UNIVERSIDAD DE GRANADA

Facultad de Ciencias

GRADO EN ESTADÍSTICA

TRABAJO DE FIN DE GRADO

Análisis Multivariante aplicado a datos sobre proyectos STEAM en educación secundaria

Presentado por:

Inmaculada Agredano Espinal

Tutorizado por:

Nuria Rico Castro

Departamento de Estadística e Investigación Operativa

Curso académico 2023-2024

Análisis Multivariante aplicado a datos sobre proyectos STEAM en educación secundaria

Inmaculada Agredano Espinal

Inmaculada Agredano Espinal *Análisis Multivariante aplicado a datos sobre proyectos STEAM en educación secundaria*.

Trabajo de fin de Grado. Curso académico 2023-2024.

Responsable de tutorización	Nuria Rico Castro *Departamento de Estadística e Investigación Operativa*	Grado en Estadística Facultad de Ciencias Universidad de Granada

Declaración de originalidad

Dña. Inmaculada Agredano Espinal

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2023-2024, es original, entendido esto en el sentido de que no he utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 4 de julio de 2024

Fdo: Inmaculada Agredano Espinal

Índice general

Agradecimientos

En primer lugar, me gustaría agradecer a Nuria, mi tutora,por su compromiso, su dedicación, las pautas que me ha ido indicando y su apoyo a lo largo de todo este curso.

Agradecer, de igual forma, al equipo docente del Departamento de Estadística e Investigación Operativa por su enseñanza durante estos años.

Y, sin duda, a mis padres, a mi hermana y a toda mi familia, por su apoyo diario y la confianza que han puesto en mí en todo momento.

Autoevaluación

El objetivo de este Trabajo de Fin de Grado ha sido investigar sobre el perfil de los alumnos que participan en los proyectos STEM, observando si existen grupos con respuestas homogéneas o si tienen características similares que permitan describir estos grupos. Una vez concluido este estudio, se considera que el objetivo ha sido cumplido en su totalidad.

Este problema se ha abordado tanto de manera teórica como práctica. Para ello, en la primera parte del trabajo se establecen las bases teóricas que sustentan el análisis cluster que después se aplicará al conjunto de datos. Se observan los principios fundamentales de esta técnica; la selección de variables y su tratamiento, las medidas que pueden utilizarse para observar la distancia o similaridad entre individuos y los procedimientos que se siguen habitualmente en los métodos jerárquicos y no jerárquicos, así como diferentes mecanismos para la selección del número de clusters y la comparación entre diferentes soluciones.

Tras esta primera parte teórica, se aborda en la segunda parte la aplicación a datos reales. Los datos son el resultado de una encuesta a estudiantes de educación secundaria de la provincia de Granada y de Melilla. Algunos de ellos han tenido la oportunidad de participar en proyectos STEAM financiados por la Junta de Andalucía. Se pretende obtener una clasificación de los estudiantes que ponga de manifiesto qué tipos de respuesta pueden esperarse en este cuestionario.

Para realizar el análisis se utiliza el software estadístico R, introduciendo previamente las funciones principales que se utilizan. Este software ofrece la ventaja de ser libre y multiplataforma, con gran cantidad de librerías a disposición del usuario para realizar multitud de análisis.

Tras realizar el análisis, se establece un último apartado de conclusiones que pueden derivarse del mismo. Así, se concluye que los estudiantes pueden dividirse en 5 grupos según las respuestas que consignan en la encuesta. A grandes rasgos, el primer grupo estaría conformado por estudiantes con bajo interés en ciencia y tecnología, el segundo por estudiantes con alto interés en ciencias pero bajo en ingeniería, el tercero con bajo en ciencias pero alto en ingeniería, el cuarto con alto en ambos campos y el quinto con interés medio pero baja confianza en sus habilidades.

Personalmente, he disfrutado mucho haciendo este análisis y considero que merezco la máxima calificación, acorde con la autonomía y capacidad de trabajo que he demostrado.

Resumen

En este Trabajo Fin de Grado se hace un estudio teórico acerca del Análisis Cluster y se realiza una aplicación del mismo a un conjunto de datos reales. Así, el trabajo consta de dos partes diferenciadas e interconectadas; en la primera de ellas se aborda la técnica de Análisis Multivariante de una forma teórica, exponiendo las bases que lo sustentan y las principales metodologías que se siguen en la práctica. En la segunda parte se aplican diferentes funciones del paquete estadístico R [28], a un conjunto de datos reales para obtener agrupaciones de individuos que faciliten un mejor conocimiento de los mismos.

La primera parte de este trabajo consiste en la exposición de las bases teóricas de la técnica. En esta parte del trabajo se realiza, en primer lugar, un resumen de diferentes técnicas multivariantes, destacando las características principales de cada una de ellas y destacando las del análisis cluster, así como los objetivos que persigue y cómo se sitúa dentro del conjunto de técnicas de Análisis Multivariante existentes. Esta contextualización permite situar esta técnica y observar claramente cómo y cuándo debe utilizarse, qué datos son necesarios y qué hipótesis se presuponen.

Una vez establecidos los principios de la técnica, se abordan diferentes aspectos relevantes, como la selección de variables y su tratamiento, aspectos fundamentales para asegurar que los resultados son interpretables y reflejan la estructura subyacente de los datos observados.

Otro aspecto relevante que se expone en este trabajo es la medición a partir de las observaciones de la proximidad entre individuos o variables, recorriendo la forma de medición mediante distancias y similaridades entre vectores.

Esta medición da pie a abordar un aspecto fundamental, que es conocer cómo funcionan diferentes métodos de clustering existentes. Se realiza el estudio observando con atención los métodos jerárquicos aglomerativo del linkage simple y el linkage completo, así como el método no jerárquico llamado de las k-medias. Para una mejor comprensión, se ejemplifican estas técnicas con matrices de datos sencillas.

Cerrando esta primera parte, se revisan las funciones de R necesarias para la implementación del análisis y cómo se pueden validar los resultados. Se

elige el software R por su gran versatilidad, la gran cantidad de funciones implementadas y su carácter de software libre, que lo sitúan como el software estadístico más relevante para el análisis de datos hoy en día.

En la segunda parte del trabajo se trabajará con los datos de una encuesta dirigida a estudiantes de Enseñanza Secundaria Obligatoria de diferentes centros andaluces. Esta encuesta trata de averiguar si la participación en programas de proyectos STEM subvencionados por la Junta de Andalucía ha afectado en la percepción de las disciplinas implicadas en los estudiantes. Nuestro objetivo será averiguar, a partir de las respuestas registradas, cuál es el perfil de los estudiantes que participan, si existen grupos con respuestas homogéneas y si comparten características que nos permitan describirlos fácilmente.

Se aplica, pues, el Análisis Cluster a los datos para observar grupos homogéneos de estudiantes, esto es, cómo se pueden clasificar los estudiantes en diferentes grupos y qué características son las que describen dichos grupos.

Como resultado se puede destacar que el análisis muestra la existencia de cinco grupos de escolares con respuestas similares. Tanto un análisis jerárquico como un no jerárquico nos dan resultados coincidentes. Se aprecian en ambos casos: un grupo que claramente manifiesta un interés muy bajo y puntúa con valores en general por debajo de la media; otro grupo cuya característica más relevante es que muestra puntuaciones muy altas en un grupo de variables que conforman el bloque segundo del cuestionario y puntuaciones menores en las preguntas del bloque 3. Esto indicaría un elevado interés en las ciencias pero no tanto en las ingenierías. Se puede apreciar otro grupo de escolares que presenta un patrón simétrico al anterior, es decir, registra valores altos en las preguntas del bloque 3 y menos altas en las preguntas del bloque 2, manifestando así un mayor interés por la ingeniería que por las ciencias. Un cuarto grupo muestra un elevado interés en todos los ámbitos por los que se les pregunta y un quinto grupo que contesta mostrando un interés tibio en las áreas de ciencias y tecnología y que tienen menor confianza en sus aptitudes, ya que presentan valores menores en las preguntas del bloque cuarto.

Este trabajo se ha desarrollado dentro del marco del proyecto de generación del conocimiento "Proyectos de Educación STEAM y aprendizaje escolar", PID2021-128261NB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE.

Summary

In this Final Degree Project, a theorical study about Cluster Analysis is done and it is applied to a real data group. Thus, the project consists of two different and interconnected parts: in the first one the Multivariant Analysis' technique in a theoretical way is tacked, exposing the basis that supports it and the main methodologies that they follow in the practice; while in the second part, different functions of the R statistical package are applied to a group of real data in order to obtain group of individuals which facilitates a better knowledge of themselves.

The first part of this project consists of the exposition of the technique's theoretical bases. In this part of the project, firstly a summary of the different multivariant techniques is written, highlighting the main characteristics of each of them and the ones belonging to the cluster analysis, as well as the objectives it pursues and how they are placed inside the existent Multivariant Analysis' group of techniques. This contextualization allows to place this technique and to observe in a clear way how and when it should be used, as well as which data is necessary and which hypothesis they presuppose.

Once the technique's principles are stablished, different relevant aspects like variable selection and its treatment are tackled which are fundamental aspects to guarantee that results are interpretable and that they reflect the observed data's underlying structure.

Another relevant aspect that is exposed in this project is the measuring from the proximity observations between individuals and variables, going through the measuring way by distances and similarities between vectors.

This measuring leads to a fundamental aspect, that is to know how different clustering methods that exist work. The study is carried out by observing carefully the hierarchical agglomerative simple linkage and complete linkage methods, as well as the non-hierarchical method named/known as k-medias. To obtain better comprehension, these techniques are exemplifying with simple data matrixes.

Closing this first part, the necessary R functions for the implementation of the analysis and how result can be validated are revised. Software R is chosen because of its versatility, the great amount of implemented functions and its free software condition, which positions it as the most relevant statistics

software for analysis data nowadays.

In the second section of this project, data from a survey aimed at Compulsory Secondary Education students from different Andalusian centres will be used. This survey tries to find out if participation in STEM projects' programmes financed by the Junta de Andalucía (Autonomous Government of Andalusia) has affected in the perception of the implied on student's disciplines. Our objective will be to figure out, based on the registered answers, which is the profile of the students who participate, if there exist homogenous answers groups and if they share features that allows us describe them easily.

Therefore, cluster analysis is applied to data in order to observe homogenous groups of students; this is, how can be students classified in different groups and which characteristics describe the groups themselves.

As a result it can be highlighted that the analysis shows the existence of five group of students with similar answers. Both a hierarchical analysis and a non-hierarchical one show concordant results. We can appreciate in both cases: a group that clearly reveals not many interest and punctuates with below-average values in general; another group of which most relevant feature is that it shows very high punctuations in a group of variables that shape the second section of the questionnaire and the lower punctuations of the 3rd section. This would indicate an elevated interest in science but not so much in engineering. Another group of students that present a symmetric pattern respecting the previous one can be appreciated, that is, it makes a record of high values in section 3's questions and less elevated in the ones belonging to section 2, stating in this way a higher interest in engineering respecting to science. A fourth group shows a high interest in all the fields they are asked for, as well as a fifth one answers showing a lukewarm interest in the scientific and technological areas and that have less confidence in their abilities, since they present lower values in the fourth section's questions.

This work has been carried out within the framework of the research project PID2021-128261NB-I00 (PROESTEAM), financed by MICIN/AEI/10.13039/501100011033 and by FEDER, EU.

Introducción

En el año 2021, se pone en marcha el Proyecto de Investigación «Proyectos de Educación STEAM y Aprendizaje Escolar (PROESTEAM)», llevado a cabo por un conjunto interdisciplinar de docentes de la Universidad de Granada. El objetivo principal de este proyecto es abordar la evaluación de la repercusión de las iniciativas formativas en el ámbito STEM tanto en los estudiantes como en los profesores, dedicando especial atención a las vocaciones científicas y tecnológicas despertadas en las niñas y jóvenes.

En estos proyectos, según la resolución de 22 de julio de 2021, de la Dirección General de Formación del Profesorado e Innovación Educativa, por la que se convocan a los centros docentes sostenidos con fondos públicos de Educación Infantil, Primaria y Secundaria para el desarrollo del «Proyecto STEAM: Robótica aplicada al aula» durante el curso escolar 2021-2022 [8], se investiga la importancia de implementar una serie de programas, por ejemplo, como en el curso 2021-2022 de Investigación Aeroespacial Aplicada al aula y Robótica Educativa, para la formación de alumnos de Educación Secundaria Obligatoria en la Provincia de Granada.

Con el propósito de conocer el impacto de las actividades STEM que financia la Junta de Andalucía en el estudiantado que participa en ellas, se elabora una encuesta que pretende observar diferentes aspectos en los estudiantes. La planificación, recogida de datos y análisis preliminar de los resultados de la encuesta forman parte de un Trabajo Fin de Grado [27] que tiene como meta proporcionar e implementar el diseño estadístico. En este trabajo se dejan líneas abiertas para el análisis de datos, una de las cuales es realizar un análisis multivariante para observar e identificar grupos de respuesta común o, equivalentemente, grupos de estudiantes con respuestas similares en la encuesta.

Este trabajo de igual manera que el de [27], tiene, pues, su marco en el proyecto de investigación mencionado y se basa en los datos recogidos en la encuesta llevada a cabo con el objetivo de arrojar luz en la evaluación de la repercusión y el impacto de diferentes iniciativas públicas en cuatro ámbitos: la formación del profesorado, el aprendizaje escolar y el despertar de vocaciones científicas en niñas y jóvenes, y el aprovechamiento de las inversiones públicas en programas de formación.

La realización de este Análisis Multivariante, como es el Análisis Cluster nos permitirá analizar si existe alguna forma de agrupar los estudiantes según hayan tenido un perfil de respuesta determinado, así como cuántos perfiles de respuesta es adecuado considerar. De esta forma, será más sencillo entender los resultados que se observen, estableciendo cuál es el perfil más habitual y cuál es más específico, e intentando interpretar las respuestas de cada grupo estudiando sus caracaterísticas fundamentales.

Este trabajo se estructura en dos partes. En la primera parte, como hemos dicho, explicaremos conceptualmente el Análisis Multivariante. Antes de adentrarnos en ello, hablaremos de cuáles fueron sus inicios y cómo se ha ido desarrollando, de igual manera, explicaremos la clasificación de las técnicas que se usan para este análisis. Nos centramos en una de ellas: análisis cluster. En la segunda parte, como hemos dicho, llevaremos a cabo la aplicación a los datos reales.

Parte I.

Análisis Cluster

1. Introducción al Análisis Multivariante de datos

Se conoce como Análisis Estadístico Multivariante a un conjunto de técnicas estadísticas que analizan múltiples variables observadas sobre el mismo conjunto de individuos. Según, Balzarini *et al.* en su artículo de 2015 [5], estos métodos *permiten explorar, describir e interpretar datos que provienen del registro de varias variables sobre un mismo caso objeto de estudio*.

En palabras de Kendall [15], el Análisis Multivariante (Multivariable o Multivariado) *es una rama de la estadística que se interesa en el estudio de la relación entres series de variables dependientes de los individuos que las sustentan*.

Según se desprende del artículo de Lozares y López-Roldán de 1991 [16], el Análisis Multivariante se caracteriza por partir de una información que viene dada en soporte matricial. Esto implica que la matriz de datos tiene un soporte algebraico que es susceptible de someterse a la lógica del lenguaje matemático y sus métricas, siempre bajo el objeto de procurar una simplificación consistente.

Los objetivos últimos de las diferentes técnicas que conforman el Análisis Multivariante son variadas y dependerán no solo de la intención de la investigación, sino de la propia naturaleza y definición de los datos, los supuesto que los rigen y el cumplimiento o no de diferentes hipótesis necesarias.

En esta sección se realiza una breve introducción histórica al Análisis Multivariante y se expone una clasificación de los diferentes tipos de técnicas dentro del Análisis Multivariante según el objetivo perseguido.

1.1. Evolución histórica del Análisis Multivariante

El Análisis Multivariante, según resume Peña [20], nace en el siglo XIX de la mano de Galton [11], quien fue la primera persona que desarrolló una técnica para evaluar una relación estadística, incluyendo conceptos básicos hoy en día como *recta de regresión* o *correlación entre variables*. Galton en sus investigaciones, observó la relación entre generaciones y concluyó que existía una herencia de rasgos genéticos cuantitativos multifactoriales, concretamente que el descendiente de los padres que se encuentran en las colas de la distribución tiende a estar más cerca del centro (la media) de la distribución. Al cuantificar esta tendencia inventó el análisis de regresión lineal, al que

llamó *regresión a la media* por haber comprobado que los datos extremos tendían, con el paso de generaciones, a «volver» a un lugar central, más cercano a la media. Estos descubrimientos surgieron en las investigaciones sobre la transmisión de rasgos hereditarios, como propuesta de su primo, Darwin, en 1859, quien estaba estudiando la teoría de la evolución de las especies.

Años más tarde, Edgewort, quien estudiaba la normal multivariante y la matriz de correlación, usó como aplicación para las ciencias sociales, este concepto de correlación.

Mientras, el estadístico británico Pearson, que desarrolló el contraste de χ^2, obtuvo el estimador del coeficiente de correlación. Con esto, se encargó de encarar el problema de encontrar si grupos de personas de las que tenía medidas físicas eran pertenecientes a la misma raza. Hemos de aclarar la cuestión racial, ya que, en 1950, la UNESCO, en una declaración [30], recomendó sustituir la noción de raza humana, considerada no científica y confusa, por la de etnia, basada más en las diferencias culturales.

Estos planteamientos llamaron la atención al joven Hotelling, quien viajaba a la estación de investigación agrícola en Reino Unido y se interesó por comparar tratamientos agrícolas en función de más de una variable. En 1933, cuando regresó a la Universidad de Columbia, uno de sus profesores le planteó el problema de encontrar factores capaces de explicar resultados obtenidos por un grupo de personas en un cuestionario de inteligencia, Hotelling ideó la metodología de los componentes principales que más tarde son los que darán lugar al análisis factorial.

Años más tarde, en 1940, después de estudiar la solución al problema de encontrar factores que representaran todos los datos, Fisher implantó nuevas ideas que lo llevaron al análisis discriminante. Estas nacieron de querer resolver el problema de de discriminación de cráneos en antropología. El problema era clasificar un cráneo encontrado en una excavación arqueológica como perteneciente a un homínido o no.Para ello, primero, inventa un método general basado en hacer un análisis de la varianza, luego, junto a Mahalanobis, busca relaciones entre la medida de Mahalanobis y sus resultados del análisis, y siguiendo los resultados de Hotelling sobre contraste de medias de poblaciones univariantes consiguen resolver el problema. Mahalanobis, pocos años después, extiende el análisis discriminante a más de dos poblaciones.

A partir de aquí, se desarrollaron con rapidez diferentes métodos de clasificación de observaciones. Con esto, llega el nacimiento del análisis cluster. En 1967, MacQueen en [18], introdujo el algoritmo de k-medias (aunque hay registros de otras investigaciones que proponen metodologías similares con anterioridad) el cual, hoy en día es uno de los algoritmos de aprendizaje no supervisado más simples que resuelve el conocido problema de agrupamien-

to.

El análisis cluster, que desarrollaremos en los siguientes capítulos, aborda el problema de encontrar una agrupación natural en los datos. Este tipo de análisis a veces se confunde con el análisis factorial, pero se diferencia porque lo que busca son grupos entre los que hay características similares. Por otro lado, también nos podría llevar a confusiones con el análisis discriminante, la diferencia es que el análisis cluster no conoce en un principio cuál es su estructura, sino, que es eso a lo que aspira.

El nacimiento del ordenador transforma de manera radical los métodos multivariantes. Estos surgen un cambio creciente desde los años 70 hasta la actualidad que siguen sufriendo constantes transformaciones. Estas transformaciones son debidas; por un lado, a la gran masa de datos que se deben de manejar y, por otro, a las hipótesis sobre las que se trabaja. Sin duda, se espera, debido a las crecientes posibilidades de cálculo proporcionadas por los ordenadores actuales, que el Análisis Multivariante siga creciendo y que se amplíe el campo de aplicación de estos métodos a problemas más complejos o globales.

1.2. Clasificación de técnicas del Análisis Multivariante

El Análisis Multivariante, según J [13], se basa en el uso de técnicas que estudian, analizar, representan e interpretan datos, al mismo tiempo, de múltiples variables. Las técnicas de análisis de datos multivariantes pueden clasificarse en términos generales en dos tipos principales, según Tukey [29]:

(i) exploratorias o descriptivas, lo que significa que el investigador no tiene modelos o hipótesis preespecificados pero quiere comprender las características generales o la estructura de los datos de alta calidad, y

(ii) confirmatorios o inferenciales, lo que significa que el investigador quiere confirmar la validez de una hipótesis/modelo o un conjunto de supuestos dados los datos disponibles.

Con este análisis se puede abordar por una parte el objetivo de reducir las dimensiones de los datos sin perder información, por otra parte permite la agrupación de los datos, y también, realiza predicciones de unas variables en función del conocimiento de otras.

Cuando realizamos un análisis de este tipo hemos de tener muy claro cuál es nuestro objetivo principal, ya que esto condiciona al método que vamos a aplicar. Los siguientes esquemas representan una clasificación de las técnicas de Análisis Multivariante, que es dada en el artículo de Lozares y López-Roldán [16], la podemos ver en la Figura 1.1, Figura 1.2 y Figura 1.3.

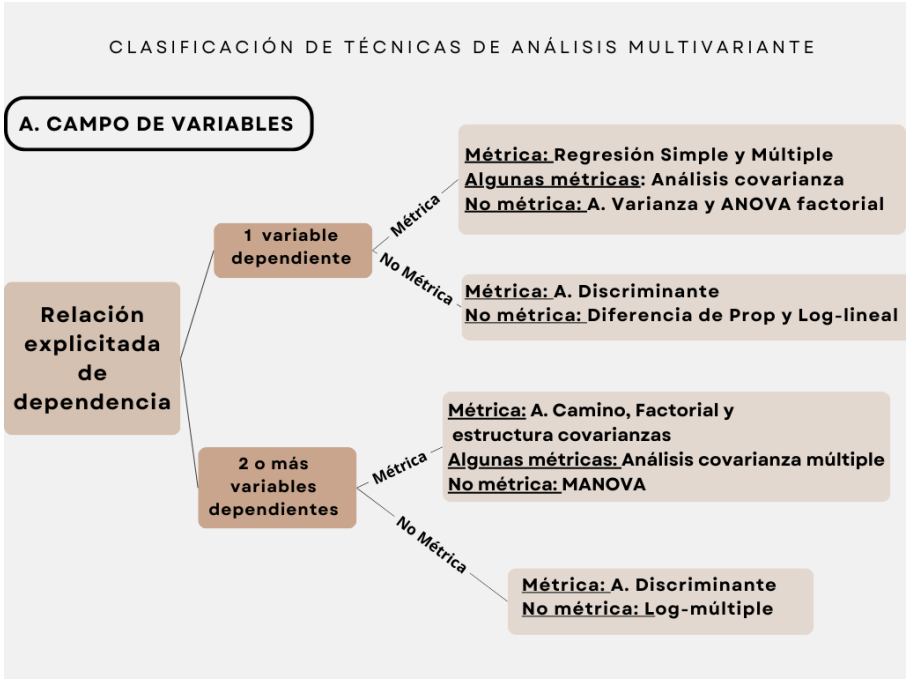


Figura 1.1.: Esquema de la clasificación (parte 1)

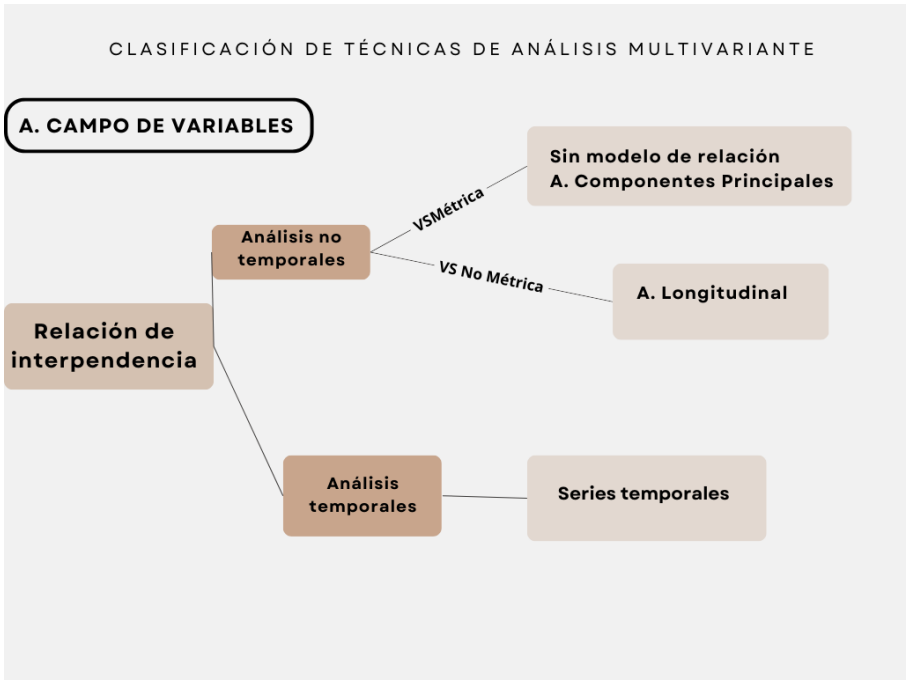


Figura 1.2.: Esquema de la clasificación (parte 2)

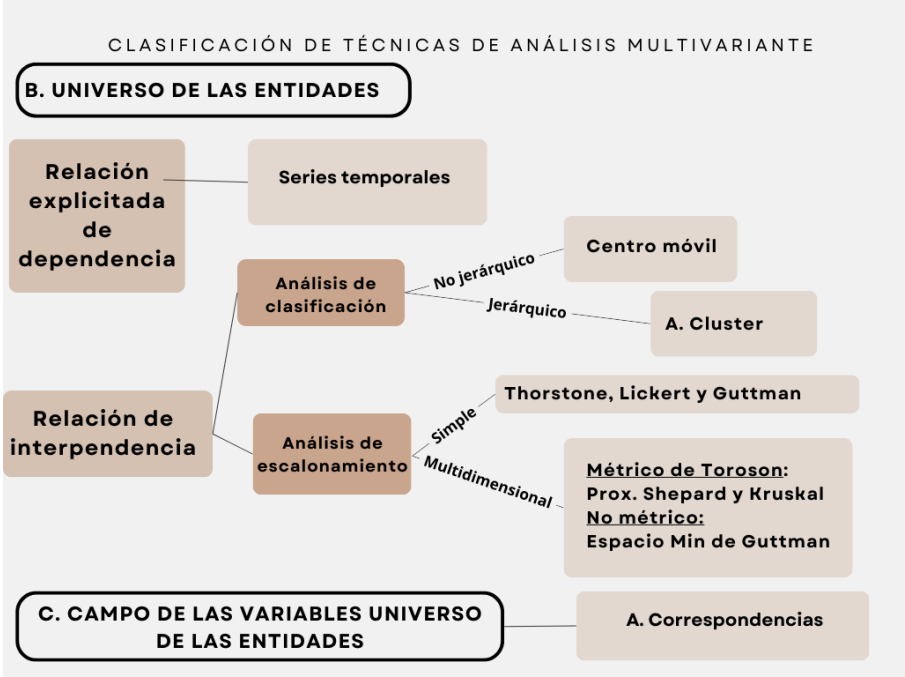


Figura 1.3.: Esquema de la clasificación (parte 3)

A continuación enunciaremos los métodos que se pueden usar en el Análisis Multivariante. Los podemos separar en dos grandes grupos, según López Pérez [17]:

1. Métodos de interdependencia. Tienen como principal objetivo describir o profundizar en el conocimiento de todas la variables en conjunto. Para ello, se intenta, por una parte, reducir la dimensión de la matriz de datos sin perder información y, por otra parte, explicar la estructura de los mismos. Dentro de este grupo de técnicas:
   - Análisis factorial. Tiene como objetivo, según Cuadras [7], expresas las variables como una combinación lineal de variables latentes, denominadas factores.
   - Análisis de componentes principales. Tiene como objetivo, según Peña [20], analizar si se puede representar la información con un número menor de variables que hayan sido construidas como combinaciones lineales de las originales.
   - Análisis de conglomerados o cluster. Tiene como objetivo, según Peña [20], agrupar los elementos en función de la similitud que hay entre ellos.

