alumnos para los que ni la vocación, ni la calidad y prestigio de la universidad son importantes a la hora de elegir titulación. Las notas de acceso de estos alumnos y las notas de corte de las titulaciones a las que acceden son algo más altas que los del grupo 1, pero bastante más bajas que las del resto de grupos.

Los alumnos del grupo 3 indican que ambos factores han sido influyentes en la elección.

Por último, los alumnos del grupo 4 indican que la calidad y prestigio de la universidad ha sido determinante y que tomaron la decisión a última hora sin tener en cuenta aspectos vocacionales. Estos alumnos tienen las notas más altas de acceso y las titulaciones elegidas tienen mayor nota de corte que el resto de grupos.

### *Prestigio social de la titulación vs Vocación*

Realizamos el análisis agrupando por las preguntas relacionadas a los dos factores indicados.

**Tabla 32**

| Factor | Pregunta |
|---|---|
| 1 | p_32 - El título está prestigiado socialmente |
| | p_33 - El título es reconocido internacionalmente |
| 3 | p_23 - Siempre he querido estudiar esta carrera |
| | p_24 - Fue una decisión de última hora |

Determinamos el número idóneo de clusters:

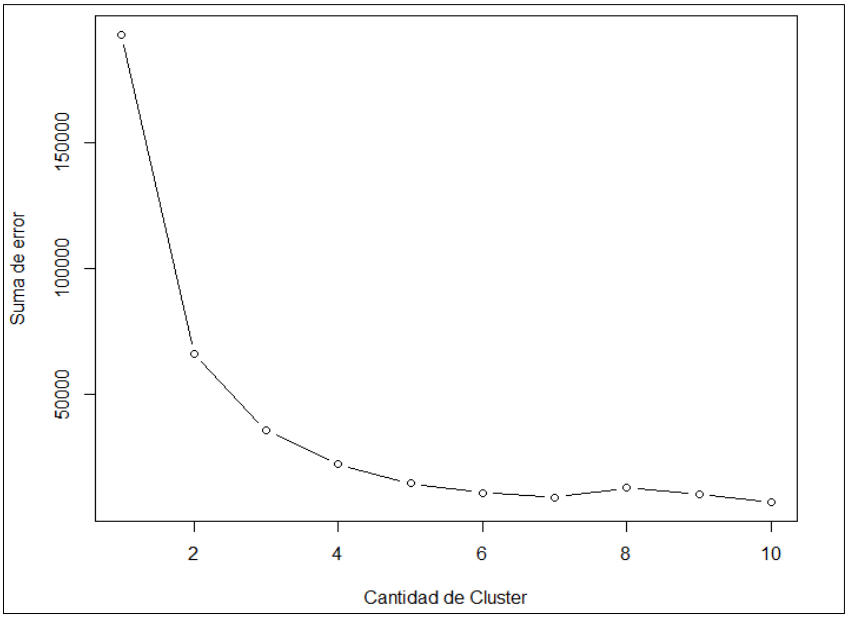


**Figura 35**

En la gráfica vemos cómo a partir de 4 clusters la reducción de la suma de error no es significativa. En el siguiente dendrograma mostramos los grupos definidos.