2. Métodos de dependencia. Son aquellos que buscan un modelo, incluyen una variable (o varias variables) que claramente depende de las demás y perseguimos un modelo. En este grupo, por otro lado, se encuentran:
   - Análisis discriminante. Tiene como objetivo, según Aldás y Uriel [1], explicar la pertenencia de distintos individuos a grupos alternativos a partir de valores de un conjunto de variables que describen a los individuos que va a clasificar.
   - Análisis de regresión logística. Tiene como objetivo, según López Pérez [17], analizar la relación entre una variable dependiente métrica y varias independientes métricas utilizando los valores conocidos de las independientes.
   - Análisis de correlación canónica. Tiene como objetivo, según López Pérez [17], utilizar las variables independientes, de las cuales conocemos sus valores, para predecir variables dependientes.

De entre todas estas técnicas, vamos a centrarnos en el análisis cluster, que es definido por una enciclopedia británica publicada por Merriam-Webster [9], como *una técnica de clasificación estadística para descubrir si los individuos de una población se dividen en diferentes grupos mediante comparaciones cuantitativas de múltiples características*. Así, el objetivo que podemos perseguir al usar este tipo de análisis multivariante es agrupar los elementos de una base de datos con características o con comportamientos similares en grupos a los que vamos a llamar clusters.

2. Planteamiento y selección de variables

Según describe Peña en [20], el análisis cluster tiene como objetivo agrupar elementos similares entre ellos. Aunque normalmente son los individuos los que se agrupan, también pueden agruparse variables. El análisis cluster estudia tres tipos de problemas:

Partición de los datos. Una forma de abordar un análisis de conglomerados es cuando se persigue agrupar los individuos en base a los datos que hemos recogido en un número de grupos, que como hemos dicho nombraremos clusters, de manera que cada individuo pertenezcan solo a un grupo, todos estén clasificados y el grupo sea de elementos homogéneos según las características que se hayan observado.

Jerarquías. Otro objetivo que puede abordarse mediante esta técnica multivariante es la construcción de jerarquías, es decir, agrupar ordenando según el grado de similitud. Para conseguir esto, debemos observar una metodología jerárquica de agrupación. Cuando se lleva a cabo un método jerárquico se representan en un gráfico llamado dendrograma, similar a un árbol, podemos ver en la Figura 2.1, cuyas ramas se unen (o separan si se trata de un método jerárquico disociativo) y que resume el procedimiento: pueden verse las uniones y la medida entre grupos que propicia cada una de las uniones de clusters.

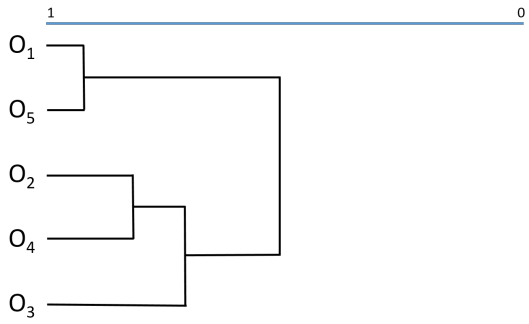


Figura 2.1.: Dendrograma de ejemplo

Clasificación de variables. Un tercer problema que podemos abordar utilizando esta técnica es la agrupación no de individuos sino de variables, de forma que tener variables similares pueden darnos pie a reducir la cantidad de estas que son necesarias para caracterizar a la población y como consecuencia se podría reducir la dimensión de los datos de los que se dispone.

Para realizar un análisis cluster, según el libro de Backhaus et al. [3], primero hemos de decidir a qué conjunto de variables vamos a utilizar para agrupar un conjunto de objetos. En el segundo paso, tenemos que decidir cómo se va a determinar la similaridad o la distancia entre los elementos. En el tercer paso elegimos un método de agrupación ya que pueden utilizarse diferentes métodos para combinar objetos similares en un cluster.

Discutiremos estas diferentes etapas necesarias que se usan.

2.1. Planteamiento

El punto de partida, como explican los autores Gutiérrez *et al.* [12], es una matriz X, esta será de dimensión $n \times p$ donde n es el número de elementos y p el número de variables observadas en ellos. La matriz de datos multivariantes es:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}$$

Según Cuadras [7], en el caso de trabajar sobre n elementos, en los cuales se han observado p variables, debemos definir que:

- Se obtiene el vector medias a partir de $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = \overline{x_j}$, tenemos $\overline{x} = (\overline{x_1}, \overline{x_2}, \dots, \overline{x_j}, \dots, \overline{x_p})'$ como el vector columna de las medias de las variables.
- Se obtiene la matriz simétrica pxp de covarianzas muestrales,

$$S = \begin{pmatrix} s_{11} & \dots & s_{1j} & \dots & s_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ s_{i1} & \dots & s_{ij} & \dots & s_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ s_{n1} & \dots & s_{nj} & \dots & s_{np} \end{pmatrix}$$

siendo la covarianza entre $j-$ésima y j'-ésima : $s_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ij'} - \overline{x_{j'}})$

- Se calcula la matriz simétrica pxp de correlaciones muestrales,

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

siendo $r_{jj'} = cor(x_j, x_{j'})$ el coeficiente de correlación muestral entre $x_j, x_{j'}$, que viene dado, siendo $s_j, s_{j'}$ las desviaciones típicas, por: $r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}}$

A partir de la información recogida en esta matriz, la idea de la técnica es establecer las diferencias entre los elementos para agrupar los elementos menos alejados entre sí, o bien observar una medida de similaridad para agrupar a los más similares dentro del mismo grupo.

2.2. Selección de variables

Dado que agruparemos los individuos que presenten características similares, y esto se conocerá a partir de calcular una distancia o una similaridad entre ellos, debemos tener en cuenta que la elección de las variables juega un papel crucial en la aplicación de esta técnica. Por una parte, debemos observar que la información recogida en las variables es la que va a dar lugar a la clasificación, por lo que la naturaleza de la similaridad entre individuos vendrá completamente determinada por las variables observadas. La clasificación que se obtenga estará íntimamente ligada a la información que se haya observado. Por otra parte, a la hora de establecer una similaridad entre individuos, debemos asegurarnos que las magnitudes de las variables son directamente comparables entre sí y, en caso de no serlo, tendremos que realizar alguna transformación para conseguir la comparabilidad.

A continuación, siguiendo los pasos de Milligan y Cooper [19], se resumen las diferentes formas en que se pueden transformar los valores observados para asegurar que las magnitudes son comparables, atendiendo a la naturaleza de las mismas.

- Variables dicotómicas. Estas variables tienen dos categorías a las que pueden pertenecer los datos, por ejemplo: *Sí/No, Hombre/Mujer*. Por ello,

podemos decir que son las que menos información nos proporcionan. Estas variables no se suelen ser transformadas.

- Variables categóricas nominales. Estas variables tienen más de dos categorías nominales, es decir, los datos no obedecen a un esquema de orden, por ejemplo, *Sí/No sé/No, Soltero/Casado/Separado/Viudo*. Estas variables pueden pasarse a dicotómicas según sus categorías, por ejemplo *Soltero/Casado/Separado/Viudo*, puede pasar a tener dos categorías siendo *Soltero = (Sí/No)*.

- Variables categóricas ordinales. Estas variables tienen más de dos categorías ordinales, es decir, los datos obedecen a un esquema de orden, por ejemplo, *Alto/Medio/Bajo, Primaria/Secundaria/Bachillerato/Grado*. Estas variables no se suelen transformar.

- Variables cuantitativas. Estas variables son aquellas que se expresan numéricamente, por ejemplo algunas variables cuantitativas son *el número de hijos, el peso, la talla, etc*.Estas se suelen transformar normalizándolas, de esta forma, se consigue tratar con la misma escala toda la matriz de datos que contiene diferentes unidades. Hay diferentes formas de normalizarlas:

  - Calcular valores tipificados. Para tipificar restamos la media y dividimos entre la raíz cuadrada de la varianza, es decir, entre la desviación típica los valores de la variables.

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x_j}}{\sigma_j}.$$

  - Homogeneizar a un intervalo entre 0 y 1. Por ello, restamos el valor mínimo a los valores de la variables, y posteriormente, se divide entre el rango de la variable:

$$x_{ij}^{*} = \frac{x_{ij} - min\{x_j\}}{max\{x_{ij}\} - min\{x_j\}}.$$

  - Igualar a un rango entre -1 y 1,

$$x_{ij}^{*} = \frac{2x_{ij} - (max\{x_{ij}\} + min\{x_{ij}\})}{max\{x_{ij}\} - min\{x_{ij}\}},$$

    i siendo el individuo y j la variable.

- Modificar los valores para conseguir que el máxima de todos sea 1.

$$x_{ij}^{*} = \frac{x_{ij}}{max\{x_{ij}\}}.$$

- Llevar a las variables a poseer una media unitaria. Por ello, los dividimos entre la media cada uno de ellos (teniendo en cuenta que no puede ser ninguna media 0).

$$x_{ij}^{*} = \frac{x_{ij}}{\bar{x_j}}.$$

- Si X tiene todas las variables en la misma escala, no es necesaria la estandarización.

De forma general, es recomendable hacer uso de la normalización cuando no tenemos las mismas unidades de medida en las variables.

Según el artículo de Vilà *et al.* [31], también es importante observar, al inicio de nuestro análisis, si encontramos en nuestros datos *valores perdidos y/o atípicos*, ya que estos nos pueden deforman las distancias y producir clusters unitarios. Una manera de evitar que los valores atípicos den confusión en nuestras agrupaciones es seleccionando un número de observaciones en cada cluster relevante. Así como un análisis previo de la multicolinealidad, para ver las correlaciones entre las variables ya que pueden perjudicar un análisis cluster.

Una cuestión que debemos tener en cuenta antes de empezar qué elementos vamos a agrupar, es decir, si son individuos o variables, ya que nos condiciona el modo de usar la matriz inicial. La diferencia, de manera técnica, entre agrupar individuos o agrupar variables se traduce en utilizar la matriz original o la matriz traspuesta. A la hora de agrupar objetos, según Backhaus *et al.* [3], las principales cuestiones que se plantean son cuántos clusters utilizar y si existe un *número óptimo* de clusters. La decisión sobre un determinado número de clusters suele ser también decidida de manera que nos facilite lo mayor posible el análisis (número reducido de clusters) y la homogeneidad (número elevado de clusters).

3. Medidas de proximidad

Una vez se han observado las variables y se ha normalizado, en caso de ser necesario, pasamos a analizar cómo será esa similitud y distancia entre las variables. Para ello, utilizaremos tanto medidas de proximidad (similaridades) como de alejamiento (distancias).

3.1. Definiciones básicas

Definiremos, de antemano, la distancia métrica y la similaridad.

Una distancia métrica, siguiendo las ideas de Peña [20], se define entre elementos de un conjunto, que nosotros llamaremos A, como una función $d : A \times A \to \mathbb{R}^+$ que verifica que para cualesquiera dos elementos x e y pertenecientes al conjunto A, se cumplen las siguientes propiedades:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \Leftrightarrow x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

Las dos primeras propiedades implican que d es una función definida positiva, la tercera es la propiedad de simetría y la cuarta es conocida como la desigualdad triangular.

Una similaridad, siguiendo a los autores Gutiérrez *et al.* [12], se define para los elementos de U, un conjunto finito o infinito de elementos, como $s : U_{\times} U \to \mathbb{R}$, cumpliendo que para cualesquiera dos elementos x e y pertenecientes a U, y s_0 un valor real arbitrario:

1. $s(x, y) \leq s_0$
2. $s(x, y) = s_0 \Leftrightarrow x = y$
3. $s(x, y) = s(y, x)$

Siendo el valor s_0 normalmente la unidad, que va a representar la similaridad máxima entre los objetos, la cual se alcanza cuando se mide la similaridad entre el objeto x y él mismo.

3.2. Medidas habituales de distancia entre individuos

Partiendo de la matriz X, donde cada fila es un vector que contiene las p observaciones de un individuo y cada columna es la observación de una variable en todos los individuos, podemos medir la distancia entre cada dos vectores fila x_r y x_s (correspondiente a dos individuos o a dos vectores traspuestos de dos variables), para ver la cercanía o lejanía entre ellos.

Expondremos, según Gutiérrez *et al.* [12], a continuación, las distancias más usuales.

Distancia de Minkowski. Se define la distancia entre los individuos r y s:

$$d_q(x_r, x_s) = \left(\sum_{h=1}^{p} |x_{rh} - x_{sh}|^q \right)^{1/q}$$

- Para $q = 2$ es la **distancia euclídea** . Es la más usada:

$$d_2(x_r, x_s) = \sqrt{\sum_{h=1}^{p} (x_{rh} - x_{sh})^2}. \tag{3.1}$$

- Para $q = 1$ es la **distancia Manhattan**, se define:

$$d_1(x_r, x_s) = \sum_{h=1}^{p} |x_{rh} - x_{sh}|.$$

- En el caso de la distancia de Manhattan cuando los valores son cercanos al 0, y se define la **distancia Camberra**:

$$d_C(x_r, x_s) = \sum_{h=1}^{p} \frac{|x_{rh} - x_{sh}|}{|x_{rh}| + |x_{sh}|}.$$

- Para $q = \infty$, es **distancia Chebychev o del máximo** se define:

$$d_\infty(x_r, x_s) = \max_{h=1,\dots,p} \{|x_{rh} - x_{sh}|\}.$$

Una generalización de la distancia euclídea es la **distancia de Pearson**:

$$d_p^2(x_r, x_s) = \sum_{h=1}^{p} \frac{(x_{ih} - x_{jh})^2}{\sigma_h^2}.$$

En la Figura 3.1 se representa gráficamente cuál sería la diferencia entre las distancias euclídea, de Manhattan y de Chebychev para dos puntos situados en el plano, lo que equivaldría a dos individuos para los cuales se tiene información de dos variables.

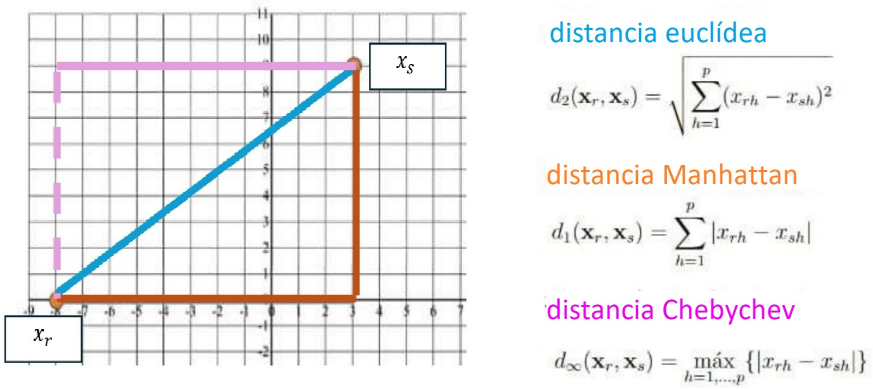


Figura 3.1.: Representación gráfica de las distancias

Distancia de Mahalanobis.

$$d_M(x_r, x_s) = \sqrt{(x_r - x_s)' S^{-1}(x_r - x_s)},$$

siendo S la matriz definida positiva de covarianzas muestrales antes definida.

Distancia de Gower. Se usa para establecer similaridad entre individuos y el conjunto de variables sea mixto (tenga varios tipos de variables diferentes).

$$d(x_r, x_s) = \frac{\sum_{h=1}^{p}(1 - \frac{|x_{rh} - x_{sh}|}{G_h} + a + \alpha)}{p_1 + (p_2 - d) + p_3},$$

siendo:

- p_1: número de variables métricas no binarias
- p_2: número de variables métricas binarias
- p_3: número de variables no métricas
- a: número de coincidencias (1,1,) y d número de coincidencias (0,0) en las variables binarias
- α: número de coincidencias en las variables no métricas

- G_h: rango entre el máximo y el mínimo de la h-ésima variable métrica

3.3. Medidas habituales de similaridad entre variables

En el caso de la asociación entre variables podemos definir diferentes medidas de similaridad, según Gutiérrez *et al.* [12].

3.3.1. Medidas de asociación para variables cuantitativas

Coseno del ángulo de los vectores. Es la mejor medidas para saber si dos vectores son paralelos, ya que lo son si este coseno toma valor 0.

$$cos(\widehat{x_r, x_s}) = \frac{\sum_{k=1}^{m}(x_{kr}x_{ks})}{\sum_{k=1}^{m}x_{kr}^2 \sum_{k=1}^{m}x_{ks}^2}.$$

Coeficiente de correlación lineal de Pearson. Siendo R la matriz simétrica pxp de correlaciones muestrales anteriormente definida, el coeficiente de correlación lineal de Pearson entre x_r y x_s, viene dado,

$$\rho(x_r, x_s) = \frac{Cov(x_r, x_s)}{\sqrt{Var(x_r)Var(X_s)}} = \frac{\sum_{k=1}^{m}(x_{kr} - \bar{x_r})(x_{ks} - \bar{x_s})}{\sqrt{\sum_{k=1}^{m}(x_{kr} - \bar{x_r})^2 \sum_{k=1}^{m}(x_{ks} - \bar{x_s})^2}}$$

3.3.2. Medidas de asociación para variables binarias

Para conocer la similaridad entre variables de tipo binario, que quedan registradas como 0 (ausencia) y 1 (presencia), presentaremos la tabla de doble entrada Tabla 3.1. Tendremos en cuenta la cantidad de individuos observados que presentan cada pareja de valores posibles en las variables r y s.

VARIABLES BINARIAS			
x_r / x_s	Presencia(1)	Ausencia(0)	TOTAL
Presencia (1)	a	b	$a+b$
Ausencia (0)	c	d	$c+d$
TOTAL	$a+c$	$b+d$	$a+b+c+d=n$

Tabla 3.1.: Tabla de doble entrada (datos binarios)

Medida de Ochiai. Esta medida es la precisión del coseno del ángulo formado por los vectores, para variables binarias:

$$Sim(x_r, x_s) = \frac{a}{\sqrt{(a+b)(a+c)}}.$$

Medida Φ Esta medida es el coeficiente de correlación lineal de Pearson para variables dicotómicas:

$$Sim(x_r, x_s) = \frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}.$$

Medida de Russel y Rao. Esta medida se basa en la probabilidad de coincidencia $(1,1)$ en dos variables:

$$Sim(x_r, x_s) = \frac{a}{m}.$$

Parejas simples. Esto hace referencia a la probabilidad de coincidencia de cualquier tipo de $(1,1)$ o $(0,0)$ en dos variables:

$$Sim(x_r, x_s) = \frac{a+d}{m}.$$

Medida de Jaccard. Esta medida es la probabilidad de coincidencia de tipo $(1,1)$ condicionada a que no hay coincidencia de $(0,0)$ en dos variables:

$$Sim(x_r, x_s) = \frac{a}{a+b+c}.$$

Medida de Dice. Esta medida es la variación de la medida Jaccard, donde la

coincidencia $(1,1)$ se pondera doble:

$$Sim(x_r, x_s) = \frac{2a+d}{2a+b+c}.$$

Medida de Rogers Tanimoto. Esta medida es la probabilidad de coincidencia entre las dos variables teniendo en cuenta la duplicación en la ponderación de las no coincidentes:

$$Sim(x_r, x_s) = \frac{a+d}{a+d+2(b+c)}.$$

Medida de Kulcynski. Esta medida es la proporción de coincidencias tipo $(1,1)$ entre las no coincidencias:

$$Sim(x_r, x_s) = \frac{a}{b+c}.$$

3.3.3. Medidas de asociación para variables ordinales

En el caso de tener variables de tipo ordinal, presentamos la tabla de doble entrada Tabla 3.2 con la que aclararemos conceptos básicos que usaremos en estas medidas. Además, explicaremos una serie de conceptos que se pueden dar en este tipo de asociación.

VARIABLES ORDINALES				
x_r / x_s	Modalidad 1	Modalidad 2	...	Modalidad q_r
Modalidad 1	A		...	
Modalidad 2	D		...	C
...	...	...	...	...
Modalidad q_s		B	...	

Tabla 3.2.: Tabla de doble entrada (ordinales)

- Dados dos individuos, decimos que son una **pareja concordante** si uno puntúa en ambas variables con valores menores al otro. Por ejemplo, si una variable tiene las categorías *poco*, *bastante* y *mucho* (variable ordinal) y otra variable tiene las categorías *suspenso*, *aprobado* y *sobresaliente*, una pareja de individuos donde uno presente la observación (*poco*, *suspenso*) y el otro la observación (*bastante*, *sobresaliente*) se llama pareja concordante.

- Dados dos individuos, decimos que son una **pareja discordante** si uno puntúa en un individuo valores menores en una variable mientras que el otro individuo presenta valores mayores en esa variable Para el mismo ejemplo mencionado, una pareja de individuos donde uno presente la observación (*poco*, *sobresaliente*) y el otro la observación (*bastante*, *suspenso*) se llama pareja discordante.
- Dados dos individuos, decimos que son **pares empatados** si al menos una de las variables no presenta valores mayores ni menores. Para el mismo ejemplo mencionado, una pareja de individuos donde uno presente la observación (*poco*, *aprobado*) y el otro la observación (*poco*, *suspenso*) se llama pares empatados.

Estas ideas de concordancia, discordancia y empates nos permite definir las siguientes medidas:

Coeficiente τ de Kendall. Este coeficiente se define como:

$$\tau(x_r, x_s) = \frac{n_c - n_d}{n_c + n_d},$$

siendo

- n_c :número de parejas concordantes
- n_d :número de parejas discordantes.

Coeficiente de concordancia $\tau - c$ de Kendall. Este coeficiente es la versión más usada para el cálculo de concordancia. Se define de la siguiente manera:

$$\tau - c(x_r, x_s) = \frac{2k(n_c - n_d)}{m^2(k-1)},$$

siendo

- n_c :número de parejas concordantes.
- n_d :número de parejas discordantes.
- k : es mínimo entre el número de modalidades de la variable r y las modalidades de las variables s, $k = \min\{(q_r, q_s)\}$.
- m: número total de parejas.

Coeficientes de correlación de Spearman. Este coeficiente representa el grado de asociación entre rangos a valores de variables analizadas. Para ello, se definen los valores ordinales $x_{(il)}$ que van a estar asociados a los

correspondientes x_{il}, y con ello, se define el coeficiente de correlación de Spearman:

$$\rho(x_r, x_s) = 1 - \frac{6\sum_{i=1}^{m}(x_{(il)} - x_{(ik)})^2}{m(m^2-1)}$$

Donde m es el número total de parejas.

4. Métodos de análisis cluster

Existen diferentes técnicas que podemos aplicar para hacer un análisis cluster. Estos métodos se pueden clasificar en dos grandes grupos: métodos jerárquicos y métodos no jerárquicos.

4.1. Métodos jerárquicos

Estos métodos toman la matriz de datos y a partir de ella obtienen una matriz de distancias o similaridades, que será cuadrada y simétrica, donde se establece la medida existente entre cada dos elementos. Con esto, según Gutiérrez *et al.* [12], su objetivo es formar o separar clusters, de manera que se haga mínima alguna de las funciones distancia o que se maximice alguna de las funciones de medida de similitud.

Los métodos jerárquicos se subdividen en aglomerativos y disociativos.

- Aglomerativos. En primer lugar, partiremos de tantos clusters como elementos se tenga que clasificar. Después, observaremos en la matriz de distancias, dónde se localiza la menor de estas (si fuera una matriz de similaridades, tendríamos que tomar la similaridad máxima), es decir, cuáles son los clusters más similares y los juntaremos para formar uno solo. De esta forma, vamos haciendo grupos, una vez formado el cluster en cada paso, la menor de las distancias hasta que todos los elementos estén englobados en un mismo conglomerado. Un ejemplo gráfico de cómo se lleva a cabo este método aglomerativo lo podemos ver en la Figura 4.1.

- Disociativos. Este caso es el inverso de caso de aglomerativos. Se inicia con un conglomerado que engloba a todos los elementos. Y desde este grupo inicial se van formando, a través de divisiones, clusters cada vez más pequeños. Hasta que se tienen tantos clusters como elementos iniciales. Un ejemplo gráfico de cómo se lleva a cabo, de manera análoga al método aglomerativo, el método disociativo lo podemos ver en la Figura 4.2.

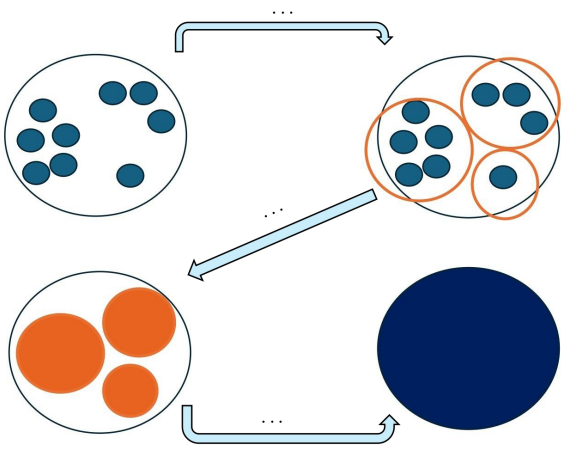


Figura 4.1.: Ejemplo de caso aglomerativo

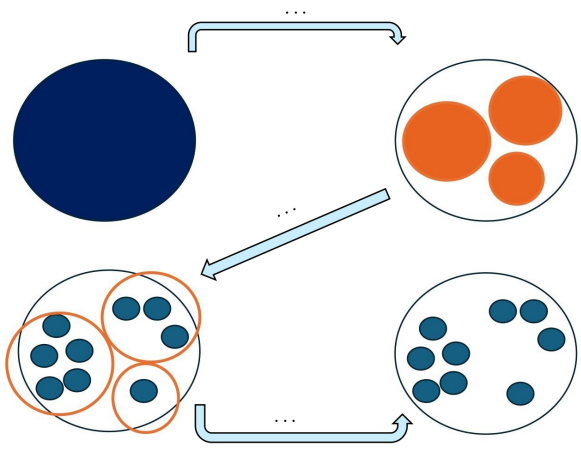


Figura 4.2.: Ejemplo de caso disociativo

Independientemente de la manera de agrupar que usemos, hemos de estudiar cómo se pueden recalcular las matrices de medidas después de cada paso en el que se ha formado o deshecho un cluster.

Veamos a continuación algunos de los métodos más habituales.

4.1.1. Método del vecino más próximo (Linkage simple)

En este método, como explican Aldás y Uriel [1], se considera que la distancia entre dos grupos será la mínima que exista entre elementos de ambos grupos, es decir, se toma como distancia de todo el grupo la de aquellos objetos de los grupos que estén más cerca, en el caso de estar considerando similaridades, se toma como similaridad entre grupos la mayor que haya entre dos elementos de grupos diferentes. Podemos verlo gráficamente en la Figura 4.3, lo que hacemos es medir entre grupos de la manera más cercana.

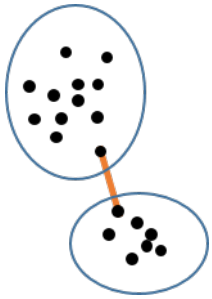


Figura 4.3.: Esquema de método de selección de distancias entre grupos linkage simple o vecino más cercano

Ejemplo. Como ejemplo, suponemos que se dispone de 5 objetos $(O_1, \ldots, O_5)$ para clasificar, y se obtiene una matriz de similaridades para ellos:

$$\begin{pmatrix} & O_1 & O_2 & O_3 & O_4 & O_5 \\ O_1 & 1 & 0,59 & 0,44 & 0,61 & 0,93 \\ O_2 & & 1 & 0,21 & 0,83 & 0,49 \\ O_3 & & & 1 & 0,71 & 0,57 \\ 0_4 & & & & 1 & 0,02 \\ 0_5 & & & & & 1 \end{pmatrix}$$

Por tanto, en primer lugar consideramos que cada objeto es en sí mismo un cluster, y vemos cuáles son los cluster más cercanos, que en este caso son los que tengan una similaridad mayor: O_1 y O_5. Entonces, esos dos clusters se unen formado solo uno y volvemos a calcular la matriz teniendo en cuenta

que la similaridad entre un cluster y otro será la máxima entre elementos de los clusters.

Así, la similaridad entre el nuevo cluster $O_1 - O_5$ y el cluster O_2 será la máxima entre la similaridad de la pareja $O_1 - O_2$ y la similaridad entre $O_5 - O_2$, en este caso la máxima es la similaridad entre O_1 y O_2. (En caso de que la matriz contuviera distancias, elegiríamos la distancia menor entre las dos posibles.) Llegamos entonces a una nueva matriz de similaridades:

$$\begin{pmatrix} & \{O_1, O_5\}, & O_2 & O_3 & O_4 \\ \{O_1, O_5\} & 1 & 0,59 & 0,57 & 0,61 \\ O_2 & & 1 & 0,21 & 0,83 \\ O_3 & & & 1 & 0,71 \\ 0_4 & & & & 1 \end{pmatrix}$$

En el siguiente paso, los cluster que se unen, por ser los más similares según la medida elegida, son O_2 y O_4,teniendo ahora que volver a calcular la matriz de similaridades, que quedará como:

$$\begin{pmatrix} & \{O_1, O_5\}, & \{O_2, O_4\} & O_3 \\ \{O_1, O_5\} & 1 & 0,61 & 0,57 \\ \{O_2, O_4\} & & 1 & 0,71 \\ 0_3 & & & 1 \end{pmatrix}$$

Siguiendo con este procedimiento, los clusters más similares son el $O_2 - O_4$ y O_3 , por lo que forman un cluster nuevo, y se recalcula la matriz de similaridades eligiendo la máxima entre los elementos de los clusters que se acaban de unir:

$$\begin{pmatrix} & \{O_1, O_5\}, & \{O_2, O_4, O_3\} \\ \{O_1, O_5\} & 1 & 0,61 \\ \{O_2, O_4, O_3\} & & 1 \end{pmatrix}$$

En el último paso, los dos clusters se unen formando uno solo con todos los objetos iniciales. De esta forma hemos realizado 4 uniones: la primera de los objetos 1 y 5, la segunda de los objetos 2 y 4, la tercera los objetos 2 y 4 se unen con el 3, y por último se unen todos los objetos en un solo cluster. El método del vecino más próximo indica que en cada paso la similaridad entre grupos sea la máxima entre los individuos, y en caso de tener distancias, elegiremos la mínima (la más favorable).

Como hemos dicho anteriormente, el método jerárquico se representan en un dendrograma; podemos ver la representación de los pasos dados en este

ejemplo en la Figura 4.4.

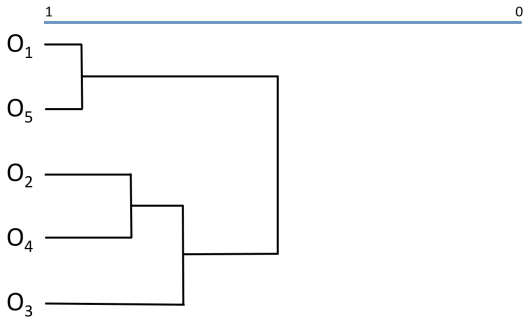


Figura 4.4.: Dendrograma resultante del ejemplo

4.1.2. Método del vecino más lejano (Linkage completo)

En este otro método, como también vemos en el libro de Aldás y Uriel [1], establecemos como distancia entre dos grupos la distancia que máxima que exista entre dos elementos uno de cada grupo. Podemos verlo gráficamente.

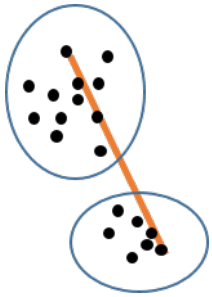


Figura 4.5.: Esquema de método de selección de distancias entre grupos linkage completo o vecino más lejano

Así, si resolvemos el ejemplo anterior utilizando este método tendremos que, partiendo de la misma matriz de similaridades, el primer paso coincidirá, puesto que siempre se unen los objetos más similares, y por tanto se unen los objetos 1 y 5, pero al calcular de nuevo la matriz de similaridades se tiene que establecer la similaridad entre este nuevo cluster y los demás, tomando el

caso más desfavorable (la menor similaridad) y por tanto quedará la matriz

$$\begin{pmatrix} & \{O_1, O_5\}, & O_2 & O_3 & O_4 \\ \{O_1, O_5\} & 1 & 0,49 & 0,44 & 0,02 \\ O_2 & & 1 & 0,21 & 0,83 \\ O_3 & & & 1 & 0,71 \\ 0_4 & & & & 1 \end{pmatrix}$$

En el siguiente paso se unen los clusters O_2 y O_4, teniendo ahora que volver a calcular la matriz de similaridades, que quedará

$$\begin{pmatrix} & \{O_1, O_5\}, & \{O_2, O_4\} & O_3 \\ \{O_1, O_5\} & 1 & 0,02 & 0,44 \\ \{O_2, O_4\} & & 1 & 0,21 \\ 0_3 & & & 1 \end{pmatrix}$$

En el siguiente paso uniremos los cluster $\{O_1, O_5\}$ y O_3, por lo que forman un cluster nuevo, y con él se vuelve a calcular la matriz de similaridades:

$$\begin{pmatrix} & \{O_1, O_5, O_3\}, & \{O_2, O_4\} \\ \{O_1, O_5, O_3\} & 1 & 0,02 \\ \{O_2, O_4\} & & 1 \end{pmatrix}$$

4.1.3. Otros métodos jerárquicos

Existen otros métodos que no consideran la distancia entre clusters como la mínima o la máxima entre los objetos, sino que promedian las medidas o establecen una medida entre los centroides de cada cluster, la cual es la más realista.

Hemos de mencionar el **método de Ward**, que según López Pérez [17],se basa en buscar la mínima varianza dentro de cada grupo a la hora de unir clusters, por tanto, en cada paso se unirán los clusters que incrementen menos el valor total de la suma de los cuadrados de las diferencias entre cada individuo del cluster al centroide de los clusters unidos. Este método es uno de los más usados ya que da lugar a los clusters pequeños y de tamaños similares y es capaz de acercarse más que otros métodos a la clasificación óptima. Una vez comprendido el mecanismo de las diferencias metodológicas, se puede observar que los mecanismo definidos se pueden generalizar según la siguiente fórmula de recurrencia:

Fórmula de recurrencia de Lance y Williams De forma matemática, Lance y Williams, según López Pérez [17], desarrollaron una fórmula general que

se puede usar en los distintos tipos de enlaces de los métodos jerárquicos algomerativos. Se define, según Gutiérrez *et al.* [12], ,siendo A, B y C clusters con n_A, n_B y n_C elementos, se supone la unión en una etapa o en la etapa anterior de los clusters B y C. Entonces, la fórmula para calcular la distancia entre A y el nuevo cluster formado por B y C es:

$$d(A,\{B,C\}) = \alpha_B d(A,B) + \alpha_C d(A,C) + \beta d(B,C) + \gamma |d(A,B) - d(A,C)|$$

De esta forma, tomando $\alpha_B = \alpha_C = \frac{1}{2}$; $\beta = 0$ y $\gamma = \frac{-1}{2}$, se obtiene una fórmula que nos lleva al método del vecino más cercano. Análogamente, tomando los valores $\alpha_B = \alpha_C = \frac{1}{2}$; $\beta = 0$ y $\gamma = \frac{1}{2}$, nos lleva a la fórmula de vecino más lejano. Podemos ver como sustituyendo otros valores en la fórmula de Lance y Williams, podemos llegar a otros métodos.

En la Tabla 4.1 mostramos un resumen de los métodos a los que podemos llegar y la forma de llegar a ellos.

Otros métodos conocidos

Método	α_B	α_C	β	γ
Vecino más próximo	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{-1}{2}$
Vecino más lejano	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Media no ponderada	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Media ponderada	$\frac{n_B}{n_B+n_C}$	$\frac{n_C}{n_B+n_C}$	0	0
Del centroide	$\frac{n_B}{n_B+n_C}$	$\frac{n_C}{n_B+n_C}$	$-\alpha_b\alpha_C$	0
De Ward	$\frac{n_B+n_A}{n_B+n_C+n_A}$	$\frac{n_C+n_A}{n_B+n_C+N_a}$	$\frac{-n_A}{n_B+n_C+N_a}$	0

Tabla 4.1.: Resumen de métodos jerárquicos siguiendo la fórmula de Lance y Williams

4.1.4. Valoración de la idoneidad

En esta sección vamos a describir algunas coeficientes, según Gutiérrez *et al.* [12]. Estos coeficientes nos van a ayudar a determinar la idoneidad de los resultados en el caso de los métodos jerárquicos.

Coeficiente de aglomeración. El coeficiente de aglomeración es una medida de ajuste de la aplicación de un método de clasificación jerárquico aglomerativo a los datos. Para cada objeto i se denota μ_i al cociente entre la distancia de i al primer grupo con el que se ha fusionado y la distancia de i al último grupo formado. El coeficiente de aglomeración se calcula realizando la media aritmética de todos los valores $1 - \mu_i$. Este coeficiente toma valores entre 0 y 1. Cuanto más cercano al 1, mejor será el ajuste del método escogido para nuestros datos.

Coeficiente de división o de disociación. Para métodos jerárquicos disociativos, se define d_i el cociente entre el diámetro del último grupo que se ha formado y el diámetro de todo el conjunto de datos. El coeficiente de disociación es la media aritmética de todos los valores $1 - d_i$. Este coeficiente toma valores entre 0 y 1. Cuanto más cercano sea este valor a la unidad, mejor será el ajuste del método escogido a nuestros datos.

Coeficiente de correlación cofenético. Este coeficiente es un coeficente de correlación lineal de Pearson entre dos matrices; la de medidas entre objetos inicial y la matriz coefenética. Esta última es la matriz que se va utilizando en las etapas del análisis sin reducir la dimensionalidad de la matriz original.

Cuando se tengan valores del coeficiente de correlación cofenético cercanos a 1 o -1, será indicativo de que la distribución en grupos de los datos es adecuada. Sin embargo, valores cercanos a 0 estarán indicando lo contrario, es decir, una estructura en grupos poco adecuada.

4.1.5. Elección del número de clusters

Cuando se lleva a cabo un método jerárquico, el número de grupos en que quedará la partición no tiene por qué estar establecido de antemano. Es habitual realizar esta selección una vez llevado a cabo alguno de los métodos jerárquicos, teniendo en cuenta algunas consideraciones según Kassambara [14].

Elección subjetiva. El procedimiento más básico consiste en cortar el dendrograma de forma subjetiva, en el lugar en que la experiencia en la investigación o el criterio experto considere más apropiado. Este criterio es poco satisfactorio, puesto que depende exclusivamente de la opinión del equipo experto. Algunas ideas intuitivas pueden apoyar la decisión, como por ejemplo, seleccionar en el dendrograma el paso en el que aumenta la distancia de fusión reflejada en el eje.

Método del codo. Este método consiste en minimizar la suma de las distancias al cuadrado de cada individuo con su centroide, lo que se conoce como WSS por ser las siglas de *within sum of squares*, es decir, la suma de las varianzas internas de los grupos. El algoritmo del codo pretende realizar una representación gráfica del WSS para diferentes valores del número de grupos k. La aparición de un codo en la gráfica indica una variación significativa en el WSS y señala el número idóneo de clusters.

Coeficiente de silueta. El método del coeficiente de silueta o sombra determina que el número óptimo de grupos es aquel valor que maximiza el coeficiente de sombra, que es una estimación de la distancia media entre grupos, representando cuán cerca están los individuos de un grupo de las observaciones de los grupos vecinos. Para cada objeto i, el coeficiente de silueta S_i se calcula como

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$$

donde a_i representa la distancia media entre el objeto i y los demás objetos del mismo cluster y b_i se calcula como $b_i = min\{d(x_i, x_s\}$, con x_s el resto de objetos del mismo cluster.

Cuando el valor del coeficiente de silueta se aproxima a la unidad, indica que el objeto i está bien clasificado. Valores cercanos a 0 indican que el objeto probablemente se encuentre entre dos grupos. Si el valor es negativo, es síntoma de que el objeto no ha quedado clasificado en el grupo que le corresponde. El algoritmo entonces consiste en calcular, para cada valor de número de grupos k, el valor medio de los coeficientes de sombra y buscar qué valor de k lo hace máximo.

Estadístico GAP. Este estadístico se utiliza para comparar un valor concreto de k grupos con el que se obtendría bajo la distribución de una hipótesis nula apropiada. El objetivo es estar lejos de la distribución uniforme, ya que esta proporciona la colección de puntos más alejados los unos de

los otros. El algoritmo concluye con la regla de seleccionar el valor de k para el cual el valor del estadístico sea mayor.

4.2. Métodos clásicos (no jerárquicos)

Estos métodos están diseñados para clasificar a individuos en k clusters, cantidad que debe ser conocida desde el principio ya que la metodología consiste en realizar una partición y a partir de ella hacer una resignación de individuos a los clusters para obtener una partición mejorada. Nos centramos en observar de distintas maneras de obtener la partición inicial y de reasignar los individuos a los clusters

4.2.1. Método de las k-medias

El método de las k-medias, según Gutiérrez *et al.* [12], consiste en partir de k semillas y asignar, según la proximidad de cada individuo al cluster con semilla más cercana. Tras cada asignación, la semilla se recalcula como el centroide del cluster formado. El cálculo de las semillas iniciales puede realizarse de varias formas:

1. Asignar los primeros k individuos como semillas.

2. Hacer un muestreo sistemático dentro de los individuos, seleccionando aquellos que estén en las posiciones $[im/k], i = 1, ..., k$ donde $[x]$ representa la parte entera de x.

3. Realizar un muestreo aleatorio simple de tamaño k de entre los m individuos a clasificar.

4. Tomar una partición pre-establecida en k grupos. Esto por ejemplo puede realizarse tras un análisis jerárquico que indique una partición posible para los individuos. Las k semillas son entonces los k centroides de los grupos iniciales establecidos.

5. Seguir el algoritmo propuesto por Astrahan [2], que consiste en estudiar la densidad de cada individuo, entendida como el número de casos que los rodean hasta una distancia d_1, para ordenar las densidades y seleccionar como primera semilla el núcleo que posea mayor densidad; se continúa escogiendo los núcleos según este criterio, con el añadido de que deben distar unos de otros una distancia d_2, hasta obtener las k semillas deseadas.

6. La opción propuesta por Ball *et al.* [4] es tomar como primer punto semilla el vector de medias de los datos y como subsiguientes semillas cualesquiera puntos que disten de los anteriores al menos una distancia d.

Una vez establecidas las semillas, se procede a la partición del grupo de individuos: se clasifican según su proximidad a los núcleos, clasificando dentro del mismo grupo aquellos con una menor distancia o con mayor proximidad al centroide del cluster.

Una vez se ha realizado la asignación, se repite el proceso de nuevo: los individuos pueden cambiar de grupo, dado que los centroides de los mismos han cambiado tras la incorporación de nuevos individuos al cluster. El procedimiento es iterativo, y se repite hasta que los centroides no cambien en una iteración, o bien cuando se alcance un número de iteraciones máximo pre-establecido.

Existen algunas variantes de este método, como la variante llamada k-medoides donde en lugar de utilizar los puntos medios (centroides) se utilizarán los medoides, que son los puntos pertenecientes al grupo de datos (individuos) que minimizan la distancia media entre los demás individuos, es decir, lo más centralizados, mientras que los centroides no tienen por qué pertenecer al conjunto. En el ámbito computacional se conoce como **PAM** (Partitioning Around Medoids).

Otra variante es el llamado método k-medians, que escoge las medianas como puntos semilla en lugar de los centroides. Una variante más es la conocida como k-means$+++$ el cual busca una optimización de las semillas iniciales, de entre todas las posibles combinaciones, para minimizar la varianza dentro de grupos en la elaboración de los clusters. Para ello se eligen los núcleos en base a una distribución de probabilidad ponderada que determinarán con probabilidad máxima los puntos que minimizan esa varianza.

4.2.2. Otros métodos no jerárquicos

Dentro de los llamados métodos de reasignación, como es el propio método de las k-medias, un individuo asignado en un paso determinado a un cluster puede ser asignado a otro grupo en un paso posterior si con ello se consigue una mejor clasificación. El proceso termina cuando no quedan individuos cuya reasignación permita optimizar el resultado. Dentro de estos métodos, además del ya expuesto, se encuentran otros, que hemos recopilado del libro de Gutiérrez *et al.* [12] como **Quick-Cluster analysis**, método de **Forgy** o método **de las nubes dinámicas**. De estos podemos destacar el método Forgy

[10], que consiste, según Gutiérrez [12],dado un conjunto de puntos de semilla asignamos cada caso a un cluster construido sobre el punto más próximo dejando los puntos de semilla estacionarios.

Existen otros métodos, además del mencionado de Astrahan [2], que son de búsqueda de la densidad, y tienen como objeto la formación de grupos buscando las zonas en las cuales se dé una mayor concentración de individuos, como el método de **Wishart**, **Taxmap** o **Fortin**. Podemos destacar el método de **Wishart** que consiste, según Gutiérrez *et al.*[12], en reducir el número de clusters final usando de base el método de las $k-$medias convergente.

Otro método de búsqueda de densidad, el método de **Wolf** aborda el problema considerando que las variables siguen una ley de probabilidad, según la cual los parámetros varían de un grupo a otro y trata de encontrar los individuos que pertenecen a la misma distribución.

Los métodos llamados directos permiten clasificar simultáneamente a los individuos y a las variables. El algoritmo más conocido dentro de este grupo es el **Block-Clustering**.

Por su parte, los conocidos como métodos de reducción de dimensiones consisten en la búsqueda de factores en el espacio de los individuos; cada factor corresponde a un grupo. Se les conoce como *análisis factorial tipo Q*.

5. Implementación en R

Para llevar a cabo un análisis cluster se pueden usar diferentes programas de ordenador, en nuestro caso, vamos a usar el lenguaje de programación de R.

R es un software, según la página web R-Pubs [25], de libre acceso y está disponible de manera gratuita. Según Crawley [6], R presenta un lenguaje con alto nivel y un campo amplio para el análisis de datos y/o gráficos.

Descargar e instalar R es sencillo. Basta con acceder a http://cran.r-project.org/ y hacer clic en "Download R"para el tipo de sistema que tenga nuestro ordenador. Además, en la página web de R-Project [21] , se ofrece información necesaria, podemos encontrar documentación de paquetes disponibles, ayuda sobre el uso de funciones, artículos o manuales. R-Studio, según Kassambra [14], es un entorno de desarrollo integrado para R que facilita el uso de R. Ya que, este incluye una consola, un editor de código y herramientas para trazar gráficos. La instalación de R-Studio, se lleva a cabo una vez se instala el software R en http://www.rstudio.com/products/RStudio/ Una vez, instalado, inicie R-Studio y da comienzo a utilizar R dentro de R-Studio.