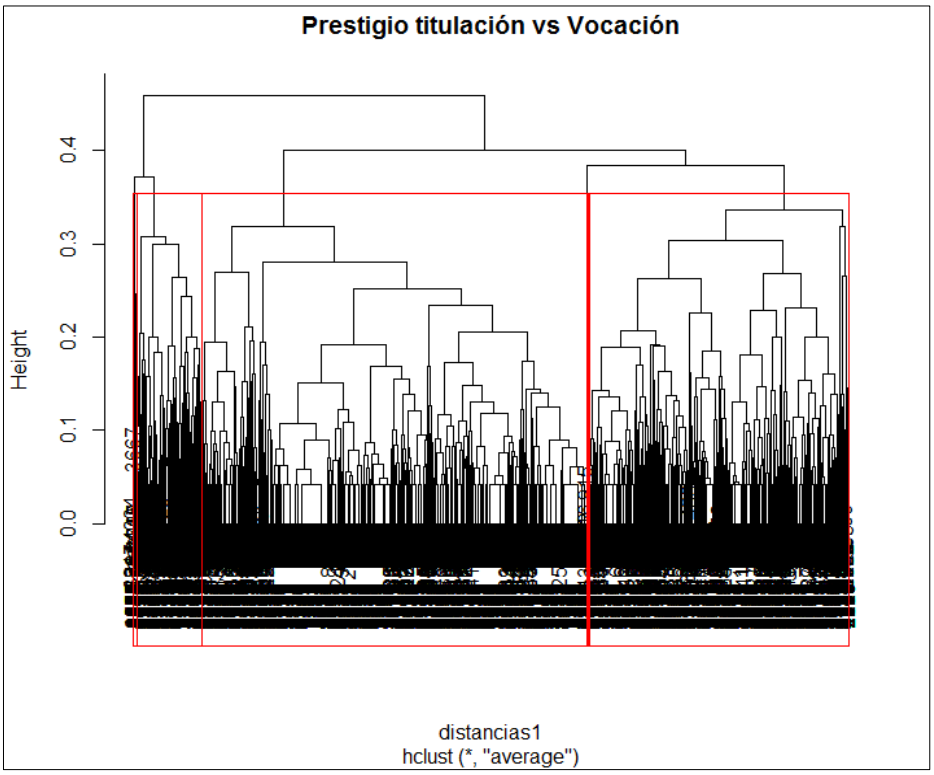


**Figura 36**

Agrupamos con k = 5 y caracterizamos los clusters:

**Tabla 33**

| Grupo | n | Prestigio social de la titulación | | Vocación | | Sexo | Nota Acceso Media (dt) | Nota Corte Media (dt) | Dif. Notas Media (dt) |
|---|---|---|---|---|---|---|---|---|---|
| | | p_32 Mediana (iqr) | p_33 Mediana (iqr) | p_23 Mediana (iqr) | p_24 Mediana (iqr) | | | | |
| 1 | 1334 | 4(1) | 5(1) | 5(2) | 1(1) | F 502<br>M 780<br>N 52 | 8.93(2.04) | 6.91(1.82) | 2.02(2.22) |
| 2 | 1243 | 5(2) | 5(2) | 3(2) | 4(2) | F 524<br>M 683<br>N 36 | 8.87(2.08) | 6.99(1.96) | 1.88(2.15) |
| 3 | 845 | 4(3) | 5(4) | 1(1) | 7(1) | F 413<br>M 407<br>N 25 | 9.16(2.09) | 7.01(1.94) | 2.15(2.23) |
| 4 | 981 | 7(1) | 7(1) | 6(2) | 1(0) | F 349<br>M 584<br>N 38 | 9.22(2.14) | 7.19(2.06) | 2.03(2.28) |
| 5 | 512 | 1(1) | 3(2) | 5(3) | 1(2) | F 247<br>M 242<br>N 23 | 8.53(1.97) | 6.69(1.61) | 1.84(2.3) |

El grupo 1 agrupa a alumnos que indican cierta influencia del prestigio social de la titulación y también indican que aspectos vocacionales han influenciado mucho en su elección.

En el grupo 2 observamos que el prestigio de la titulación ha tenido bastante influencia en la elección mientras que los aspectos vocacionales no han tenido tanta y presentan valores medios - bajos.

En el grupo 3 tenemos a alumnos que indican que la elección fue una decisión de última hora, sin atender a aspectos vocacionales, y que el prestigio de la titulación tuvo cierta influencia.

Los alumnos del grupo 4 indican que el prestigio de la titulación y aspectos vocacionales han sido claves en su elección, estos alumnos son los que presentan mayor nota media de acceso y las titulaciones elegidas tienen mayor nota de corte que las de los otros grupos.

Por último, en el grupo 5 tenemos a alumnos para los que el prestigio de la titulación no tiene importancia y lo que les ha llevado a elegir la titulación son aspectos vocacionales. La nota media de acceso de estos alumnos es la más baja de todos los grupos y la nota de corte de las titulaciones elegidas es también la menor de todos los grupos.

# 5. Conclusiones

## 5.1. Conclusiones

Nuestro objetivo era analizar qué factores influyen en la toma de decisiones de los estudiantes en el momento de elegir estudios universitarios y buscar qué perfiles de los alumnos existen con respecto a la influencia de dichos factores.

El objeto de estudio ha sido un cuestionario desarrollado a partir de ~~en~~ una tesis doctoral donde se definían estos factores y un TFM donde se reducen el número de factores para determinar las preguntas significativas. Con esto, disponíamos de los datos y variables necesarias para realizar nuestro análisis.

De todos los factores identificados en los trabajos mencionados, nosotros nos quedamos con 4:

- Aspectos vocacionales
- Empleabilidad
- Prestigio social de la titulación
- Calidad y prestigio de la Universidad

Se han estudiado estos 4 factores y los hemos enfrentado 1 vs 1 para agrupar a los alumnos en base a ellos. Para profundizar, hemos analizado el comportamiento por ramas en los análisis "Vocación vs Empleabilidad" y "Prestigio social de la titulación vs Calidad y prestigio de la Universidad".

A partir de estos análisis hemos definido los perfiles existentes según los factores analizados añadiendo como variables observables el sexo, la nota de acceso y la nota de corte. En base a esto hemos constatado que existen diferencias en estos perfiles entre las distintas ramas de estudio.

En los análisis **"Calidad y prestigio de la Universidad vs Vocación"** y **"Prestigio de la titulación vs Vocación"** hemos detectado la existencia de 4 perfiles: alta influencia de ambos factores, alta influencia de un factor y baja del otro; y baja influencia de ambos factores. En ninguno de los perfiles detectados en ambos análisis se observan diferencias por género. Cabe destacar de los resultados del análisis "Calidad y prestigio de la Universidad vs Vocación" la manera en la que el perfil que indica Alta influencia del prestigio de la universidad y baja influencia de aspectos vocacionales determina que elegir una titulación u otra fue una decisión de última hora.

En cuanto a las notas de acceso y corte, los resultados indican que los alumnos cuya influencia principal es el prestigio de la universidad por encima de aspectos vocacionales tienen notas de acceso más altas y se matriculan en titulaciones con notas de corte también más altas que el resto de los perfiles. Por contra, el perfil con menor nota de acceso y de corte se corresponde con aquellos alumnos para los que los aspectos vocacionales son la principal influencia en el momento de la elección, sin dar importancia al prestigio de la universidad.

En el análisis **"Prestigio de la titulación vs Vocación"** el perfil con mayor nota de acceso y de corte es el definido con valores altos en ambos factores. Por contra, los valores más bajos los tiene el perfil donde los aspectos vocacionales son la principal influencia sin dar importancia al prestigio de la titulación.

El siguiente análisis realizado ha sido **"Prestigio social de la titulación vs Calidad y prestigio de la Universidad"** y nos aporta 3 perfiles distintos: alumnos que consideran muy influyente el prestigio de la titulación y de la universidad, alumnos que no consideran influyentes estos factores y alumnos que dan mucho valor al prestigio de la titulación y muy poco al de la universidad. En lo que respecta a las notas de acceso y corte, los alumnos con valores altos en ambos factores presentan una nota de acceso y corte mayor que el resto. Todo lo contrario a los alumnos con valores bajos en ambos factores, cuya nota de acceso y corte son las menores.

En el análisis por ramas de esos factores detectamos un nuevo perfil en las ramas "Ciencias" y "Ciencias Sociales y Jurídicas" donde los alumnos dan mucho valor al prestigio de la universidad por encima de la titulación.

Atendiendo a las variables nota de acceso y nota de corte, el comportamiento en "Artes y Humanidades" y en "Ciencias de la salud" es similar a los resultados generales aunque se detectan variaciones en los valores de nota de corte. El perfil con valores altos en ambos factores presenta una nota de corte menor que el resto, y el perfil con valores bajos en ambos factores presenta una nota de corte mayor que el resto.

En "Ingeniería y Arquitectura" y "Ciencias Sociales" se detecta que los alumnos con menor nota de acceso y nota de corte se corresponden con los alumnos del perfil el cual el prestigio de la titulación es muy importante y el prestigio de la universidad no influye en su elección.

En "Ciencias", existen dos perfiles con valores altos en ambos factores y las notas de acceso y corte altas se corresponden el perfil más moderado, siendo el perfil con valores más altos el segundo en cuanto a estas notas.

Por último, el análisis "Empleabilidad vs Vocación" nos aporta de nuevo los 4 perfiles que cabría esperar combinando la influencia o no de los dos factores. Es

en el análisis por ramas donde detectamos variaciones en todas las ramas excepto "Ingeniería y Arquitectura" que coincide con la rama con mayor número de observaciones.

Con respecto a las notas de acceso y de corte de cada uno de los perfiles, en el análisis general la mayor nota de acceso y de corte se corresponde con valores altos en el factor empleabilidad y bajos en aspectos vocacionales. En el estudio por ramas vemos cómo en las ramas "Ingeniería y Arquitectura", "Ciencias de la Salud" y "Ciencias Sociales y Jurídicas" las notas de acceso y corte altas se corresponden también con el perfil indicado para el análisis general. En cambio, para las notas de acceso y corte bajas el perfil de los alumnos indica valores altos en aspectos vocacionales y bajos en empleabilidad.

El comportamiento en la rama "Ciencias" difiere con respecto al análisis general. Las notas de acceso y corte más altas corresponden al perfil con valores más altos en ambos factores, mientras que las bajas se corresponden con valores más bajos en ambos factores.

Como último detalle, en la rama "Artes y Humanidades" vemos cómo las notas de acceso y corte más bajas se relacionan con el mismo perfil que en el análisis general mientras que las más altas corresponden al perfil con valores altos en aspectos vocacionales y medios en aspectos relacionados con la empleabilidad.

Con todo esto, podemos concluir que se ha conseguido el objetivo propuesto al inicio del trabajo y hemos podido identificar y categorizar los perfiles y alumnos con respecto a los factores que influyen en su elección de estudios universitarios.