A continuación se discutirá el proceso de preparación de los datos, las herramientas que podemos emplear para realizar el análisis y cómo podemos asegurar la validez de nuestros resultados.

5.1. Preparación de los datos

Antes de comenzar debemos asegurarnos que los datos que vamos a analizar están en óptimas condiciones.

Los datos los tendremos almacenados en un objeto que llamaremos `datos`. Este puede ser un objeto tipo `data.frame` o un archivo `excel` o una matriz de datos. Inicialmente, mostramos este objeto donde están almacenados nuestros datos. En caso de tener gran cantidad de datos y no siendo necesario mostrarlos todos, podemos hacer uso de la función `head(datos)`, que como podemos ver en una de las páginas web de ayuda de R [23],nos muestra los 6 primeros datos. Seguidamente, como se ve en la misma [23], debemos ver el tipo de variables que tenemos en nuestros datos y para ello usamos la función `str(datos)`.

Posteriormente, es importante comprobar si hay presencia de datos faltantes o valores perdidos en nuestro archivo. Y además, cuál es la proporción de datos que faltan, para ello, usamos la siguiente función:

```
is.na(datos)
proporcion<-sum(is.na(datos))/totaldedatos
```

Según Aldás y Uriel [1], si existen datos faltantes o hay valores que se han perdido, se dispone de dos grandes procedimientos que se suelen llevar a cabo para el tratamiento de los datos: la eliminación de los casos que contienen valores perdidos o la imputación de una estimación para ese valor. El procedimiento más usado es el de la eliminación, pero esto puede provocar estar eliminando casos que resultan muy significativos sin darnos cuenta. Esta es una de la principales limitaciones que se presentan en este procedimiento. Otra alternativa, es eliminar solo el valor que se ha perdido, aunque existe el problema de que va a variar el tamaño muestral entre variables. El caso de la imputación, también trae consigo desajustes que afectarán, sobre todo, a la varianza de la variable.

5.2. Funciones de R para el análisis cluster

En esta sección explicaremos, por un lado, funciones que podemos usar para realizar la matriz de similaridad o de distancias, y, por otro lado, funciones que usaremos para llevar a cabo las técnicas de análisis cluster dentro del paquete básico `stats`.

Presentaremos antes de comenzar, algunos paquetes los cuales, según, la página web R-bloggers [22], son útiles y básicos para el análisis cluster:

- El paquete `stats`:

  ```
  library(stats)
  ```

  Este paquete es uno de los más básicos para este análisis. Algunas de sus funciones que incluye son `hclust` y `kmeans`.

  Sin embargo, para un análisis más en profundidad existen diferentes paquetes como:

  ```
  library(cluster)
  library(clues)
  library(bayesclust)
  library(pvclust)
  ```

```
library(clustofvar)
library(NbClust)
```

- El paquete `cluster`:

```
library(cluster)
```

Este paquete nos puede ayudar a hacer un análisis más profundo. De él, según la página web R-Pubs [25], podemos destacar dos funciones:

1. La función `agnes`: La usaremos cuando queramos realizar el método jerárquico aglomerativo.

```
agnes(x=datos,stand= TRUE, metric = "euclidean",method = "
    ward")
```

2. La función `diana`: La usaremos cuando queramos realizar el método jerárquico disociativo.

```
diana(datos, stand = TRUE,metric = "euclidean" )
```

- El paquete `graphics`:

```
library(graphics)
```

Este paquete lo usamos para hacer gráficos, en este caso, por ejemplo, un dendrograma.

- Otros paquetes para construir gráficos, según R-Documentation [26], son:

```
library(factoextra)
```

De este paquete usaremos varias funciones.

Haremos uso de la función `fviz_dist` para representan la matriz de distancias, explicaremos, a continuación, sus argumentos. También aplicaremos `fviz_dend`, que nos devuelve un dendograma separando cada clusters por colores. Además, usaremos la función `fviz_cluster` representa los clusters formados. Y por último, usaremos `fviz_nbclust` que nos ayudará a determinar el número óptimo de clusters, más tarde, veremos las distintas salidas que nos proporciona.

```
library(dendextend)
```

De este paquete haremos uso de la función `tanglegram`, la cual nos ayudará a comparar dendogramas.

```
library(ggradar)
```

De este paquete haremos uso de la función `ggradar`, la cual nos muestra un gráfico con varias líneas indica los diferentes casos que se presentan. Profundizaremos en ella más adelante.

```
library(randomForest)
cm <- table(clusterward,km$cluster)
cm
clusterward 1  2  3  4  5
          1  0  0  0 40  3
          2  7 46  3  4  0
          3 44  0  2  2  5
          4  5  0 11 17 79
          5  5 20 56 17  2
```

Haciendo uso del paquete `randomForest`, vamos a mostrar un cruce de métodos según la página R-Pubs [25]. Aquí nos muestra donde clasifica cada método a cada individuo.

5.2.1. Cálculo de la matriz de distancias

Función `dist`. Esta función, que según Aldás y Uriel [1], nos permite calcular la matriz de distancias para los elementos de la matriz de datos original. Devuelve el triángulo inferior de la matriz de distancias:

```
dist(x,method="euclidean",diag=FALSE,upper=FALSE,p=2)
```

donde:

- `x` puede ser una matriz o una hoja de datos númericos.
- `method` es la medida de asociación que vamos a usar que puede ser `euclidean`, `maximum`, `manhattan`, `canberra`, `binary` o `minkowski`.
- `diag` es el valor lógico que indica si la diagonal de la matriz de distancia debe ser impreso por `print.dist` Por defecto el valor es `FALSE`, por lo que no se imprime.
- `upper` se iguala al valor lógico por defecto `FALSE`. Indica si el triángulo superior de la matriz de distancias debe ser impreso.

- `p` es la potencia en el caso de la distancia de Minkowski.

En caso de tener que estandarizar nuestros datos, según Aldás y Uriel [1], podemos hacer uso de la función `scale(x)`.

Función `fviz_dist`. Esta función es muy útil para representar matrices de distancias. Está dentro del paquete `factoextra`. Con ella, según la página web R-Pubs [25], visualizamos un gráfico en el que vemos varios colores. Basándonos en una escala de colores dependiendo de si el valor es mayor o menor, podemos ver si la distancia entre esos individuos es mayor o menor.

```
fviz_dist(dist, gradient = list(low = "white", mid="cadetblue2"
   ,high="blue"))
```

donde:

- `dist` es la matriz de distancias que hemos calculado.
- `gradient` es una lista de los colores que queremos para una escala de distancias.

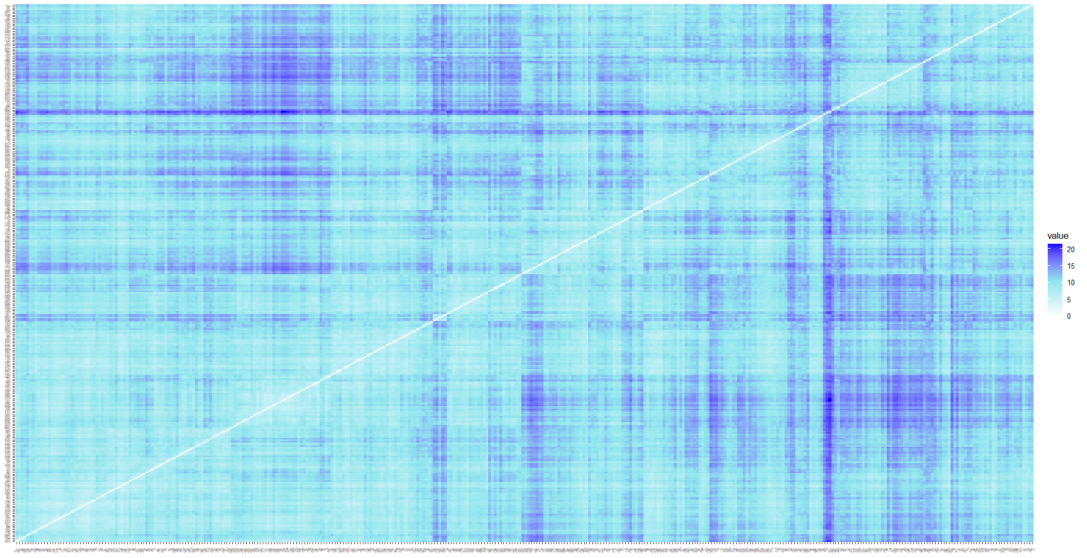


Figura 5.1.: Representación de la distancia euclídea de los datos

La representación que nos daría sería la Figura 5.1. Como vemos el color blanco que presenta en diagonal, que indica que la distancia entre los individuos es 0, nos revela que es el mismo individuo. En los casos que solemos ver, los individuos que son iguales están la diagonal principal.

En los casos de los mapas de calor, como es este, según vemos en R-Pubs [25], se muestran en el otro sentido. Es decir el primer individuo ubicado en la primera posición del eje horizontal, en los casos que solemos ver, coincidiría con la posición más alta del eje vertical (la distancia sería 0 y sería él mismo), en los mapas de calor, es al contrario. Por ello, el primer individuo del eje horizontal es el mismo que el individuo más bajo del eje vertical.

5.2.2. Funciones del paquete básico `stats`

Este paquete contiene funciones tanto para el análisis jerárquico como el no jerárquico, por ello haremos distinción en dos subapartados. Explicaremos cada una de las funciones según la página web R-Blogger [22].

5.2.2.1. Método jerárquico

Función `hclust`. Esta función usar el método del vecino más cercano o más lejano o cualquier otro tipo de enlace de clusters. Hemos de destacar que hay que partir de la matriz de distancias o similaridades calculada con `dist`:

```
hclust(d,"method=complete")
```

donde `d` es la matriz de distancias que se puede calcular con la función `dist` y `method` que se puede usar: `ward.D`, `ward.D2`, `single`, `complete`, `average`, `median` o `centroid`.

Esta devuelve un objeto que proporciona los siguientes elementos:

- `merge`: nos devuelve una matriz de dimensión $(m-1) \times 2$ donde cada fila explica la fusión de los clusters en las sucesivas etapas.
- `height`: es un conjunto de $n-1$ valores reales que proporciona los valores de las alturas de las distancias de la fusión.
- `Labels`: son las etiquetas de cada uno de los elementos que se agrupan.
- `call`: nos devuelve la llamada que produjo ese resultado.
- `method`: indica el método que se ha usado.
- `dist.method`: indica la distancia que se ha utilizado para crear `d`.

Función `cutree`. Esta función permite cortar los árboles a partir de objetos que se obtienen de la función `hclust`. La función es:

```
cutree(tree,k=NULL,h=NULL)
```

donde:

- `tree`: es el árbol que obtenemos de la función `hclust`.
- `k`: es el número entero de clusters que queremos hacer.
- `h`: es el entero o vector con las alturas donde ha de cortarse el árbol.

Función `rect.clust`. Esta función sirve para representar en el dendograma unos rectángulos para diferenciar los grupos. La función es:

```
rect.clust(tree,k=NULL,which:NULL, x=NULL, h=NULL, border=2,
   cluster=NULL)
```

donde:

- `tree`: es el árbol que obtenemos de la función `hclust`.
- `k` y `h`: valores númericos para indicar el corte del grupo.
- `which` y `x`: son los vectores que seleccionan los grupos sobre los que dibujan el rectángulo (`which` los selecciona por números y `x` por las respectivas coordenadas horizontales.
- `border`: es el vector que pinta de colores los bordes de los rectángulos.

5.2.2.2. Método no jerárquico

Función `kmeans`. La sintasis de esta función es:

```
kmeans(x,centers,iter.max=10,nstart=1,algorithm="Hartingan-Wong
   ",trace=FALSE)
```

donde:

- `x`: es la matriz de datos.
- `centers`: se puede indicar un valor k de número de grupos o conjunto de centros iniciales o semillas para la primera partición de grupos.
- `iter.max`: es el número máximo de iteraciones que se pueden hacer.

- `nstart`: es el número de veces que se va a realizar el proceso, cada vez con una asignación aleatoria diferente. Se recomienda que sea elevado.
- `algorithm`: son los diferentes métodos de la k-`medias` que podemos seleccionar.
- `trace`: en caso de tomar el valor `TRUE`, se devuelve un seguimiento de las iteraciones del proceso.

Esta función, nos devuelve un objeto con los siguientes resultados, que son los principales que produce el proceso.

- `clusters`: es un vector de números enteros que indica a qué cluster es asignado cada punto.
- `centers`: es una matriz de los centroides de los clusters.
- `totss`: es la suma total de cuadrados.
- `withinss`: es el vector de suma de cuadrados dentro de los grupos.
- `tot.withniss`: es la suma total de los `withniss`.
- `betweens`: es la suma de los cuadrados entre los grupos.
- `size`: es el tamaño, el número de elementos en cada grupo.
- `iter`: es el número de iteraciones realizadas.
- `infault`: es el indicador de un posible error en el algortimo.

5.2.3. Funciones para gráficos

Función `plot`. Esta función se aplica a la salida de la función anterior, `hclust`. Se usa para hacer gráficas, en este caso, nos devuelve un dendrograma.

Función `ggradar`. Está función está dentro del paquete `ggradar`. La sintasis de esta función, según R-Charts [24], es:

```
# install.packages("devtools")
# devtools::install_github("ricardo-bion/ggradar")
library(ggradar)

ggradar(df)
```

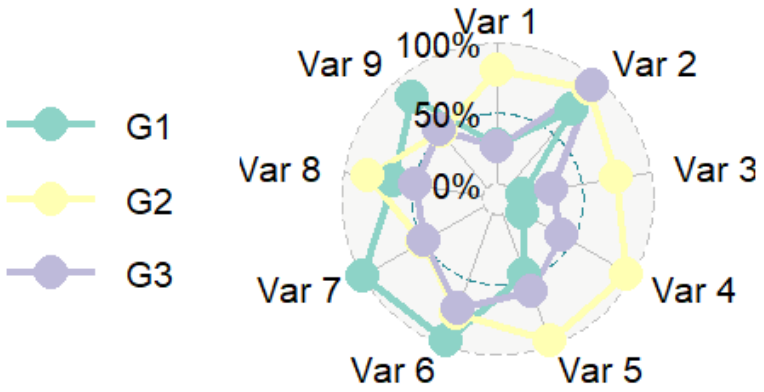


Figura 5.2.: Gráfico de radar

El objeto `df` es un `data.frame` que contiene tantas fila como grupos haya con los valores medios de cada uno de ellos. En el siguiente ejemplo tendríamos 3 grupos que representar.

```
set.seed(4)
df <- data.frame(matrix(runif(30), ncol = 10))
df[, 1] <- paste0("G", 1:3)
colnames(df) <- c("Grupo", paste("Var", 1:9))
ggradar(df)
```

Esta dibuja el gráfico radar como en la Figura 5.2. Según la página de R-Charts [24], consiste en dibujar varias líneas a partir los grupos que tengamos en un marco de datos. Debe estar compuesto por más de tres variables como ejes y las filas indican casos como series.

5.3. Funciones de R para la validación del análisis cluster

La validación de este tipo de análisis se va a llevar a cabo tanto gráficamente como utilizando los coeficientes definidos anteriormente en el apartado Subsección 4.1.4, observando la elección del número adecuado de grupos y de metodología.

Para ello, primeramente, verificaremos hasta qué punto el dendograma que se realiza refleja sus datos. Para ello, según Kassamabra [14], calculamos la correlación entre las distancias cofenéticas y las distancias originales.

Si la agrupación es válida la correlación debería de ser fuerte.

Vamos a medir el coeficiente de correlación coefenético, cuya función, según Kassamabra [14] es `cophenetic`.

También haremos uso de la función `cor` que sirve para calcular la correlación.

```
distancias
metodoward<-hclust(distancias,method="ward.D")
cof1<-cophenetic(metodoward)
cor(distancias, cof1)
```

donde:

- *distancias*: es la matriz de distancias que hemos calculado.
- $cof1$: es es coeficiente de correlación coefenético con un método concreto.

Un coeficiente cercano a 1, indica que representa en buenas condiciones el agrupamiento.

Además, podremos hacer comparaciones de los dendogramas, según Kassambra [14], creando una lista de los dendogramas jerárquicos, para ello, usaremos el paquete `dendextend` :

```
library(dendextend)
hc1 <- hclust(distancias, method = "average")
hc2 <- hclust(distancias, method = "ward.D2")
dend1 <- as.dendrogram (hc1)
dend2 <- as.dendrogram (hc2)
dend_list <- dendlist(dend1, dend2)
tanglegram(dend1, dend2)
cor_cophenetic(dend1, dend2)
```

Con esto, tendremos una comparación de ambos método con el gráfico con nos da la función `tanglegram` podremos ver el cruce de dos dendogramas. Y con la correlación podremos ver si se parecen o no los dendogramas, mientras más se parezcan más cerca de 1 estarán.

Para la elección del número de cluster podemos hacer uso del paquete `factoextra`. Dentro de este vamos a usar la siguiente función:

```
library(factoextra)
fviz_nbclust(x, FUNcluster=kmeans, method = c("silhouette", "wss", "
    gap_stat")
```

donde, según Kassambra [14],

- `x`: es la matriz de datos.
- `FUNcluster`: es una función de partición. Los valores permitidos son para los métodos jerárquicos: `kmeans`, `pam`, `clara` y `hcut`.
- `method`: es el método que vayamos a usar al que queremos determinarle el número óptimo de cluster.

Parte II.

Aplicación a Datos sobre Proyectos STEM en Educación Secundaria

6. Descripción e implementación de los datos

Los datos reales sobre los que se va a trabajar han sido recopilados a través de un cuestionario que fue presentando en el TFG de [27]. Este cuestionario se ha pasado a 368 alumnos de diferentes institutos. En este cuestionario se pregunta, primeramente, la participación en los proyectos STEAM, la cual nos ayudará a conocer el impacto en el alumnado. Para ello, seguidamente, se cuestiona, sobre el interés en las asignaturas pertenecientes al ámbito de las ciencias, así como el interés en general en las ciencias. Por último, también se plantean cuestiones sobre sus perspectivas en un futuro.

Las preguntas que se han realizado se van a separar en 7 bloques:

- Bloque 0: Recoge las tres primeras preguntas que son el centro en que estudia, el sexo y la participación en algún proyecto STEAM.
- Bloque 1: Este bloque es sobre las **Matemáticas**. En él se recogen preguntas sobre la dificultad y el gusto que tienen por ellas.
- Bloque 2: Este bloque es sobre las **ciencias**. En él se recogen preguntas sobre el gusto y la dificultad, como en el anterior bloque, y además sobre la utilidad en un futuro.
- Bloque 3: Este es el bloque sobre **Ingeniería y Tecnología**. En él se recogen preguntas sobre el gusto que tienen por ellas, y las aplicaciones futuras a las que pueden aspiran en este campo.
- Bloque 4: En este bloque se recogen preguntas sobre las **Habilidades del siglo XXI** a la hora de movitar y trabajar solo y en equipo.
- Bloque 5: Este bloque se llama: **Tu futuro**. En él se recoge el interés que tienen los alumnos en dedicarse en su futuro a diferentes ramas de las ciencias.
- Bloque 6: En el último bloque se pregunta sobre cómo esperan que se le den distintas asignaturas al siguiente año, si quieren dar clases avanzadas y si conocen a científicos. Este se llama **Acerca de ti**.

En nuestro caso, antes de empezar, hemos recodificado los datos, transformando las respuestas cualitativas a numéricas siguiendo los valores que se resumen en la Tabla 6.1

Bloque	Respuesta	Valor numérico
Bloques 1, 2, 3 y 4	Muy en desacuerdo	1
	En desacuerdo	2
	Ni de acuerdo ni en desacuerdo	3
	De acuerdo	4
	Muy de acuerdo	5
Bloque 5	No me interesa en absoluto	1
	No me interesa mucho	2
	Me interesa	3
	Me interesa mucho	4
Bloque 6	No muy bien	1
	Bien/Bastante bien	2
	Muy bien	3
Bloque 6	No	0
	No estoy seguro	0.5
	Sí	1

Tabla 6.1.: Recodificación a valores numéricos de las respuestas de la encuesta

En primer lugar, para comenzar este análisis en R, es recomendable cargar, inicialmente, todos los paquetes que se vayan a utilizar para tener mayor comodidad y organización.

```
library(stats)
library(factoextra)
library(dendextend)
library(ggradar)
```

Después, leer el fichero de datos almacenado en un Excel.

```
library(readxl)
Cuestionario<- read_excel("C:/Users/rosin/OneDrive/Escritorio/ESTADÍ
   SICA UGR/AA CUARTO/TFG/Cuestionario (respuestas).xlsx")
```

En el objeto `Cuestionario`, tenemos 368 alumnos y, de cada uno de ellos, 61 variables.

Para agrupar a los alumnos, inicialmente, no vamos a usar las preguntas del bloque 0. Esta información la dejaremos reservada para usarla en posteriores análisis, cuando ya estén hechos los grupos.

```
datos<-Cuestionario[,-(1:3)]
total<-368*58
sum(is.na(datos))/total*100
## [1] 0
```

En el objeto `datos` hemos guardado el objeto `Cuestionario` eliminando el bloque 0. También, hemos comprobado que no hay datos faltantes en el objeto `datos`.

Ahora, haciendo uso de la función `str`, podemos ver el tipo de variables que tenemos:

```
str(datos)
## tibble [368 x 58] (S3: tbl_df/tbl/data.frame)
##  $ B1-Var1 : num [1:368] 2 3 4 5 1 5 5 4 4 4 ...
##  $ B1-Var2 : num [1:368] 4 4 1 1 4 4 1 2 4 1 ...
##  $ B1-Var3 : num [1:368] 3 3 4 5 2 5 5 4 4 5 ...
##  $ B1-Var4 : num [1:368] 4 4 2 1 5 2 1 3 2 1 ...
```

Todas las variables son del mismo tipo, ya que como las respuestas han sido recodificadas, se obtiene que todas son de tipo numérico.

6.1. Cálculo de la matriz de distancias

El análisis cluster con R, se realiza sobre la matriz de distancias.

En este caso, queremos calcular la distancia entre los individuos y los datos que tenemos son de tipo numérico. Por ello, utilizaremos la distancia euclídea, la medida más común que se emplea, según Cuadras [7],cuando se trabaja con datos numéricos. Esta métrica de distancias nos proporciona una medida de las diferencias entre individuos. La distancia euclídea ha sido definida en apartados anteriores (3.1).

```
distancias<-dist(datos, method = "euclidean")
head(distancias)
## [1] 11.575837 12.257651 10.500000 8.261356 11.715375 11.968709
fviz_dist(distancias,gradient=list(low = "white", mid="cadetblue2",
    high="blue"))
```

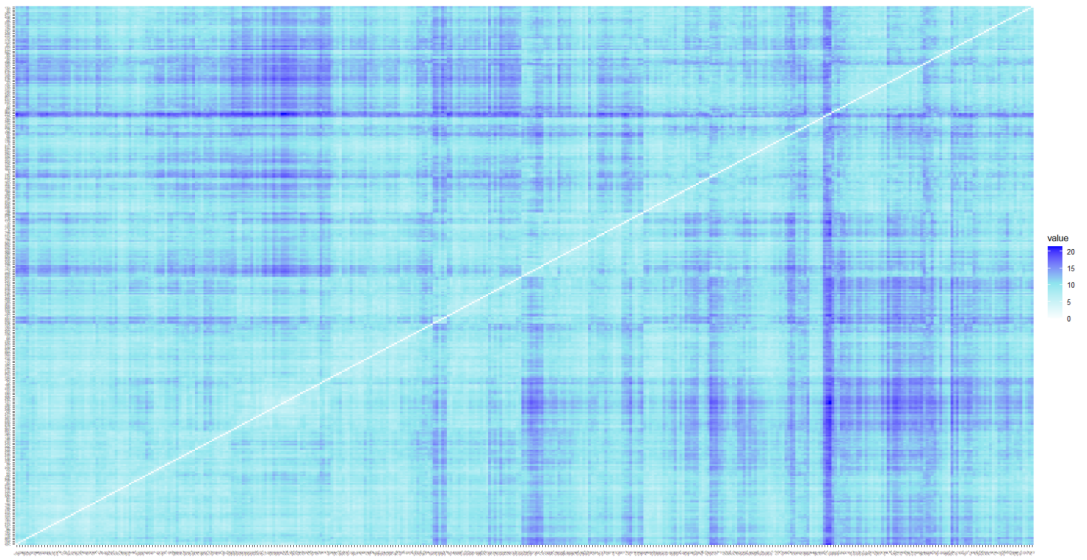


Figura 6.1.: Representación de la matriz de distancias euclídeas entre los individuos

No hemos mostrados todos los valores de la matriz distancias debido a su gran tamaño. En la Figura 6.1, podemos ver la representación de la matriz de distancias. Como indica la leyenda, identificamos de color azul fuerte las distancias más grandes, de azul claro las distancias medianas y mientras más se van acercando a 0, se van mostrando de color blanco. El color blanco que presenta en diagonal, que indica que la distancia entre los individuos es 0, nos revela que es el mismo individuo.