## 5.2. Trabajos futuros

En posibles desarrollos futuros sobre esta línea de trabajo se podría valorar trabajar sobre los siguientes aspectos:

- Profundizar en el estudio por titulaciones para ver las características concretas de los alumnos de una determinada titulación y si coincide con el pensamiento general y lo que se presupone sobre estos.
- Incorporar resultados académicos de los alumnos encuestados en sus titulaciones y repetir el análisis para así poder prever la progresión de un alumno de un perfil concreto.
- Realizar estudios periódicos para ver el comportamiento a lo largo del tiempo.
- Realizar una nueva encuesta ampliando el ámbito de aplicación y así disponer de resultados más simétricos y que cubran a la mayoría de alumnos del SUPE.

En resumen, ampliando las variables observables con información de los resultados académicos, profundizando a nivel de titulación y ampliando el ámbito de aplicación consiguiendo más resultados, consideramos que el análisis realizado puede ayudar en la toma de decisiones tanto a los alumnos como a los responsables universitarios.

# 6. Bibliografía


[1] Soriano Jiménez, Pedro Pablo. *"Un modelo estructural para el análisis de los factores asociados a la elección de estudios universitarios"* 2016.

[2] Mosteiro García, Mª Josefa, Porto Castro, Ana Mª. "*Motivos de elección de estudios en alumnos y alumnas de universidad.*" Innovación Educativa, 2000, nº 10.

[3] Sánchez García, Mª Fe. "*La orientación universitaria y las circunstancias de elección de los estudios.*" Revista de Investigación Educativa. Volumen 19, número 1. 2001.

[4] Gonzalez López, Ignacio. "*Motivación y actitudes del alumnado universitario al inicio de la carrera. ¿Varían al egresar?"* Electronic Journal of Research in Educational Psychology, *2005.*

[5] Navarro Guzmán, C., Casero Martínez, A. "Análisis de las diferencias de género en la elección de estudios universitarios." Revista Semestral del Departamento de Educación. Facultad de Filosofía y Letras. 2012.

[6] Gámez, Elena, Marrero, Hipólito. "*Metas y motivos en la elección de la carrera universitaria: un estudio comparativo entre psicología, derecho y biología.*" Anales de Psicología 2003, vol 19. Nº 1 (Junio) 121-131.

[7] Solsona Solé, Joan, Gómez Adillón, Mª Jesús, Saladrigues Solé, Ramón. "*Análisis de los factores de elección de la universidad de Lleida. La demanda universitaria en la provincia de Lleida.*" Investigaciones de Economía de la educación. Número 11. 2016.

[8] Valle, A., González, R., Cuevas, L.M. "*Patrones motivacionales en estudiantes universitarios: características diferenciales.*" Revista de Investigación Educativa. Volumen 15, número 1. 1997.

[9] Ollé, Jordi. (12/06/2017). Curso de Estadística. Lugar de publicación: Conceptos Claros. https://conceptosclaros.com

[10] Martínez, María. (25/08/2017). Cómo utilizar la escala Likert en el análisis estadístico. Lugar de publicación: Educacion.uncomo. https://educacion.uncomo.com/articulo/como-utilizar-la-escala-de-likert-en-el-analisis-estadistico-2354.html

[11] Boone, Harry N. Jr., Boone, Deborah A. "*Analyzing Likert Data*" Jounal of Extension. Volume 50 Number 2. Abril 2012. Disponible en: http://wiki.biologyscholars.org/@api/deki/files/2002/=Likert_Scale_Analysis.pdf

[12] Alonso Gimeno, Mª Jesús, Martín, Mercè. (17/02/2017). Introducción al Machine Learning. Lugar de publicación: Miriada X. https://miriadax.net/web/introduccion-al-machine-learning

[13] Parra, Francisco. (25/07/2017). Estadística y Machine Learning con R. Agrupación de la información. Lugar de publicación:Rpubs . https://rpubs.com/PacoParra/293407

[14] Wikipedia (08/07/2017). R (lenguaje de programación). Lugar de publicación: Wikipedia. https://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n)

[15] Proyecto R (25/07/2017). Paquete Stats. Lugar de documentación: Rdocumentation.org. https://www.rdocumentation.org/packages/stats/versions/3.4.1/topics/stats-package

[16] Proyecto R (25/07/2017). Paquete Cluster. Lugar de documentación: Cran.r-project.org. https://cran.r-project.org/web/packages/cluster/cluster.pdf

[17] Llauradó, Oriol. (25/08/2017). La escala de Likert: qué es y cómo utilizarla. Lugar de publicación: Netquest. https://www.netquest.com/blog/es/la-escala-de-likert-que-es-y-como-utilizarla

[18] Santana, Emmanuel (25/07/2017). Data Mining con R: Optimizar K-Means. Lugar de publicación: Apuntes R BlogSpot. http://apuntes-r.blogspot.com.es/2014/10/optimizar-k-means.html