Una vez que tenemos la matriz de distancias, como hemos indicado, se realizará sobre ella el análisis cluster.

7. Resultados

7.1. Resultados con métodos jerárquicos y no jerárquicos

Para los primeros resultados probaremos y compararemos algunos **métodos jerárquicos**. Así, haremos uso de la función definida `hclust`. Emplearemos cuatros métodos de lo que presentamos en la Tabla 4.1, concretamente el vecino más próximo (*single*), el vecino más lejano (*complete*), el método de la media no ponderada (*average*) y el método de Ward (*Ward.D*).

```
linkagecompleto<-hclust(distancias,method="complete")
plot(as.dendrogram(linkagecompleto))
abline(h=14,col="red")
metodoward<-hclust(distancias,method="ward.D")
plot(as.dendrogram(metodoward))
abline(h=30,col="red")
linkagesimple<-hclust(distancias,method="single")
plot(as.dendrogram(linkagesimple))
abline(h=5,col="red")
medianoponderada<-hclust(distancias,method="average")
plot(as.dendrogram(medianoponderada))
abline(h=10,col="red")
```

Las soluciones de este análisis, según la elección subjetiva, las veremos en los dendogramas de las figuras: Figura 7.1, Figura 7.2, Figura 7.3 y Figura 7.4.

Las líneas horizontales que se han elegido son orientativas, y con ellas podemos ver cuántos clusters habría para cada método. Como se puede observar, es difícil elegir un número de clusters simplemente mirando el dendograma con el método de linkage simple en la Figura 7.3 o determinarlos con el método de la media no ponderada representado en la Figura 7.4. Los clusters se definen de mejor manera en los dendogramas del método de linkage completo (Figura 7.1) y del método de Ward (Figura 7.2).

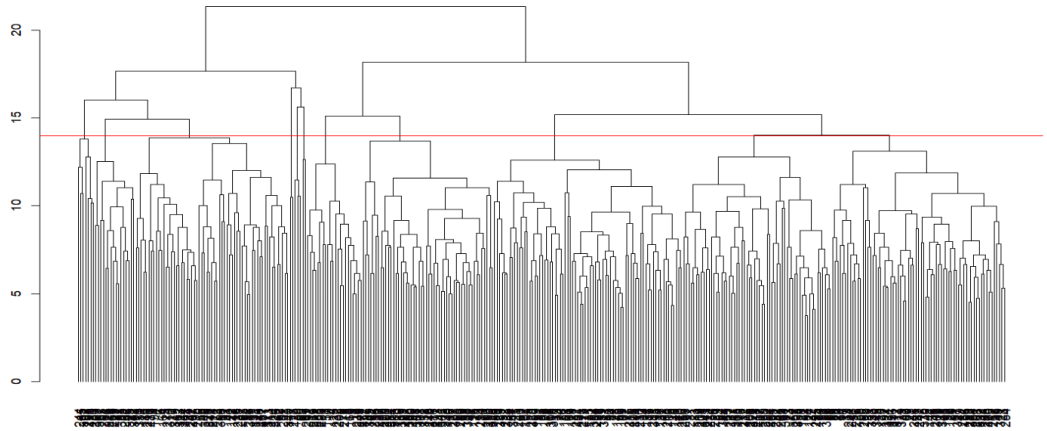


Figura 7.1.: Representación del dendograma con el método de linkage completo

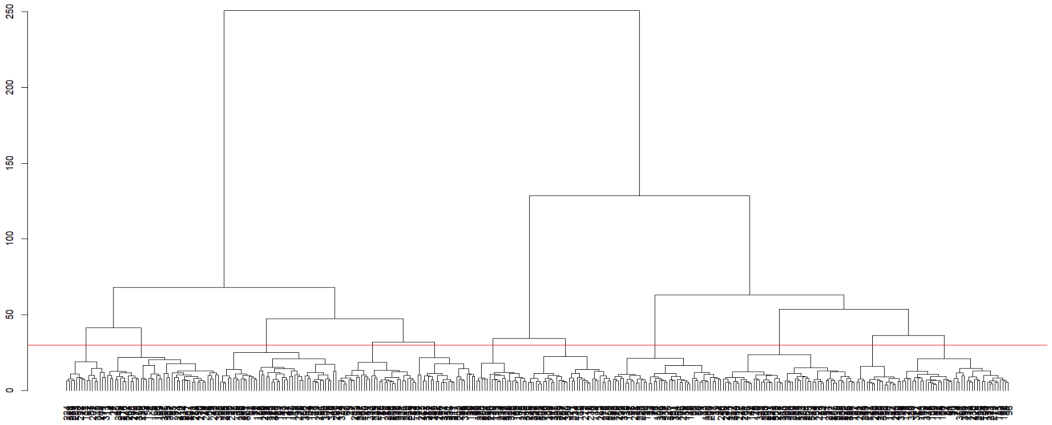


Figura 7.2.: Representación del dendograma con el método de Ward

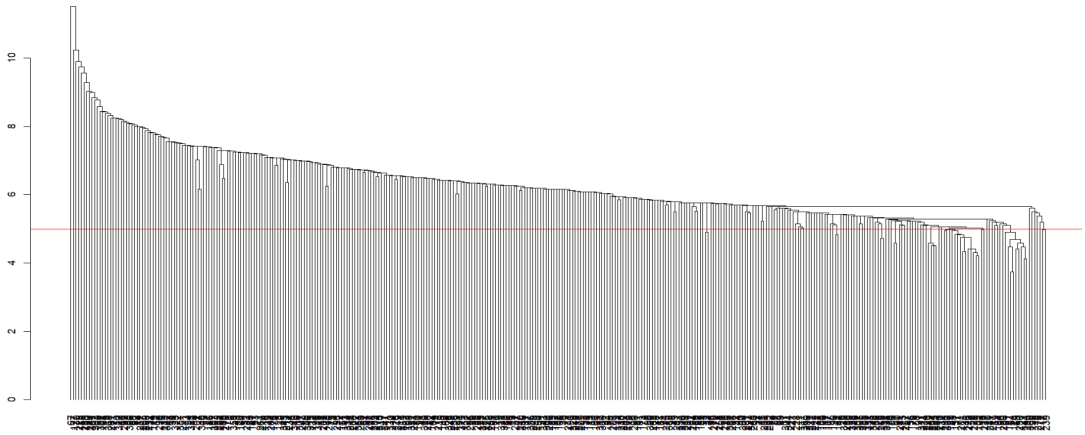


Figura 7.3.: Representación del dendograma con el método de linkage simple

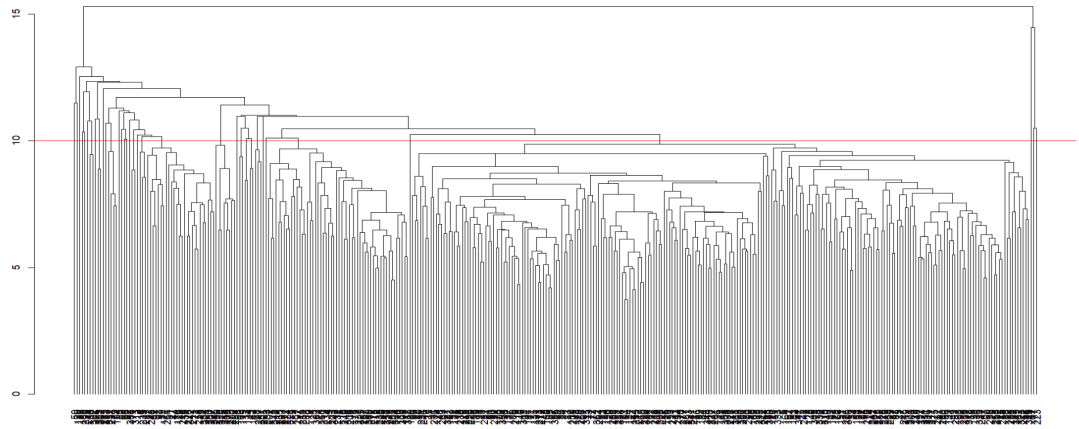


Figura 7.4.: Representación del dendograma con el método de media no ponderada

Con ambos, podríamos destacar de primeras, la agrupación en 2 grandes clusters, y con ellos parece más que suficiente. Sin embargo, esta solución es trivial; por ello, vamos a intentar dar, si es posible, un paso más: buscaremos agrupar los alumnos en más clusters. Basando nuestra elección en las líneas orientativas, podríamos agrupar, según la elección subjetiva: con el método de linkage completo en 12 clusters y con el método de Ward en 11 clusters.

A continuación, en la Figura 7.5 y la Figura 7.6, se grafican los dendogramas aplicando a cada uno de los grupos, que hemos elegido subjetivamente mirando los gráficos, un color. Para ello, haremos uso de la siguiente función

del paquete `factoextra`.

```
fviz_dend(linkagecompleto, cex = 0.5,
k =12, # grupos
palette = "jco" # Color
)
fviz_dend(metodoward, cex = 0.5,
k = 11 , # grupos
palette = "jco" # Color
)
```

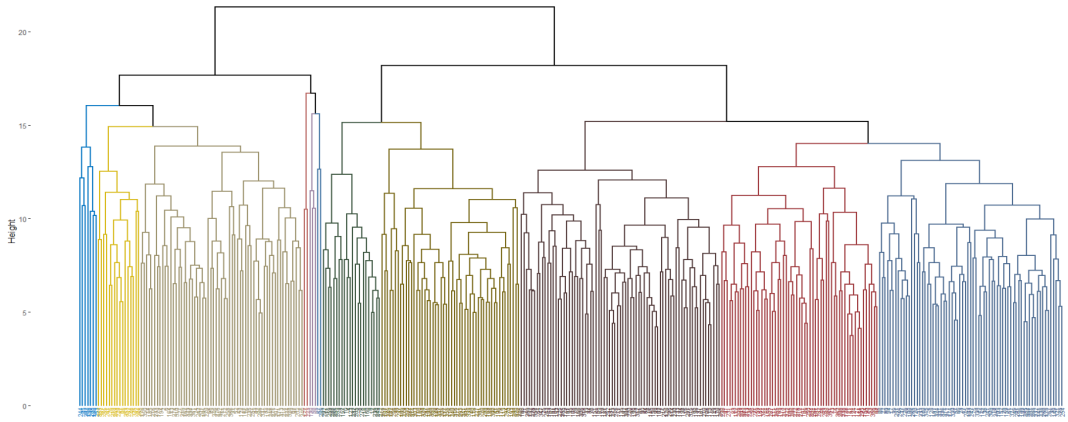


Figura 7.5.: Clusters con el método del linkage completo

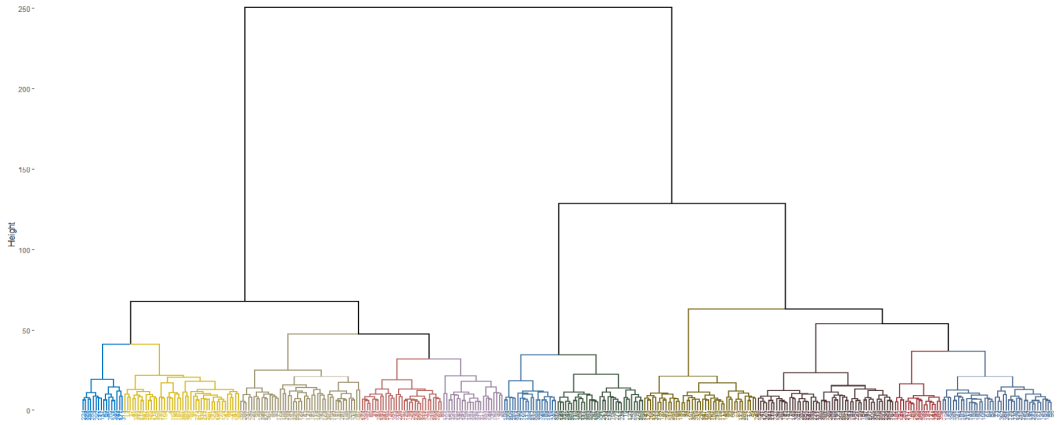


Figura 7.6.: Clusters con el método de Ward

En la Figura 7.5, podemos ver la manera en la que se agrupan los 368 individuos según el método de linkage completo en 12 clusters, cada color es un cluster. Análogamente vemos el resultado del método de Ward, agrupados en 11 clusters, en la Figura 7.6.

7.1.1. Idoneidad del modelo

Una opción para determinar que método de análisis cluster es más adecuado es comparar dendogramas y ver la correlación entre ellos.

```
hc1 <- linkagecompleto
hc2 <- metodoward
dend1 <- as.dendrogram (hc1)
dend2 <- as.dendrogram (hc2)
dend_list <- dendlist(dend1, dend2)
cor.dendlist(dend_list, method = "cophenetic")
tanglegram(dend1, dend2)

 #          [,1]      [,2]
#[1,] 1.0000000 0.3747051
#[2,] 0.3747051 1.0000000
```

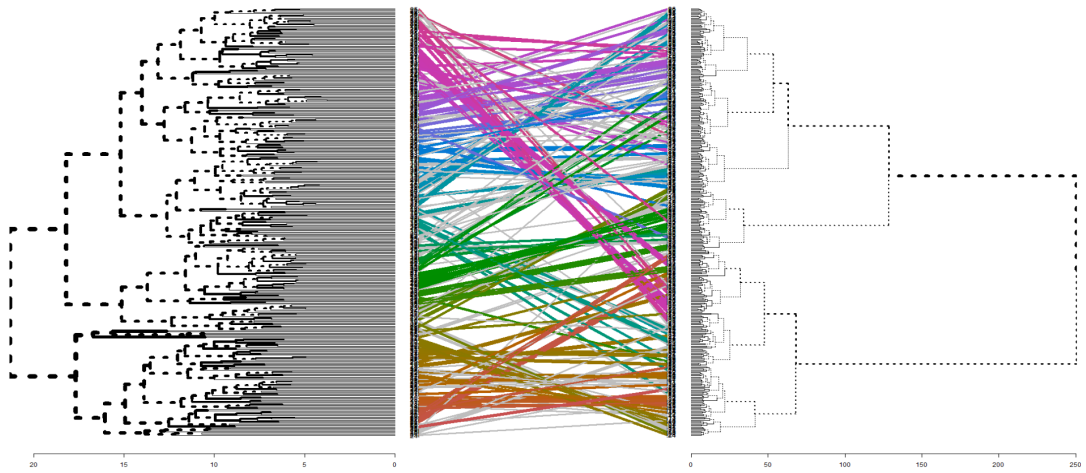


Figura 7.7.: Representación de tanglegram

Como podemos ver en la representación del cruce de los dos dendogramas, en la Figura 7.7, se muestran bastantes lineas que se cruzan, lo cual implica

muchas diferencias entre ambos. Además, al ver el coeficiente de correlación coefenético de ambos, 0.3747051, podemos decir al estar más cercano a 0 que, según el libro de Kassamabra [14], los dos dendogramas no son estadísticamente similares.

Otra opción que aplicaremos para el estudio de la correlación entre la matriz de distancias inicial y la matriz cofenética, de forma que cuanto mayor sea el valor de este coeficiente, mejor se está obteniendo la estructura subyacente en grupos con el análisis.

```
coecof1<- cophenetic(linkagecompleto)
coecof2<- cophenetic(metodoward)
coecof3<- cophenetic(linkagesimple)
coecof5<- cophenetic(medianoponderada)

list(c("complete",cor(distancias, coecof1)),
c("single",cor(distancias, coecof2)),
c("average",cor(distancias, coecof2)),
c("ward.D",cor(distancias, coecof5)))

## [[1]]
## [1] "complete"       "0.507315470820868"
##
## [[2]]
## [1] "single"         "0.366968173808702"
##
## [[3]]
## [1] "average"        "0.366968173808702"
##
## [[4]]
## [1] "ward.D"         "0.645122592318792"
```

Cuanto más próximo sea este coeficiente a 1, mejor será el análisis.

Como vemos, había que descartar el método del linkage simple y de la media no ponderada. Entre el método de linkage completo y de Ward, nos quedaríamos con el de Ward.

7.1.2. Elección del número óptimo de clusters

Antes de pasar a los resultados no jerárquicos, y explicar las características de cada grupo, debemos de comprobar si hemos elegido el número óptimo de grupos.

Para ello, usaremos la función `fviz_nbclust` del paquete `factoextra`.

```
fviz_nbclust(datos,FUNcluster=kmeans,method = c("silhouette"))+labs(
   subtitle = "Silhouette method")
fviz_nbclust(datos,FUNcluster=kmeans,method = c("wss"))+labs(
   subtitle = "Elbow method")
fviz_nbclust(datos,FUNcluster=kmeans,method = c("gap_stat"))+labs(
   subtitle = "Gap statistic method")
```

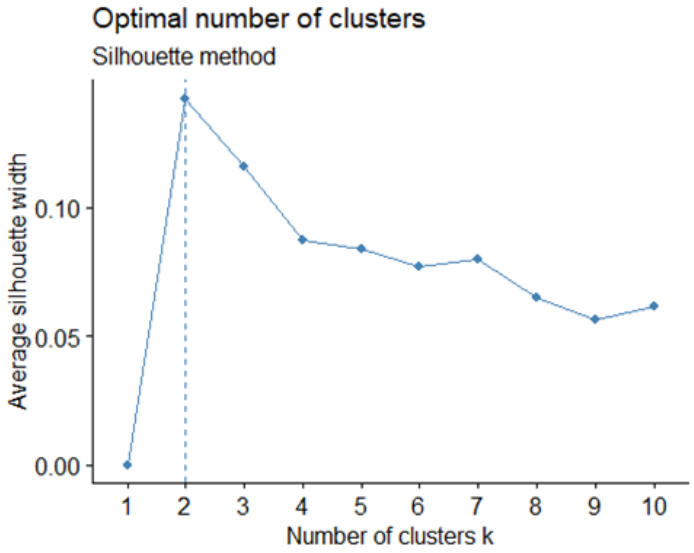


Figura 7.8.: Coeficiente de Silueta

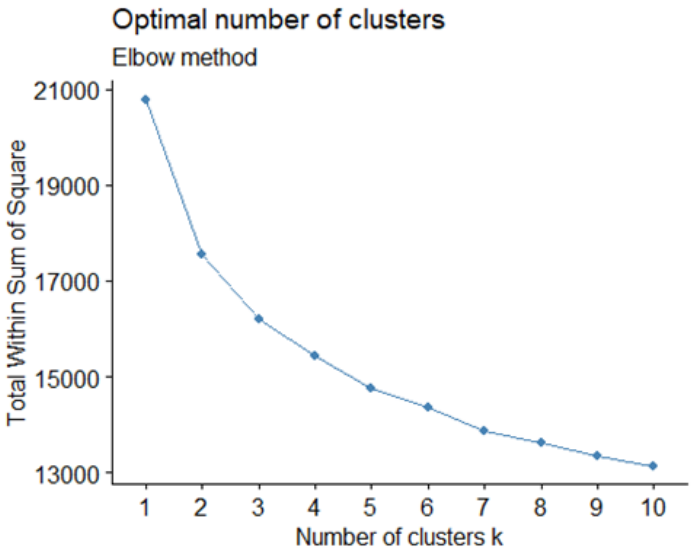


Figura 7.9.: Método del codo

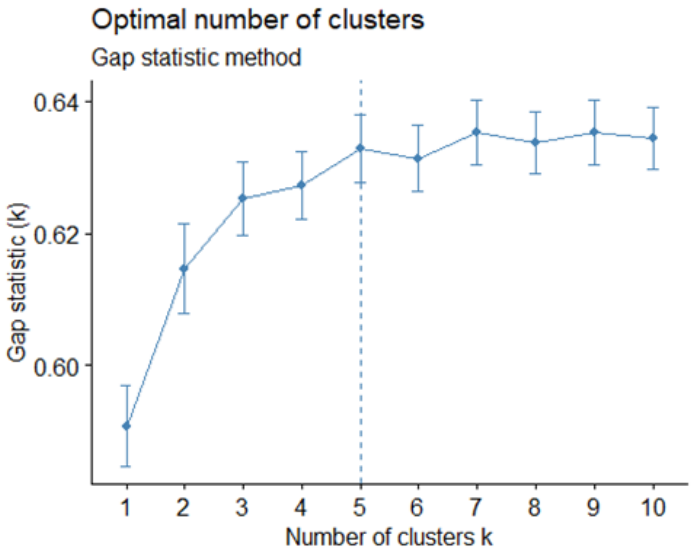


Figura 7.10.: Estadístico GAP

Podemos ver que la Figura 7.8, que es según el coeficiente de silueta, nos indica que usemos 2 clusters, sin embargo, como hemos dicho, esa solución es trivial. La figura del método del codo (Figura 7.9), en este caso, da lugar a confusión con respecto a cuántos clusters seleccionar, ya que no se ve claramente a partir de que número se empieza a estabilizar algo. Nos fijamos, por lo tan-

to, que la figura del estadístico Gap (Figura 7.10), indica que seleccionemos 5 clusters.

Visualizaremos el dendograma con el método de Ward realizando 5 clusters:

```
fviz_dend(metodoward, cex = 0.5,
k =  ,
palette = "jco"
)
```

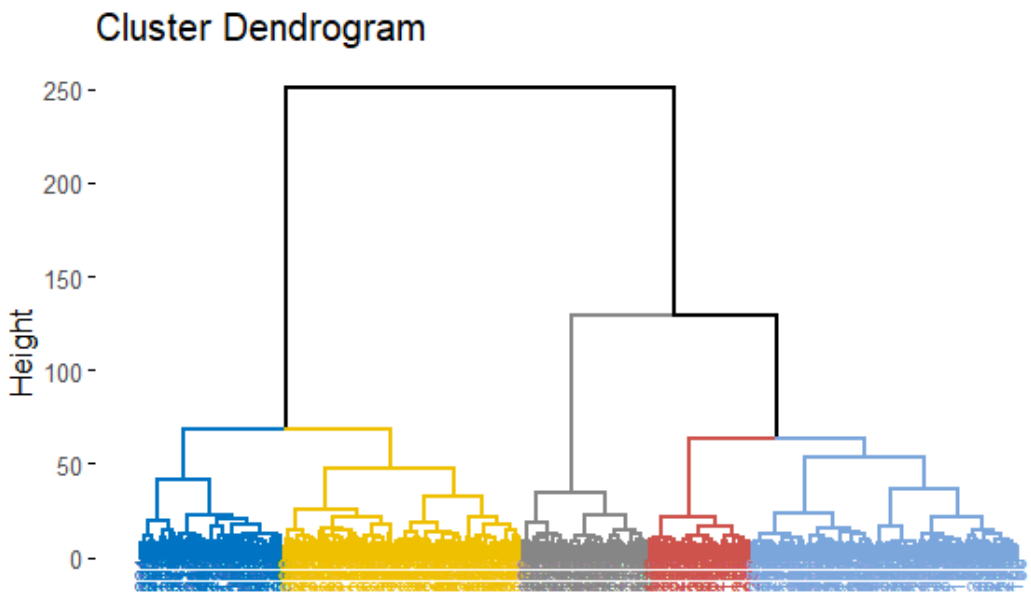


Figura 7.11.: Dendograma con 5 clusters

Con la Figura 7.11, podemos ver cómo se agruparían los 368 estudiantes en 5 grupos. Además con el siguiente código podemos a ver cuántos estudiantes hay en cada grupo:

```
solucionjerarquico <- cutree(metodoward, k =5)
table(solucionjerarquico)
#solucionjerarquico
# 1   2   3   4   5
# 43  60  53 112 100
```

La solución nos da un el primer grupo con 43 estudiantes, un segundo grupo con 60 estudiantes, un tercer grupo con 53 estudiantes, un cuarto grupo con 112 estudiantes y un quinto y último grupo con 100 estudiantes.

Vamos a realizar otro dendrogramas que muestran cómo es la agrupación. Estos se podrán ver con el paquete `factoextra`:

```
fviz_dend(metodoward, cex = 1, type = c("phylogenic"),k=5,k_colors="
    jco",ggtheme = theme_gray())
fviz_dend(metodoward, cex = 1, type = c("circular"),k=5,k_colors="
    jco",ggtheme = theme_gray())
```

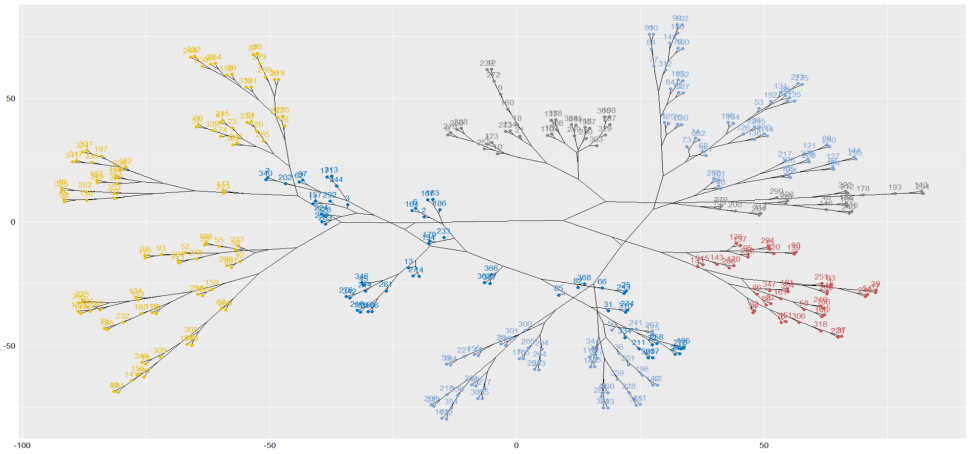


Figura 7.12.: Representación dendograma en árbol con 5 clusters

En la Figura 7.12 y la Figura 7.13 , podemos visualizar cómo vamos a agrupar los 368 individuos en los 5 clusters.

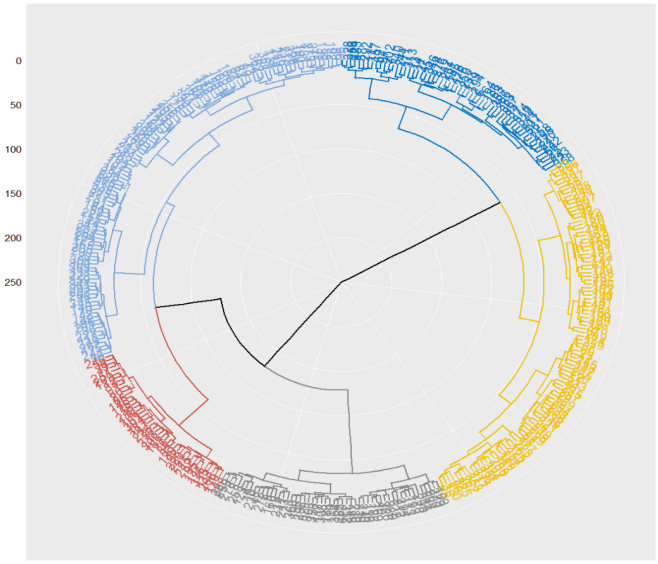


Figura 7.13.: Representación dendograma circular con 5 clusters

Para obtener resultados con **métodos no jeráquicos**, como hemos visto en la teoría, hemos conocer los k clusters en los que vamos a clasificar los individuos. A pesar, de haber visto, gráficamente y por elección subjetiva, que podemos hacer 11 grupos, y de saber que la solución trivial sería hacer 2, vamos a realizarlo con el número óptimo según el gráfico del estadístico GAP (Figura 7.10), que es 5.

Usaremos la función `kmeans` del paquete `stats`:

```
k <- 5 # Número de clústeres
km <- kmeans(datos, centers = k)
km$cluster
#[1] 1 3 2 2 5 3 3 2 2 2 2 2 3 3 5 5 5 2 5 4 2 5 5 3 5 5 3 5
# [29] 5 3 3 5 5 5 3 5 5 4 5 4 4 1 5 1 3 3 4 1 4 3 5 3 1 5 3 4
# [57] 5 5 1 5 1 5 5 5 3 3 3 1 1 4 1 3 1 1 5 5 1 5 5 5 5 3 1 1
# [85] 3 5 3 5 1 1 1 3 3 1 5 1 3 1 1 5 1 1 5 3 1 5 1 5 4 4 4 5
#[113] 4 1 5 4 2 1 5 5 1 3 2 1 4 5 1 1 3 3 1 1 5 5 1 5 5 5 5 4
#[141] 5 1 5 1 5 5 3 2 3 5 5 5 5 1 1 4 2 5 5 2 5 3 3 3 1 2 5 5
#[169] 4 1 2 5 5 5 4 3 3 4 3 2 1 4 4 1 3 5 3 4 4 1 5 1 4 3 3 4
#[197] 4 4 4 4 1 2 1 1 4 1 3 1 1 5 3 1 5 2 5 3 1 1 3 5 1 1 3 2
#[225] 3 5 5 1 5 1 1 2 3 5 5 1 2 5 2 4 5 3 3 3 4 5 5 2 2 1 5 1
#[253] 2 5 2 4 3 3 2 4 3 3 1 1 5 5 4 5 3 3 3 4 1 3 1 3 4 2 2 2
#[281] 5 1 2 4 4 5 4 2 5 4 1 2 4 4 1 1 2 2 4 4 4 3 4 5 1 5 3 2
#[309] 5 3 5 1 3 4 5 4 4 5 3 4 5 4 3 3 3 4 1 2 4 5 4 5 5 3 1 4
```

```
#[337] 5 1 5 2 4 3 1 1 4 3 5 5 3 3 1 3 3 1 3 1 2 2 5 4 3 1 3 5
#[365] 4 3 5 3
```

En esa salida podemos ver a que cluster pertenece cada uno de los estudiantes.

Vamos a visualizarlo con la ayuda de la función `fviz_cluster` del paquete `factoextra`.

```
fviz_cluster(km,datos,palette = c("#2E9FDF", "green", "#E7B800", "#
    FC4E07","pink"),ellipse.type = "euclid", # Concentration ellipse
star.plot = TRUE, # Add segments from centroids to items
repel = TRUE, # Avoid label overplotting (slow)
ggtheme = theme_minimal()
)
```

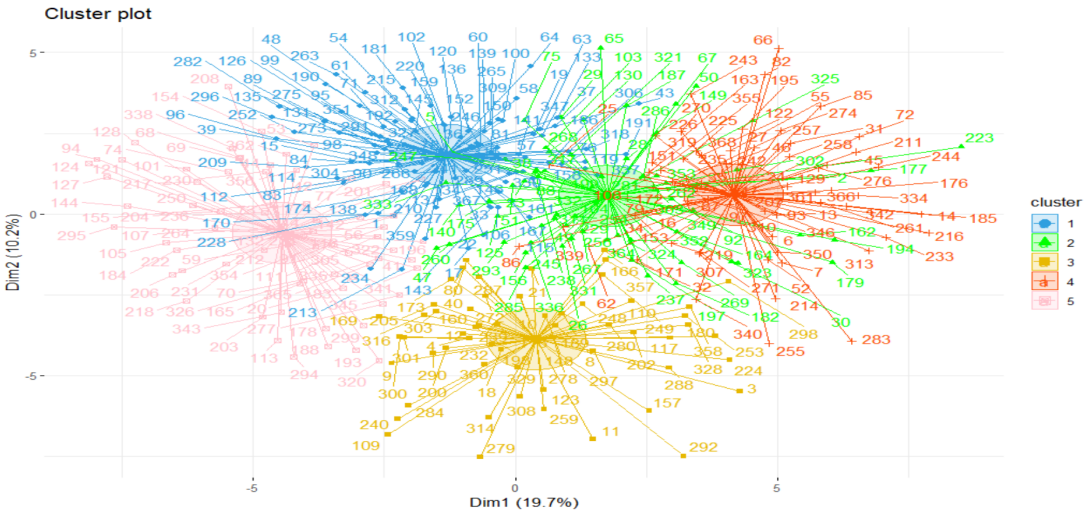


Figura 7.14.: Agrupación con el método de *k*-medias

En la Figura 7.14, podemos ver cómo se agrupan los datos cuando partimos de 5 semillas y quedan pre-establecidos en 5 grupos.

7.2. Comparación de resultados obtenidos

Para visualizar el comportamiento que tienen cada uno de los grupos, veremos la clasificación con el método jerárquico y con el método no jerárquico *k*-medias y las compararemos, haremos uso de la función `ggradar` del paquete `ggradar`:

En primer lugar mostramos el código para hacer el gráfico radar con el método jerárquico de Ward.

```
clusterward<-cutree(metodoward,k=5)
datos_cluster <- data.frame(datos, cluster = clusterward)
dfjerarquico <- aggregate(.~cluster, data = datos_cluster, mean)
lcols <- c("blue", "yellow", "orange","pink","green")
ggradar(dfjerarquico, background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2.5, 5),
        legend.title = "Grupos",
        group.point.size = 3,
        group.colours = lcols)
```

En la Figura 7.15, podemos ver el comportamiento de los 5 grupos según este método jerárquico.

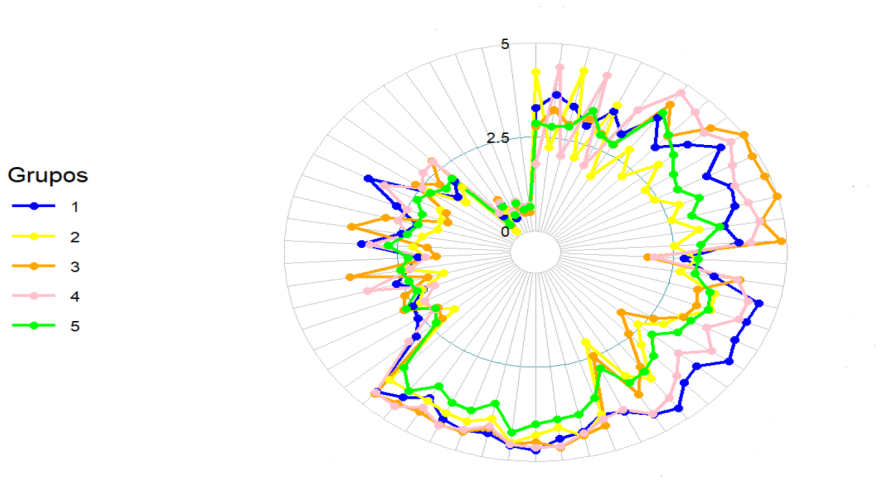


Figura 7.15.: Gráfico radar con método jerárquico

```
km <- kmeans(datos, centers = 5)
datos$cluster <- km$cluster
dfnojerarquico <- aggregate(.~cluster, data=datos, mean)
lcols <- c("blue", "yellow", "orange","pink","green")
ggradar(dfnojerarquico,
        background.circle.colour = "white",
        gridline.min.linetype = 1,
```

```
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2.5, 5),
        legend.title = "Grupos",
        group.point.size = 2.5,
        group.colours = lcols)
```

En la Figura 7.16, podemos ver el gráfico de radar con el método de las *k*-medias con el cual podremos explicar las características de cada grupo.

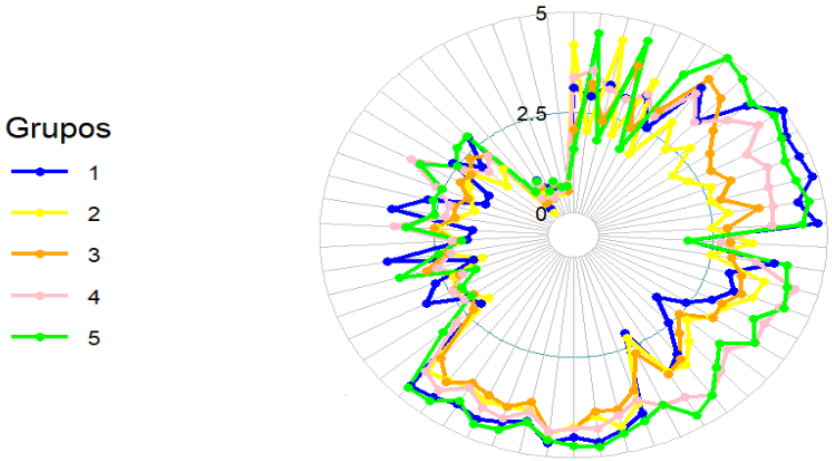


Figura 7.16.: Gráfico radar con método no jerárquico

Sin embargo, para poder realizar la clasificación con mayor exhaustividad, haremos un gráfico radar con cada método para cada uno de los bloques de preguntas.

- Bloque 1: Este primer bloque está compuesto por las siguientes preguntas:
  1. Matemáticas ha sido mi peor asignatura.
  2. Consideraría elegir una carrera en la que se usen las matemáticas.
  3. Las matemáticas me resultan difíciles.

4. Soy el tipo de estudiante al que se le dan bien las matemáticas.
5. Se me dan bien la mayoría de asignaturas, pero no tanto las matemáticas.
6. Con total seguridad, podría hacer trabajos avanzados en Matemáticas.
7. Puedo obtener buenas calificaciones en Matemáticas.
8. Se me dan bien las Matemáticas.

Y las respuestas podían ser (del 0 al 5): *muy en desacuerdo, en desacuerdo, ni acuerdo ni en desacuerdo, de acuerdo, muy de acuerdo.*

```
dfjerarquicob1<-cbind(dfjerarquico[1],dfjerarquico[,2:9])
       group.colours = lcols)
ggradar(dfjerarquicob1, background.circle.colour = "white",
       gridline.min.linetype = 1,
       gridline.mid.linetype = 1,
       gridline.max.linetype = 1,
       values.radar = c(0, 2.5, 5),
       legend.title = "Grupos",
       group.point.size = 3,
       group.colours = lcols)
dfnojerarquicob1<-cbind(dfnojerarquico[1],dfnojerarquico[,2:9])
ggradar(dfnojerarquico,
       background.circle.colour = "white",
       gridline.min.linetype = 1,
       gridline.mid.linetype = 1,
       gridline.max.linetype = 1,
       values.radar = c(0, 2.5, 5),
       legend.title = "Grupos",
       group.point.size = 2.5,
       group.colours = lcols)
```

A partir de la Figura 7.17 y la Figura 7.18, podemos describir cada grupo:

Primero, describiremos los grupos con el método jerárquico, siguiendo el gráfico de la Figura 7.17 y luego con el no jerárquico, siguiendo el gráfico de la Figura 7.18, para finalmente compararlos.

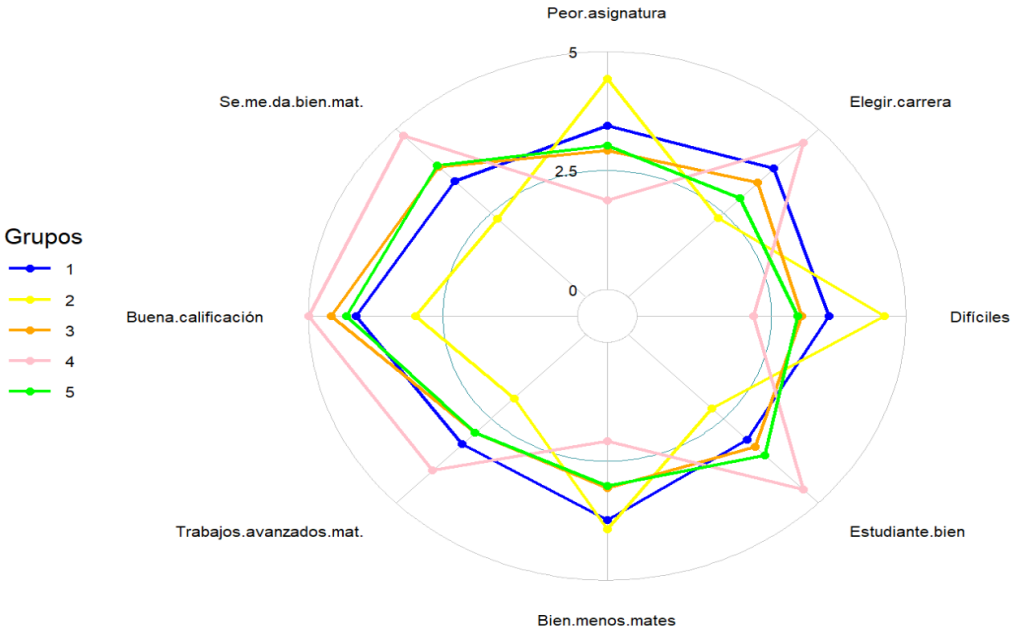


Figura 7.17.: Gráfico radar con método jerárquico bloque 1

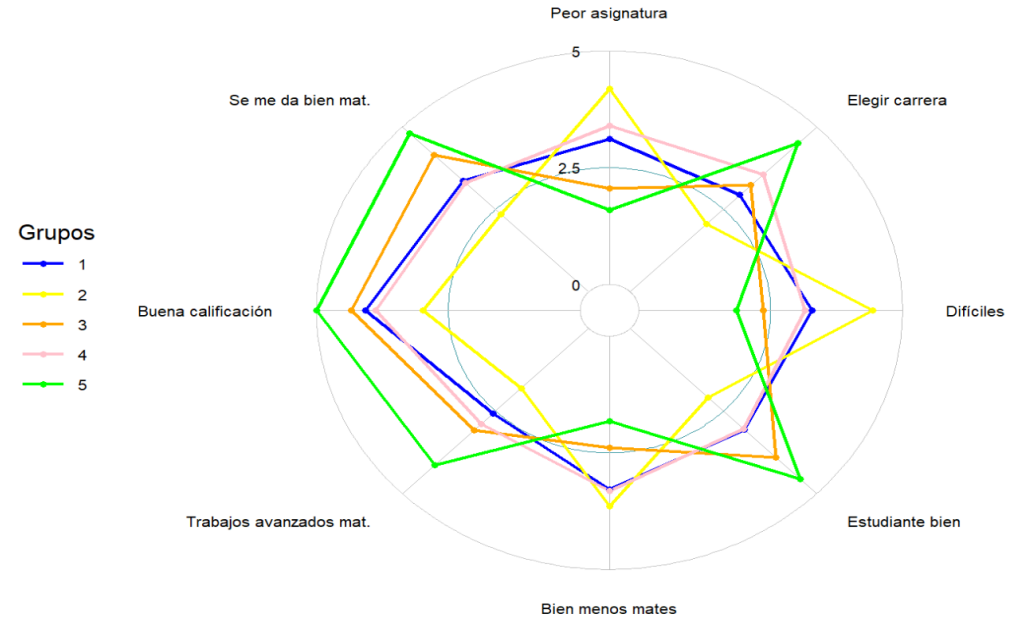


Figura 7.18.: Gráfico radar con método no jerárquico bloque 1

Grupo 1 **jerárquico.** Los escolares que pertenecen a este grupo, en general, muestran un grado de acuerdo no muy alto ni muy bajo en todas las preguntas de este bloque. Podríamos describirlo como un grupo al que se le dan bien las matemáticas, pero, que a su vez, le resulta algo difícil la asignatura. Sin embargo, considerarían el elegir una carrera de matemáticas.

Grupo 2 **jerárquico.** Este grupo es el que está más de acuerdo con que las matemáticas han sido su peor asignatura. A su vez, este grupo es el que menos consideraría estudiar una carrera de matemáticas. También, es al que más difícil le resultan las matemáticas. Es el grupo que está más en desacuerdo con que se le den bien las matemáticas, y además, considera que se le dan bien la mayoría de las asignaturas exceptuando esta. Es el grupo que está más en desacuerdo con las afirmaciones de que son capaces de hacer un trabajo avanzado en matemáticas, que puedan sacar buenas calificaciones y que se le dan bien las matemáticas. En resumen, este grupo es al que más difícil le parecen, peor se le dan y menos le gustan las matemáticas.

Grupo 3 **jerárquico.** Este bloque se muestra que tiene mucha similitud con los grupos 1,3 y el 5. Los estudiantes tienen respuestas que quedan en la media. Se podría destacar la pregunta sobre si considera que puede obtener altas calificaciones en matemáticas, donde tienen una respuesta elevada. En resumen, podríamos decir que en este bloque, es uno de los grupos que presenta acuerdo tibio.

Grupo 4 **jerárquico.** Este grupo es el que está más en desacuerdo con que las matemáticas han sido su peor asignatura. A su vez, este grupo es el que más consideraría estudiar una carrera de matemáticas y al que menos difícil le resultan las matemáticas. Es el grupo que está más de acuerdo con que se le den bien las matemáticas, y por ello, tienen menor acuerdo con la pregunta 4. Presenta el mayor acuerdo con poder hacer un trabajo avanzado en matemáticas, al igual, que con poder sacar buenas calificaciones. En resumen, este grupo es al que menos difíciles le resultan, mejor se le dan y más le gustan las matemáticas.

Grupo 5 **jerárquico.** Este grupo, vemos que, es similar al 1 y al 3. De hecho, en la mayoría de los casos queda entre medias de estos dos grupos mencionados. Podemos decir, que este grupo presenta, también, un acuerdo tibio en este bloque.

En conclusión, con el **método jerárquico**, el grupo 4 están los que más interés tienen en las matemáticas y el grupo 2 es el que menos interés tiene. Los grupos 1, 3 y 5 no muestran muchas diferencias entre sí, conteniendo estudiantes que señalan un acuerdo tibio con las preguntas de este bloque.

Grupo 1 **no jerárquico.** Este grupo vemos que muestra acuerdo en casi todo, en un grado no muy alto. Presenta un poco más de acuerdo en la pregunta 7, pero eso no le hace destacar en comparación a los demás.

Grupo 2 **no jerárquico.** Este grupo es el que está más de acuerdo con que las matemáticas han sido su peor asignatura. A su vez, este grupo es el que menos consideraría estudiar una carrera de matemáticas y al que más difícil le resulta esta disciplina. Es el grupo que está en más en desacuerdo que se le den bien las matemáticas, sin embargo, este grupo considera que se le dan bien la mayoría de las asignaturas exceptuando esta. Es el grupo que está más en desacuerdo con poder hacer un trabajo avanzado en matemáticas, al igual, que con que puedan sacar buenas calificaciones y que se le den bien las matemáticas. En resumen, este grupo es al que más difícil le resultan, peor se le dan y menos le gustan las matemáticas.

Grupo 3 **no jerárquico.** Los escolares de este grupo muestran puntuaciones un poco por debajo de las que muestran los estudiantes del grupo 5. Por ello, podemos considerar que es un grupo al que le gusta las matemáticas y no le parecen muy difíciles.

Grupo 4 **no jerárquico.** En este bloque se muestra un grado de acuerdo similar al del grupo 1. Los estudiantes de este grupo tienen un grado de acuerdo tibio, aunque en algunas preguntas muestran algo más de acuerdo, no destaca por encima de los demás.

Grupo 5 **no jerárquico.** Este grupo es el que está más en desacuerdo con que las matemáticas han sido su peor asignatura. A su vez, este grupo es el que más consideraría estudiar una carrera de matemáticas. También es al que menos difícil le resultan las matemáticas. Es el grupo que está más de acuerdo en que se le dan bien las matemáticas y, por ello, este grupo no considera que se le den bien la mayoría de las asignaturas excepto matemáticas. Es el grupo que está más de acuerdo con poder hacer un trabajo avanzado en matemáticas, al igual, que con que puedan sacar buenas calificaciones y que se le den bien las matemáticas. En resumen, este grupo es al

menos difícil, mejor se le dan y más le gusta las matemáticas.

En conclusión, con ambos métodos se tiene un grupo con respuestas que indican poco aprecio hacia las matemáticas, otro donde están los individuos con mayor interés y afecto hacia ellas, y tres grupos muy similares con individuos que muestran respuestas intermedias.

- Bloque 2: Este segundo bloque está compuesto por las siguientes preguntas:
  1. Me desenvuelvo con seguridad en Ciencias.
  2. Consideraría elegir una carrera de Ciencias.
  3. Espero usar las Ciencias cuando termine los estudios.
  4. Dominar las ciencias me ayudará a ganarme la vida.
  5. En mi futuro trabajo necesitaré las Ciencias.
  6. Se me dan bien las Ciencias.
  7. Las Ciencias serán importantes para el trabajo de mi vida.
  8. Se me dan bien las mayoría de asignaturas, pero no tanto Ciencias.
  9. Con total seguridad, podría hacer tabajos de avanzados en Ciencias.

  Y las respuestas podían ser (del 0 al 5): *muy en desacuerdo, en desacuerdo, ni acuerdo ni en desacuerdo, de acuerdo, muy de acuerdo.*

```
dfjerarquicob2<-cbind(dfjerarquico[1],dfjerarquico[,10:18])
ggradar(dfjerarquicob2, background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2.5, 5),
        legend.title = "Grupos",
        group.point.size = 3,
        group.colours = lcols)
dfnojerarquicob2<-cbind(dfnojerarquico[1],dfnojerarquico
   [,10:18])
ggradar(dfnojerarquicob2,
        background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2.5, 5),
```

```
        legend.title = "Grupos",
        group.point.size = 2.5,
        group.colours = lcols)
```

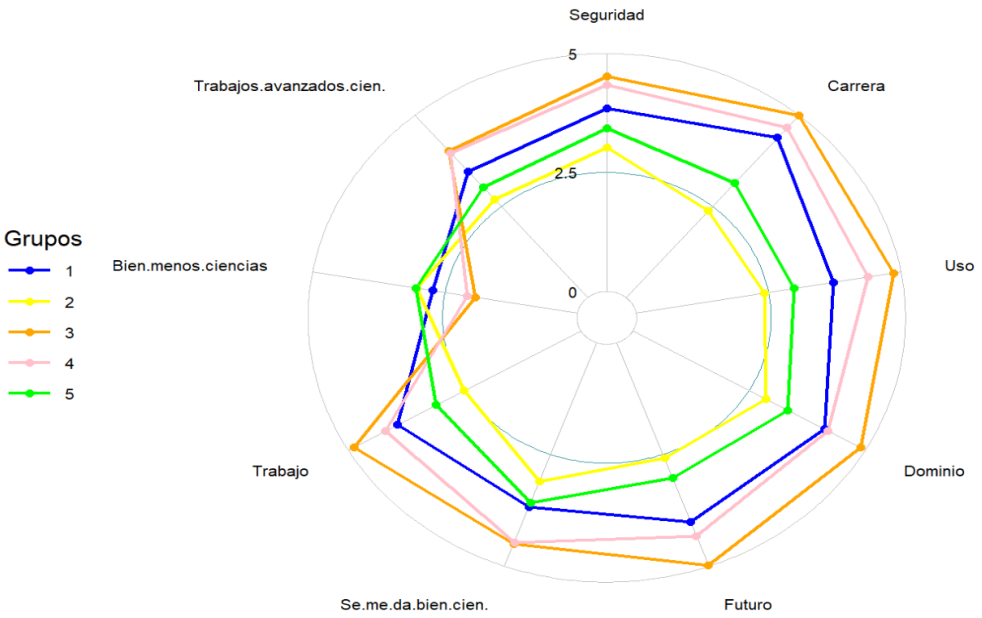


Figura 7.19.: Gráfico radar con método jerárquico bloque 2

A partir de la Figura 7.19 y la Figura 7.20, podemos describir cada grupo con el método jerárquico y no jerárquico respectivamente:

Grupo 1 **jerárquico** Este grupo responde con acuerdo alto a casi todas las preguntas, podríamos decir que muestra interés por las ciencias, y que se le dan bien.

Grupo 2 **jerárquico.** Este grupo responde a casi todo que no está ni en acuerdo ni en desacuerdo. Sin embargo, si comparamos con los demás es el que está más en desacuerdo con desenvolverse con seguridad en ciencias, elegir una carrera de ciencias, esperar usarlas en el futuro, que se les de bien y que pueda hacer trabajos avanzados en ciencias. A su vez, este grupo es uno de los que más de acuerdo está con que se le dan bien la mayoría de las asignaturas exceptuando las de ciencias. En resumen, este grupo, de entre todos, es al más difícil le parecen, peor se le dan y menos le gusta las ciencias.

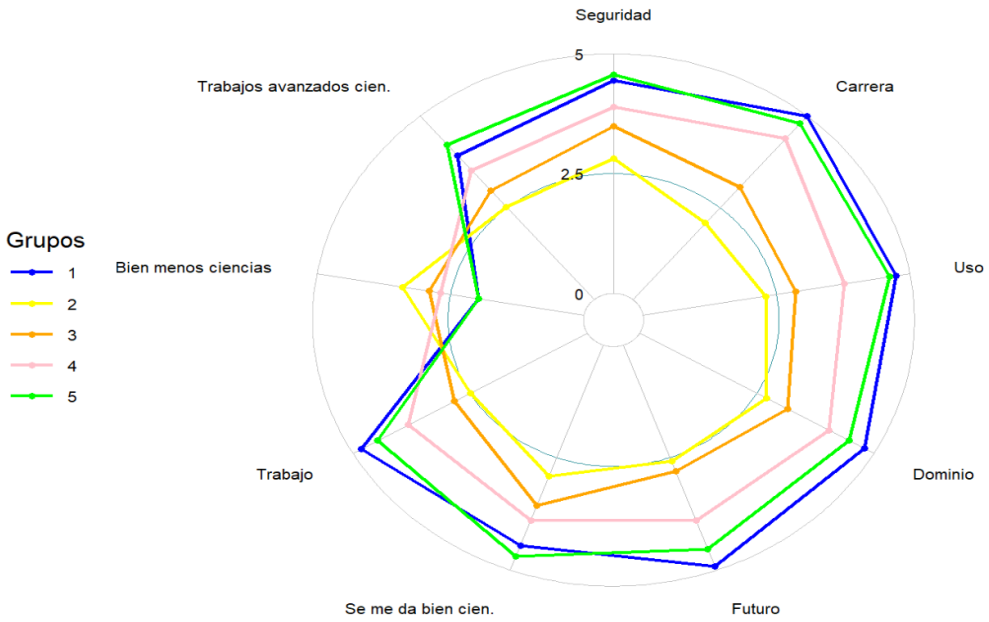


Figura 7.20.: Gráfico radar con método no jerárquico bloque 2

Grupo 3 **jerárquico.** En este bloque el grupo 3 muestra siempre el mayor interés. Además podemos destacar lo claro que tienen los alumnos de este grupo que necesitarán ciencias en su futuro. En resumen, podríamos decir que de este bloque, tiene gran interés y muestra que menos difíciles le parecen las ciencias.

Grupo 4 **jerárquico.** En este bloque se muestra mucha similitud entre el grupo 3 y el 4. Los estudiantes de ambos grupos tienen respuestas que casi siempre el el mayor interés. Sin embargo, el grupo 4 se posiciona en varias preguntas algo por debajo del 3. En resumen, podríamos decir que de este bloque, es uno de los que más interés muestra y menos difíciles le parece las ciencias.

Grupo 5 **jerárquico.** Este grupo después del grupo 2 es el que menos interés presenta en las ciencias, en comparación a los demás. Pero vemos que presenta acuerdo en todas las preguntas.

En conclusión, los grupos 3 y 4, son los que más interés tienen en las ciencias, y siguiéndolos está el grupo 1. El grupo 2, aunque al igual que el 5 (y, el 5, algo más que el 2), no muestran ni mucho ni poco interés en este bloque.

Grupo 1 **no jerárquico.** Este grupo es, junto al 5, el que más interés muestra por las ciencias. En resumen, podríamos decir que, de este bloque, es uno de los que más interés muestra y menos difíciles le parece las ciencias.

Grupo 2 **no jerárquico.** Este grupo responde a casi todo que no está ni en acuerdo ni en desacuerdo. Es muy similar al grupo 2 del jerárquico, exceptuando que muestra el mayor acuerdo en la pregunta 8.

Grupo 3 **no jerárquico.** Es similar al grupo 5 no jerárquico, que muestra un acuerdo, mayoritariamente bajo, con todas las preguntas. Podríamos decir que se acerca a ni en acuerdo ni en desacuerdo en casi todas las preguntas, menos en la 6, que indica que se le dan bien las ciencias.

Grupo 4 **no jerárquico.** Este grupo responde con acuerdo alto a casi todas las preguntas. Podemos ver que es similar al grupo 1 jerárquico. En conclusión diríamos que tienen interés por las ciencias, y que se le dan bien.

Grupo 5 **no jerárquico.** En este bloque este grupo muestra siempre el mayor interés. Además podemos destacar lo claro que tienen los alumnos de este grupo que necesitarán ciencias en su futuro. Es similar al grupo 3 jerárquico. En resumen, podríamos decir que de este bloque, son quienes tiene mayor interés y les parecen menos difíciles las ciencias.

En conclusión, los grupos 5 y 1, son los que más interés tienen en las ciencias, y siguiéndolos está el grupo 4. El grupo 2, aunque al igual que el 3 no muestran ni mucho ni poco interés en este bloque.

Resumiendo, en este bloque dos de los grupos muestran elevado interés en las ciencias, uno de ellos muestra respuestas de no estar ni de acuerdo ni en desacuerdo con las afirmaciones y dos grupos intermedios.

- Bloque 3: Este tercer bloque está compuesto por las siguientes preguntas:
  1. Me gusta imaginar la creación de nuevos productos.
  2. Si estudio ingeniería, podré mejorar cosas que la gente usa diario.
  3. Se me da bien construir y arreglar cosas.
  4. Me interesa saber cómo funcionan las máquinas.
  5. Diseñar productos o estructuras será importante en mi futuro trabajo.

6. Siento curiosidad por el funcionamiento de la electrónica.
7. Me gustaría poder aplicar creatividad e innovación en mi futuro trabajo.
8. Saber aplicar matemáticas y ciencias me permitirá inventar cosas útiles
9. Creo que puede irme bien en una carrera de ingeniería.

Y las respuestas podían ser (del 0 al 5): *muy en desacuerdo, en desacuerdo, ni acuerdo ni en desacuerdo, de acuerdo, muy de acuerdo.*

```
dfjerarquicob3<-cbind(dfjerarquico[1],dfjerarquico[,19:27])
ggradar(dfjerarquicob3, background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2.5, 5),
        legend.title = "Grupos",
        group.point.size = 3,
        group.colours = lcols)
dfnojerarquicob3<-cbind(dfnojerarquico[1],dfnojerarquico
   [,19:27])
ggradar(dfnojerarquicob3,
        background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2.5, 5),
        legend.title = "Grupos",
        group.point.size = 2.5,
        group.colours = lcols)
```

A partir de la Figura 7.21 y la Figura 7.22, podemos describir cada grupo:

Grupo 1 jerárquico. Este es el grupo que presenta más acuerdo en la ingeniería y tecnología, junto al grupo 4. Aunque, destacamos, que el acuerdo de este es el mayor de todos.

Grupo 2 jerárquico. Este muestra similitud con el 3 y el 5. Sin embargo, podemos destacar la pregunta 7 por mostrar más acuerdo entre los tres en que le gustaría aplicar su creatividad e innovación y en la pregunta 9, responde el acuerdo más bajo con respecto a creer que le puede ir bien en una ingeniería.

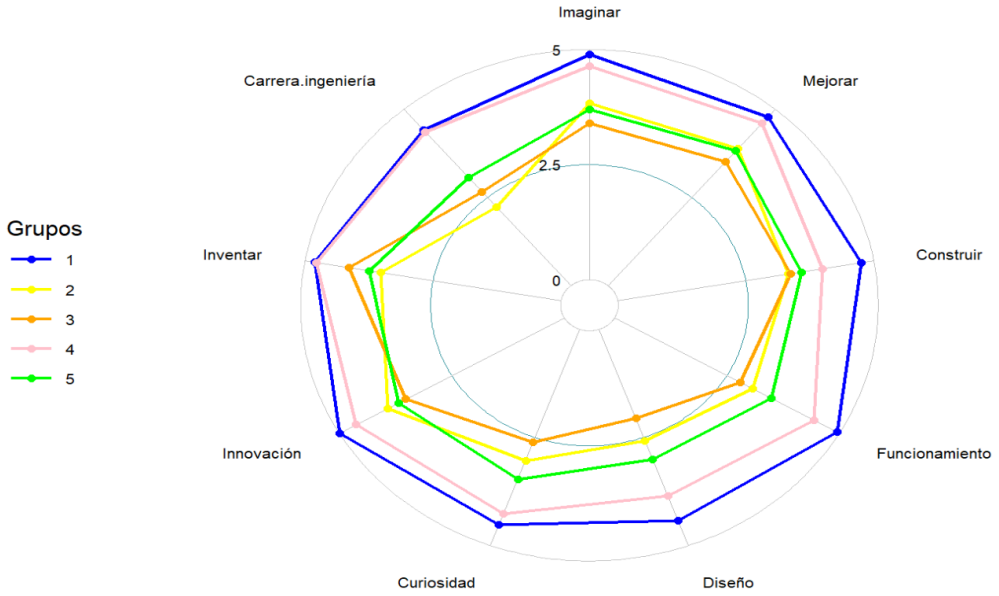


Figura 7.21.: Gráfico radar con método jerárquico bloque 3

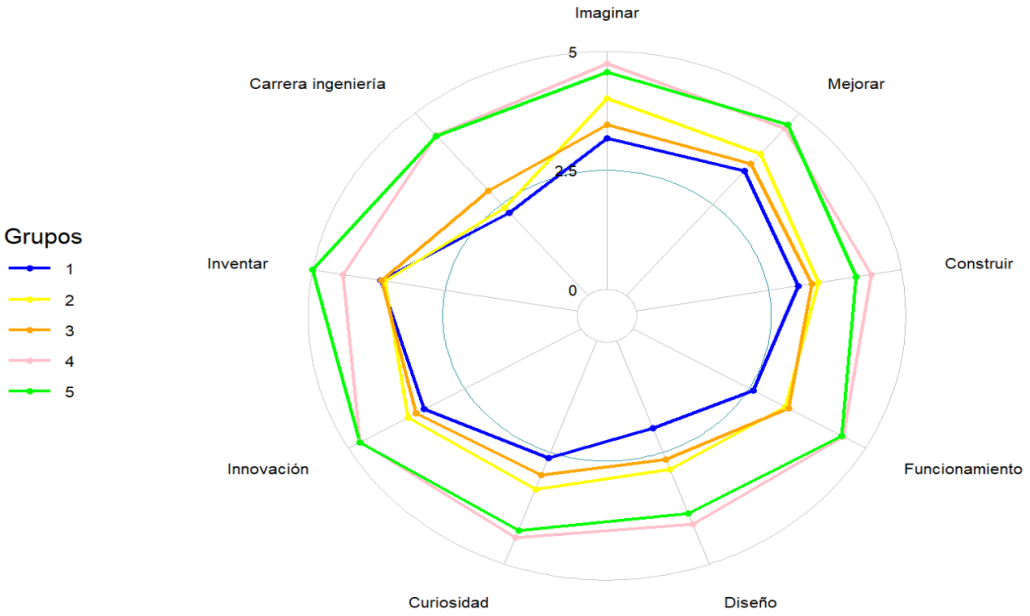


Figura 7.22.: Gráfico radar con método no jerárquico bloque 3

Grupo 3 **jerárquico.** Este grupo es bastante similar al anterior, aunque con respuestas ligeramente inferiores. Se puede señalar que solamente se muestra más de acuerdo frente al grupo 2 en las respuestas 8 y 9, referidas a que saber aplicar matemáticas y ciencias será útil para inventar cosas y que le podría ir bien en una carrera de ingeniería.

Grupo 4 **jerárquico.** Este grupo, vemos que, como hemos dicho ya, es bastante similar, al 1. A este grupo le parecen muy interesantes la ingeniería y la tecnología, aunque muestra respuestas ligeramente inferiores a las del grupo 1.

Grupo 5 **jerárquico.** Este grupo, similar al 2 y al 3, teniendo puntuaciones ligeramente mayores a ellos en la mayoría de preguntas.

En conclusión, los grupos 1 y 4, son los que más interés tienen en la ingeniería y tecnología. Y siguiéndolos están el grupo 2, 3 y el 5. En comparación, el 2 presenta menor interés.

Grupo 1 **no jerárquico.** Este grupo, similar al 2 y al 3, pero no destaca en preguntas sobre estos.

Grupo 2 **no jerárquico.** Este muestra similitud con el 1 y el 3, en este caso. Es muy similar al grupo 2 no jerárquico. Destaca en algunas preguntas.

Grupo 3 **no jerárquico.** Este grupo, vemos que, es bastante similar, al 1 y al 2. A este grupo le parecen muy interesantes la ingeniería y la tecnología. Se parece al grupo 3 en los jerárquicos.

Grupo 4 **no jerárquico.** Este es el grupo que presenta más acuerdo en la ingeniería y tecnología, junto al grupo 5. Es muy similar al grupo 4 jerárquico.

Grupo 5 **no jerárquico.** Este es el grupo que presenta más acuerdo en la ingeniería y tecnología, junto al grupo 4. Es muy similar al grupo 1 jerárquico.

En conclusión, los grupos no jerárquicos 4 y 5, son los que más interés tienen en la ingeniería y tecnología. Y siguiéndolos están el grupo 1, 2 y el 3. El 1, en comparación presenta menor interés.

De los cinco grupos, dos grupos muestran respuestas más elevadas y tres que puntúan con menor valor.

- Bloque 4: Este cuarto bloque está compuesto por las siguientes preguntas:
  1. Estoy convencido/a de que puedo guiar a otros personas para cumplir un objetivo.
  2. Estoy convencido/a de que puedo animar a otras personas para que den lo mejor de sí mismas.
  3. Estoy convencido/a de que puedo hacer un trabajo de alta calidad.
  4. Estoy convencido/a de que puedo respetar las diferencias de mis compañeros.
  5. Estoy convencido/a de que puedo ayudar a mis compañeros.
  6. Estoy convencido/a de que puedo incluir las perspectivas de los demás cuando tomo decisiones.
  7. Estoy convencido/a de que puedo hacer cambios cuando las cosas no van según lo planeado.
  8. Estoy convencido/a de que puedo establecerme mis propios objetivos de aprendizaje.
  9. Estoy convencido/a de que puedo gestionar mi tiempo con inteligencia al trabajar solo.
  10. Cuando tengo muchos trabajos que hacer, puedo elegir a cuáles dar prioridad.
  11. Estoy convencido/a de que puedo trabajar bien con estudiantes de orígenes distintos.

  Y las respuestas podían ser (del 0 al 5): *muy en desacuerdo, en desacuerdo, ni acuerdo ni en desacuerdo, de acuerdo, muy de acuerdo.*

```
dfjerarquicob4<-cbind(dfjerarquico[1],dfjerarquico[,28:38])
ggradar(dfjerarquicob4, background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2.5, 5),
        legend.title = "Grupos",
        group.point.size = 3,
        group.colours = lcols)
dfnojerarquicob4<-cbind(dfnojerarquico[1],dfnojerarquico
   [,28:38])
```

```
ggradar(dfnojerarquicob4,
        background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2.5, 5),
        legend.title = "Grupos",
        group.point.size = 2.5,
        group.colours = lcols)
```

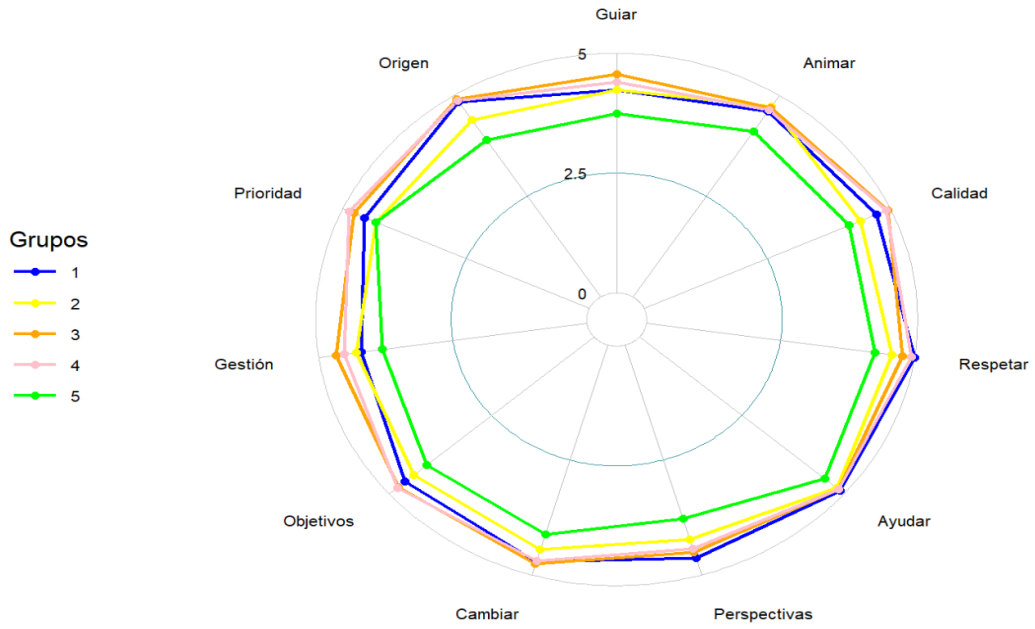


Figura 7.23.: Gráfico radar con jerárquico bloque 4

En la Figura 7.23 y Figura 7.24, podemos ver, que todos los grupos tienen respuestas bastante parecidas, de hecho, todos muestra acuerdo con las preguntas planteadas.

En general, podemos decir que todos los grupos presentan un grado de acuerdo elevado, con pequeñas diferencias entre ellos y, por tanto, que consideran que tienen las habilidades del siglo XXI que se les indican. Podemos ver que el grupo 3 es el que muestra mayor acuerdo, junto al 1 y al 4, por ello, que consideran de manera convincente tener esas habilidad, seguidos por el grupo 2. Y el grupo 5 son los que muestran

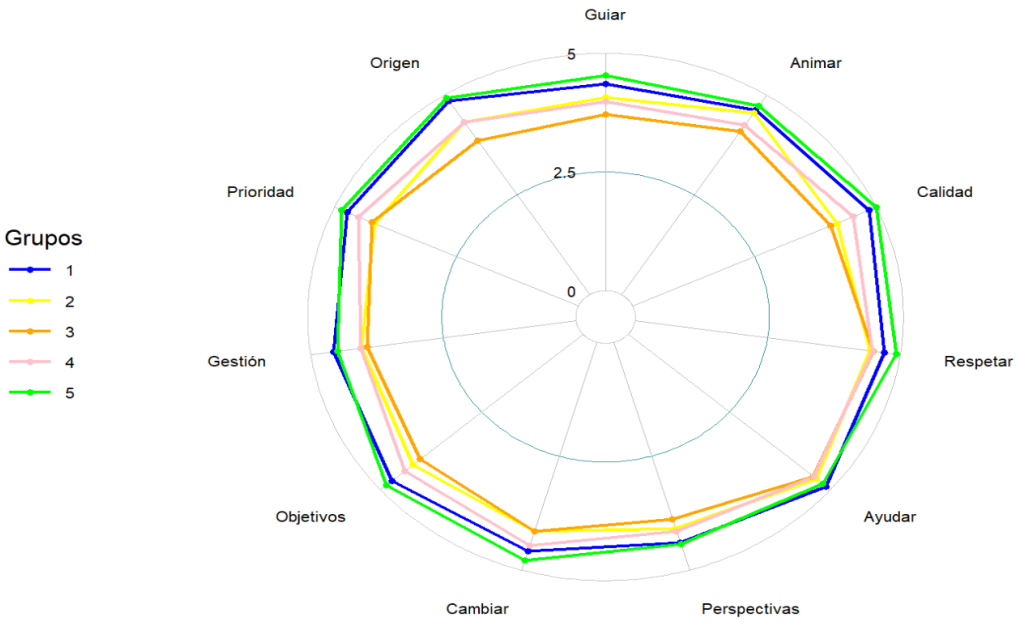


Figura 7.24.: Gráfico radar con no jerárquico bloque 4

menor grado de acuerdo, sin embargo, no es mucho menor, por ello, consideran que pueden tener las habilidades que se plantean.

Para los grupos que resultan de aplicar el método no jerárquico, podemos decir que, igual que con el método jerárquico, todos los grupos consideran que tienen las habilidades del siglo XXI que se les indican. El mayor acuerdo lo tienen el grupo 1 y 5, seguidos por los grupos 4 y 2. Y el 3 es, en comparación es que tiene un poco de menos acuerdo.

- Bloque 5: En el quinto bloque han de marcar el grado de interés por las siguientes materias:
  1. Física.
  2. Trabajo ambiental.
  3. Biología y Zoología.
  4. Trabajo veterinario.
  5. Matemáticas.
  6. Medicina.
  7. Geociencia.
  8. Informática.

9. Ciencias médicas.
10. Química.
11. Energía.
12. Ingeniería.

Y las respuestas podían ser (del 0 al 4): *no me interesa, no me interesa mucho, me interesa, me interesa mucho.*

```
dfjerarquicob5<-cbind(dfjerarquico[1],dfjerarquico[,39:50])
ggradar(dfjerarquicob5, background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2, 4),
        legend.title = "Grupos",
        group.point.size = 3,
        group.colours = lcols)
dfnojerarquicob5<-cbind(dfnojerarquico[1],dfnojerarquico
   [,39:50])
ggradar(dfnojerarquicob5,
        background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 2, 4),
        legend.title = "Grupos",
        group.point.size = 2.5,
        group.colours = lcols)
```

En la Figura 7.25 y la Figura 7.26, vemos mayor variedad entre los grupos, debido a que a cada uno tendrá interés en alguna materia concreta que pueden, o no, estar relacionadas con otras. Veamos qué destaca en cada grupo:

Grupo 1 **jerárquico.** Este grupo, presenta interés en muchas materias, podemos destacar que le gustan las matemáticas, la informática, la energía y sobre todo destacamos su interés por la ingeniería.

Grupo 2 **jerárquico.** Este grupo, no presenta mucho interés en nada. Pero tiene algo de interés en trabajo ambiental, biología y zoología, trabajo veterinario y medicina.

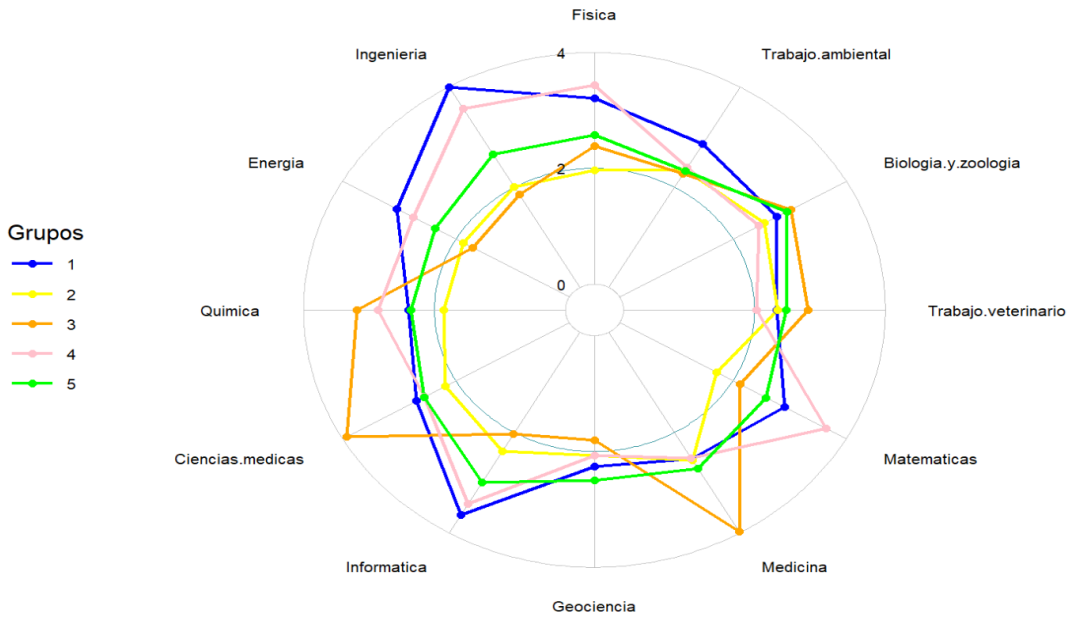


Figura 7.25.: Gráfico radar con jerárquico bloque 5

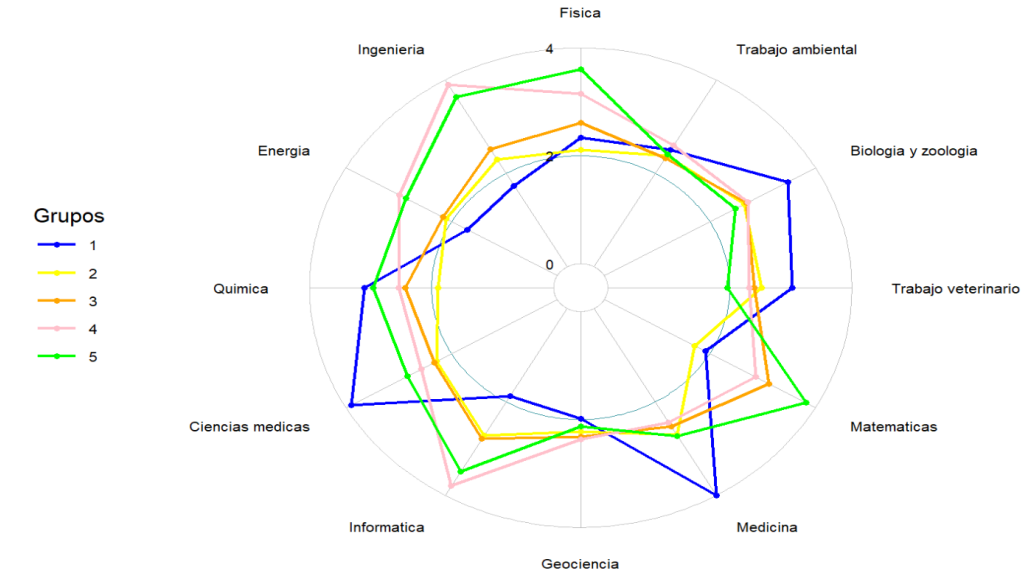


Figura 7.26.: Gráfico radar con no jerárquico bloque 5

Grupo 3 **jerárquico.** Este grupo destaca por su gran interés en medicina y ciencias médicas. Tiene, también, algo de interés en biología y

zoología, trabajo veterinario y química.

Grupo 4 **jerárquico.** Destaca por su interés en física, matemáticas informática e ingeniería, destacando por la puntuación en matemáticas y física.

Grupo 5 **jerárquico.** Este grupo tiene escaso interés en general. Podríamos destacar un poco más de valoración en biología y zoología o geociencia.

En conclusión podemos decir, que el grupo 2 presenta un grado de interés menor en todas las áreas científico-tecnológicas. El grupo 5, presenta un escaso grado de interés en todas las áreas. Los grupos 1,3 y 4 tienen mayor interés en general, aunque se diferencias en las materias. El grupo 3 destaca en su interés por: biología y zoología, trabajo veterinario, medicina, ciencias médicas y química. El grupo 4 se hace destacar especialmente, en su interés por: física, ingeniería y matemáticas. Y, el grupo 1, en su interés por: informática e ingeniería.

Grupo 1 **no jerárquico.** Este es similar al 4 del método jerárquico. Se diferencia en que tiene más interés en trabajo ambiental y en biología y zoología, y menos interés en matemáticas.

Grupo 2 **no jerárquico.** Este grupo, no presenta mucho interés en nada. Tiene un patrón parecido al grupo 2 con el método jerárquico, gustándole un poco más biología y zoología, trabajo veterinario y medicina. Y no interesándole las matemáticas.

Grupo 3 **no jerárquico.** Este grupo tiene un poco de interés en todo. Podríamos destacar en matemáticas.

Grupo 4 **no jerárquico.** Este grupo, presenta interés en muchas materias. Podemos destacar que con este método presenta mayor interés por la informática e ingeniería.

Grupo 5 **no jerárquico.** Destaca por su interés en matemáticas, informática y física.

En conclusión podemos decir, que el grupo 2 presenta un grado de interés menor, al igual que el grupo 3, en cierta medida, aunque este presenta gran interés en matemáticas. Los grupos 1, 4 y 5 tienen mayor interés en general. El grupo 1 destaca en su interés por: biología y zoología, trabajo veterinario, medicina, ciencias médicas y química. El grupo 4 se hace destacar especialmente, en su interés por: energía, ingeniería e informática. Y el grupo 5 destaca por su elevado interés por la física,

matemáticas, informática e ingeniería. El mayor interés es en la física y las matemáticas.

Se perfilan los grupos con diferente interés; uno en las ramas científicas (matemáticas y física), otro en ingeniería (ingeniería, energía e informática), otro grupo con interés en las áreas biosanitarias (medicina, ciencias médicas,...), un cuarto grupo con bajo interés en todas las ramas y un quinto con un poco más de interés en general, especialmente en matemáticas.

- Bloque 6: En el sexto bloque tenemos las siguientes preguntas:
  1. Cómo esperas que se te dé este año las clases de inglés/lengua y literatura
  2. Cómo esperas que se te dé este año las clases de matemáticas
  3. Cómo esperas que se te dé este año las clases de ciencias
  4. En un futuro, tienes planes de dar clases avanzadas en matemáticas
  5. Planeas ir a la universidad
  6. Conoces a adultos que trabajen como científicos
  7. Conoces a adultos que trabajen como ingenieros
  8. Conoces a adultos que trabajen como matemáticos
  9. Conoces a adultos que trabajen como tecnólogos

Las respuestas, de las tres primeras, podían ser (del 1 al 3): *no muy bien, bien, muy Bien*. Y de las siguientes (del 0 al 1): *sí, no, no estoy seguro/a.* Por ello, vamos a hacer un gráfico radar con cada método para las tres primeras y, otro, para las demás preguntas.

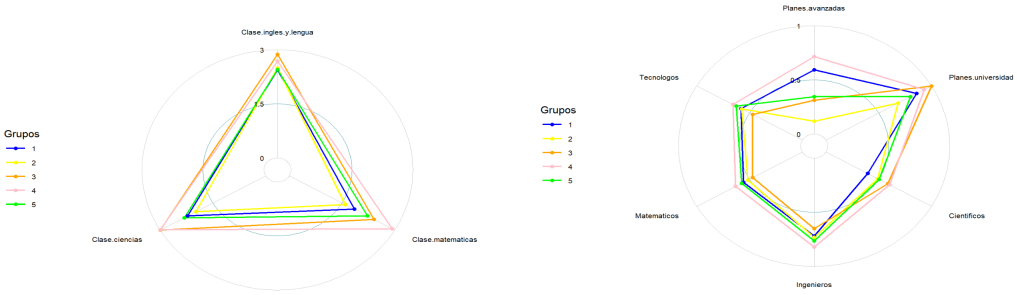
```
dfjerarquicob6<-cbind(dfjerarquico[1],dfjerarquico[,51:53])
dfjerarquicob61<-cbind(dfjerarquico[1],dfjerarquico[,54:59])
ggradar(dfjerarquicob6, background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 1.5, 3),
        legend.title = "Grupos",
        group.point.size = 3,
        group.colours = lcols)
ggradar(dfjerarquicob61, background.circle.colour = "white",
        gridline.min.linetype = 1,
```

```
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 0.5, 1),
        legend.title = "Grupos",
        group.point.size = 3,
        group.colours = lcols)
dfnojerarquicob6<-cbind(dfnojerarquico[1],dfnojerarquico
   [,51:53])
dfnojerarquicob61<-cbind(dfnojerarquico[1],dfnojerarquico
   [,54:59])
ggradar(dfnojerarquicob6,
        background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 1.5, 3),
        legend.title = "Grupos",
        group.point.size = 2.5,
        group.colours = lcols)
ggradar(dfnojerarquicob61,
        background.circle.colour = "white",
        gridline.min.linetype = 1,
        gridline.mid.linetype = 1,
        gridline.max.linetype = 1,
        values.radar = c(0, 0.5, 1),
        legend.title = "Grupos",
        group.point.size = 2.5,
        group.colours = lcols)
```

A partir de la Figura 7.27 y la Figura 7.28, podemos describir los grupos. En primer lugar, veremos las 7.27(a) y 7.28(a), para las preguntas del 1 al 3. Y posteriormente, las 7.27(b) y 7.28(b) con el resto de preguntas.

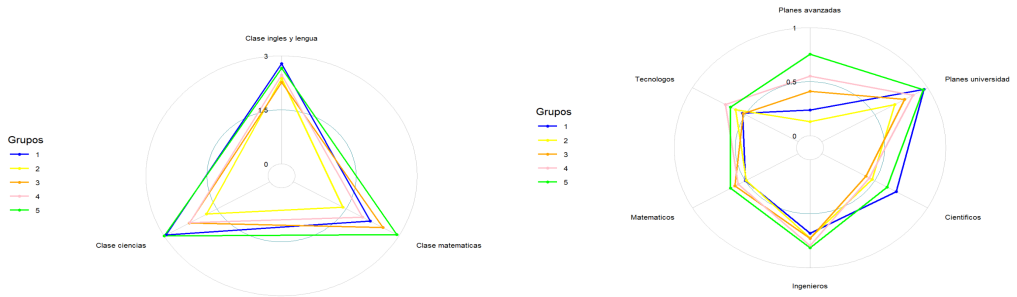
Grupo 1 jerárquico. Este grupo espera que le vaya mejor este curso en inglés/lengua y literatura y en ciencias que en matemáticas. Los escolares de este grupo planean dar clases avanzadas en matemáticas. Se puede destacar que es el que menos adultos científicos conoce.

Grupo 2 jerárquico. Este grupo, es muy parecido al grupo 1. También, espera que se le dé mejor inglés y lengua y literatura y las ciencias.



(a) Preguntas de la 1 a la 3 (b) Preguntas de la 4 a la 9

Figura 7.27.: Gráfico radar con jerárquico bloque 6



(a) Preguntas de la 1 a la 3 (b) Preguntas de la 4 a la 9

Figura 7.28.: Gráfico radar con no jerárquico bloque 6

Grupo 3 **jerárquico.** Este grupo considera que se le van a dar muy bien inglés y lengua y literatura y las ciencias, y que se le van a dar bien las matemáticas. Los escolares de este grupo son los que más planean ir a la universidad. Se puede destacar que es el que menos ingenieros, matemáticos y tecnólogos conoce, pero conoce a más científicos.

Grupo 4 **jerárquico.** Este grupo espera que se le den muy bien todas las materias. Además es el grupo que está más seguro de querer dar clases avanzadas en matemáticas. Se puede destacar que es el que más adultos conoce que trabajan en el ámbito científico-tecnológico.

Grupo 5 **jerárquico.** Este grupo espera que le den bien las tres materias.

En conclusión podemos decir, que el grupo 4 es el que mejor espera que

se le den las asignaturas este año, seguido del 3, que espera que se le den mejor inglés, lengua y literatura y las ciencias, y del 5, que espera que se le den mejor las matemáticas. El grupo 2 es el que peor espera que se le den las asignaturas, y el 1 tiene un pensamiento similar. El 4 también es el que muestra más de interés en dar clases avanzadas de matemáticas. Mencionar que todos tienen interés en ir a la universidad, aunque el que más interés tiene es el 3 y el que menos el 2. Todos los grupos, de manera general, muestran que conocen a algunos adultos que trabajen en las materias que hemos presentado, a los que más conocen en general, es a ingenieros. Podríamos destacar que el grupo 4, en comparación, es el que a más adultos trabajando en este ámbito conoce.

Grupo 1 **no jerárquico.** Este grupo espera que le den muy bien inglés, lengua y literatura y ciencias, y que se le den bien las matemáticas.

Grupo 2 **no jerárquico.** Este grupo, es el que peor espera que le vaya este curso. Este grupo es el que menos seguro está en querer dar clases avanzados en matemáticas.

Grupo 3 **no jerárquico.** Este grupo espera que se le den bien todas las materias.

Grupo 4 **no jerárquico.** Este grupo espera que se le den bien inglés, lengua y literatura y ciencias, y un poco peor las matemáticas. Este grupo considera que no le va a ir muy bien el curso.

Grupo 5 **no jerárquico.** Este grupo es el que espera que le vaya mejor este curso. Este grupo es el que más seguro está en querer dar clases avanzados en matemáticas.

En conclusión podemos decir, que el grupo 5 es el que mejor espera que las asignaturas este año, seguido del 1, que espera que se le den mejor inglés, lengua y literatura y las ciencias, y del 3, que espera que se le den mejor las matemáticas. El grupo 2 es el que peor espera que se le den las asignaturas, y el 4 tiene un pensamiento similar, aunque espera que se le den mejor inglés, lengua y literatura y ciencias. El 5 también es el que muestra más de interés en dar clases avanzadas de matemáticas. Mencionar que todos tienen interés en ir a la universidad, pero destacan el 1 y el 5. Todos los grupos muestran que lo que más conocen son ingenieros y que conocen a algunos adultos que trabajen como científicos, matemáticos o tecnólogos, pero no difieren demasiado. Podemos destacar que el grupo 4 es el que conoce a más adultos que trabajan en estos ámbitos.

En comparación, vemos que el grupo 2 es el que es similar en ambos casos. El grupo 1 jerárquico es similar al grupo 4 no jerárquico. El grupo 3 jerárquico es similar al 1 no jerárquico. El grupo 4 jerárquico es parecido al 5 no jerárquico. Y, por último, el grupo 5 jerárquico es similar al 3 no jerárquico.

A continuación, vamos a mostrar una tabla de doble entrada con el cruce de las soluciones obtenidas con el método jerárquico de Ward y el método no jerárquico de *k*-medias que hemos hecho en R. Haciendo uso del paquete `randomForest`:

```
library(randomForest)
cm <- table(clusterward,km$cluster)
cm
clusterward  1  2  3  4  5
          1  0  0  0 40  3
          2  7 46  3  4  0
          3 44  0  2  2  5
          4  5  0 11 17 79
          5  5 20 56 17  2
```

Tabla de cruce de métodos					
CLUSTER	1	2	3	4	5
1	0	0	0	40	3
2	7	46	3	4	0
3	44	0	2	2	5
4	5	0	11	17	79
5	5	20	56	17	2

Tabla 7.1.: Cruce del método jerárquico y no jerárquico

8. Conclusiones

En este trabajo se han estudiado y aplicado técnicas de análisis cluster a los datos de una encuesta a estudiantes de educación secundaria. Tras analizar los datos, se obtienen cinco grupos diferenciados en sus respuestas. Como el análisis se realiza con dos metodologías diferentes, se tienen dos formas de agrupamiento distintas.

Derivado del análisis de los datos, cabe destacar que tanto un análisis jerárquico como uno no jerárquico nos dan resultados coincidentes.

Se muestra, por un lado, en el bloque 1, un grupo que llamaremos A, que destaca por su gran interés y gusto por las matemáticas. Otro grupo, al cual llamaremos B, que sus respuestas indican que tiene poco aprecio hacia las matemáticas. Y otros tres grupos similares, C, D y E, a los cuales pertenecen escolares con respuestas tibias. Con respecto al bloque 2, se muestra un patrón parecido. Dos grupos, el A y el C, que tienen mayor gusto por las ciencias, y siguiéndole se encuentra el grupo E. El grupo B vuelve a mostrar poco interés en este bloque. El grupo D tiene una situación similar al anterior. Los grupos E y A tienen mayor gusto por la ingeniería. Los demás grupos no muestran ni mucho ni poco gusto en este bloque 3;así como también se podría destacar que el grupo B vuelve a ser el que tiene menos gusto en este bloque. El grupo C tampoco tiene mucho interés en la ingeniería. Todos los grupos consideran tener las habilidades del siglo XXI. Podemos destacar que los grupos A, C y E son los que consideran de manera más notable que las presentan. Con el bloque 5 podríamos concluir que el grupo B presenta poco interés en las materias del ámbito científico-tecnológico. El grupo D tiene un poco de interés en todo. El grupo C tiene interés por: biología y zoología, trabajo veterinario, medicina, ciencias médicas y química. El grupo A se hace destacar especialmente, en su interés por: física, ingeniería y matemáticas. Y, el grupo E en su interés por: informática e ingeniería.

Con respecto a este año, el grupo A es el que espera que se le den mejor las asignaturas. Seguidamente, el C espera que se le den mejor inglés, lengua y literatura y ciencias. Le sigue el grupo D que espera que se le den mejor las matemáticas. El grupo B es el que espera los peores resultados en las asignaturas. El grupo E tienen un pensamiento similar. Cabe mencionar, que todos tienen interés en ir a la universidad, aunque el que más interés tiene es

el C y el que menos el B. Todos los grupos, de manera general, muestran que conocen a algunos adultos que trabajen en las materias que hemos presentado, siendo la mayoría de ellos ingenieros. Podríamos destacar que el grupo A, en comparación, es el que a más adultos trabajando en este ámbito conoce.

En conclusión, podemos decir que, tenemos un grupo B que claramente manifiesta un interés muy bajo y puntúa con valores en general por debajo de la media; otro grupo C cuya característica más destacado es que muestra puntuaciones muy altas en un grupo de variables que conforman el bloque segundo del cuestionario (es decir, en ciencias), y puntuaciones menores en las preguntas de ingeniería. Tenemos otro grupo de escolares que presenta un patrón semejante al anterior, que muestra puntuaciones altas en las preguntas de ingeniería y menos altas en las preguntas de ciencias: este es el grupo E, que como vemos tienen un mayor interés por la ingeniería que por las ciencias. Un cuarto grupo muestra un alto interés en todos los ámbitos por los que se les pregunta, que es el grupo A. Y finalmente, un quinto grupo, que llamamos D, el cual contesta mostrando un interés tibio en las áreas de ciencias y tecnología y cuyos escolares tienen menor confianza en sus aptitudes.

9. Conclusions

In this project, techniques of the cluster analysis have been studied and applied to the data obtained of a questionnaire carried out by students of Secondary Education. After analysing the data, five groups distinguishable by their responses are obtained. As the analysis is carried out by means of two different methodologies, two different ways of grouping are possible.

As a product of the data analysis, it should be pointed out that a hierarchical analysis and a non-hierarchical one produce coinciding results.

It is shown, on one side, in section 1, a group that will be called A that stands out for their great interest and pleasure in mathematics. Another group, which we will call B, shows the lack appreciation towards mathematics by means of their responses. Other three similar groups C, D and E to which scholars with lukewarm responses belong. Respecting section 2, a similar pattern is shown. Two groups, A and C that prefer science, and that are followed by E. Group B shows low interest in this section again. Group D has a similar situation to the previous one. Groups E and A have more pleasure for engineering. The rest of the groups do not show high or low preference in this section 3; as well as it could be highlighted that group B's interest decreases. Group C isn't interested that much in engineering neither. All the groups consider that they have the XXI's century abilities. We can highlight that groups A, C and E highly consider that they present them significantly. We could conclude, with the help of section 5, that group B presents not much interest in the scientific-technological subjects. Group D shows some interest in all of them. Group C is interested in: Biology, Zoology, Veterinary Medicine, Medicine, Medical Science and Chemistry. Group A stands out in a special way for: physics, engineering and mathematics. And group E highlights for its interest for: computer science and engineering.

Regarding this year, group A is the one that expect the best results results on subjects. In the say way, group C expects the best results in English, Language and Literature and Science; as well as group D hope they do better with mathematics. Group B expects the worst results in their subjects. Group E has similar thoughts.

It is important to mention that all groups are interested in going to University, although the most interested group is C and the least one is B.

All groups show, in general, that they know some adult persons that work in the subjects that have been presented, being engineers most of them.

We could highlight that group A is, in contrast, the one that knows the higher number of adult workers in this field.

In conclusion, we have a group B that clearly shows a very low interest and punctuates with under the average values in general; another group C whose most relevant characteristic is that it shows very high punctuation in a group of variables that shape the questionnaire's second section, (that is, in science) and lower punctuation in engineering questions. We have another group of scholars that present a very similar pattern to the previous one, that shows high punctuation in engineering questions and less high in science ones: this is group E, that has a greater interest in engineering rather than in science. A forth group shows an elevated interest in all the fields they are asked for, that is group A. And finally, a fifth group, which we call D, the one that answers showing a lukewarm interest in science and technological fields and whose students trust less in their aptitudes.

Bibliografía

[1] Joaquín Aldás Manzano and Ezequiel Uriel Jiménez. *Análisis multivariante aplicado con R*. Ediciones Paraninfo, SA, 2017.

[2] MM Astrahan. *Speech analysis by clustering, or the hyperphoneme method*. Standford University, 1970.

[3] Klaus Backhaus, Bernd Erichson, Sonja Gensler, Rolf Weiber, and Thomas Weiber. *Multivariate analysis*. Springer, 2021.

[4] Geoffrey H Ball, David J Hall, et al. *ISODATA, a novel method of data analysis and pattern classification*, volume 699616. Stanford research institute Menlo Park, CA, 1965.

[5] Mónica Balzarini, Cecilia Bruno, Mariano Córdoba, and Ingrid Teich. Herramientas en el análisis estadístico multivariado. *Escuela Virtual Internacional CAVILA. Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba. Córdoba, Argentina*, 2015.

[6] Michael J Crawley. *The R book*. John Wiley & Sons, 2012.

[7] Carlos María Cuadras. *Nuevos métodos de análisis multivariante*, volume 20. Barcelona: CMC editions, 2007.

[8] Resolución de la Dirección General de Formación del Profesorado e Innovación Educativa. `https://www.juntadeandalucia.es/boja/2021/148/19`. 22 de julio de 2021.

[9] Merriam-Webster Dictionary. Merriam-webster. 1831.

[10] Edward W Forgy. *Cluster analysis of multivariate data: efficiency versus interpretability of classifications*. 1965.

[11] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[12] Ramón Gutiérrez, Andrés González, Francisco Torres, and José A Gallardo. *Técnicas de Análisis de datos multivariable. Tratamiento computacional*. 1994.

[13] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[14] Alboukadel Kassambara. *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda, 2017.

[15] Maurice G. Kendall. *Multivariante Analysis, p.1*. Charles Griffin b Co. LTD. London, 1975.

[16] Carlos Lozares Colina and Pedro López-Roldán. El análisis multivariado. definición, criterios y clasificación. *Papers Sociología*, 1991.

[17] César López Pérez. *Técnicas de Ánalisis Multivariante de datos.* Prentice Hall, 2004.

[18] James MacQueen et al. *Some methods for classification and analysis of multivariate observations*, volume 1. 1967.

[19] Glenn W Milligan and Martha C Cooper. A study of standardization of variables in cluster analysis. *Journal of classification*, 5:181–204, 1988.

[20] Daniel Peña. *Análisis de datos multivariantes*. McGraw-Hill, 2002.

[21] R-Bloggers. `http://r-project.org/`.

[22] R-Bloggers. `www.r-bloggers.com/2021/04/cluster-analysis-in-r/`.

[23] R-Books. `https://myrbooksp.netlify.app/`.

[24] R-Charts. `https://r-charts.com/es/ranking/ggradar/`.

[25] R-Pubs. `https://rpubs.com`.

[26] RDocumentation. `https://www.rdocumentation.org`.

[27] Marta Salas. Planificación, recogida de datos, análisis e informe para el estudio del desarrollo de proyectos stem en los centros educativos de granada, 2023.

[28] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[29] John Wilder Tukey et al. *Exploratory data analysis*, volume 2. Springer, 1977.

[30] UNESCO. Declaration race and racial prejudice. `https://www.unesco.org/en/legal-affairs/declaration-race-and-racial-prejudice`.

[31] Ruth Vilà Baños, María José Rubio Hurtado, Vanessa Berlanga, and Mercè Torrado Fonseca. Cómo aplicar un cluster jerárquico en spss. *REIRE. Revista d'Innovació i Recerca en Educació, 2014, vol. 7, num. 1, p. 113-127*, 2014.