COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales
ICADE

Segmentación de estudiantes universitarios mediante *clustering* para la planificación académica

Autor: Carlos Martín de los Santos Antoñanzas
Director: Lourdes Fernández Rodríguez

MADRID | Diciembre 2025

Resumen:

Este Trabajo de Fin de Grado surge del interés por comprender la diversidad real del alumnado que accede a la universidad y de explorar cómo el análisis de datos puede ayudar a mejorar la toma de decisiones académicas.

El objetivo principal del estudio es identificar grupos homogéneos de estudiantes a partir de información académica y sociodemográfica, con el fin de proponer estrategias de gestión y planificación basadas en evidencias. La metodología integra una revisión de la literatura con un enfoque cuantitativo aplicado al *dataset* 'UCM-Acceso'. Tras un análisis exploratorio exhaustivo, se aplicó el algoritmo *k-prototypes* como técnica principal de *clustering* para datos mixtos, complementado de forma exploratoria con *k-medoids* y distancia de Gower. La evaluación mediante *silhouette score* permitió validar la coherencia de la segmentación obtenida.

El análisis reveló la existencia de dos perfiles claramente diferenciados de estudiantes, cuyas características permiten diseñar estrategias orientadas tanto al apoyo académico como a la planificación institucional. Asimismo, el estudio identificó tendencias estructurales, como la evolución de la demanda por ramas de estudio, los cambios en el nivel socioeconómico del alumnado o el aumento de las notas de acceso, que aportan información valiosa para la gestión de recursos universitarios.

En conjunto, el trabajo pone de manifiesto la utilidad del análisis de datos propio del ámbito de *Business Analytics* para comprender la heterogeneidad del alumnado y para fundamentar decisiones académicas en datos reales, abriendo la puerta a enfoques más personalizados y a líneas futuras de investigación en predicción y análisis educativo.

Palabras clave: *Business Analytics*, *clustering*, *k-prototypes*, análisis, segmentación, datos, universidad, planificación académica

Abstract:

This Final Degree Project stems from an interest in understanding the real diversity of students entering university and exploring how data analysis can support academic decision-making.

The main objective of the study is to identify homogeneous groups of students based on academic and sociodemographic information, in order to propose evidence-based strategies for management and planning. The methodology combines a literature review with a quantitative approach applied to the 'UCM-Acceso' dataset. Following an exhaustive exploratory analysis, the k-prototypes algorithm was implemented as the primary clustering technique for mixed data, complemented by an exploratory application of k-medoids with Gower distance. The segmentation obtained was validated through the silhouette score.

The analysis revealed two clearly differentiated student profiles, whose characteristics support the design of strategies aimed at both academic support and institutional planning. In addition, the study identified several structural trends, such as the evolution of demand across fields of study, changes in students' socioeconomic level, and the rise in admission grades, which provide valuable insights for the allocation of university resources.

Overall, the results highlight the usefulness of data analysis within the field of Business Analytics as a tool to understand student heterogeneity and to inform academic decisions using real data. The findings open the door to more personalized management approaches and future research in educational prediction and analytics.

Keywords: Business Analytics, clustering, k-prototypes, analysis, segmentation, data, university, academic planning

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Objetivos del trabajo

El objetivo general de este trabajo es analizar el perfil del alumnado que accede a la Universidad Complutense de Madrid (UCM) mediante técnicas de análisis de datos y segmentación, con el fin de apoyar la toma de decisiones en materia de planificación académica y gestión universitaria.

A partir de este objetivo general, se plantean dos objetivos específicos:

1. Identificar perfiles diferenciados de estudiantes mediante técnicas de *clustering*, con el fin de proponer estrategias académicas y de orientación adaptadas a las características específicas de cada grupo.
2. Analizar las implicaciones del comportamiento histórico de la demanda y las tendencias de acceso universitario para la asignación eficiente de recursos, a partir del análisis exploratorio del *dataset*, sin depender de los perfiles generados por el *clustering*.

De este modo, el trabajo integra dos aproximaciones complementarias: por un lado, el análisis exploratorio del acceso universitario para detectar tendencias relevantes para la planificación de recursos; y por otro, la segmentación del alumnado para proponer estrategias orientadas a mejorar la experiencia académica y la organización docente.

Aunque este trabajo se basa en datos de la UCM, el procedimiento analítico podría aplicarse en otras instituciones siempre que las bases de datos (*datasets*) presenten la misma definición, codificación y estructura de variables. En este sentido, lo extrapolable es la metodología empleada, mientras que los perfiles obtenidos no serían directamente transferibles, ya que dependen de las características específicas del alumnado de cada institución.

1.2. Justificación e interés del tema

La elección de este tema está muy ligada a mi formación en el Doble Grado en ADE y

Business Analytics en ICADE, pero también a mi propia experiencia como estudiante universitario. A lo largo de estos años he visto de primera mano la variedad de perfiles que llegan a la universidad: estudiantes de fuera de la Comunidad de Madrid, trayectorias educativas muy distintas y situaciones familiares que influyen en el recorrido académico de cada persona. Entiendo bien las dificultades que aparecen al comenzar esta etapa y cómo esas diferencias pueden afectar a la adaptación inicial y al rendimiento. Esto me llevó a querer estudiar de manera más rigurosa cómo es realmente el alumnado que accede a la UCM y qué patrones se pueden identificar en sus características.

Además, mi formación en análisis de datos ha sido una motivación clave a la hora de escoger este tema. Durante la carrera he trabajado con técnicas de segmentación y análisis basado en datos, y me parecía una oportunidad perfecta aplicar estas herramientas a un entorno que vivo a diario y que considero especialmente relevante desde una perspectiva social y educativa. Poder combinar mi interés por el análisis cuantitativo con un tema tan cercano como el acceso universitario hace que este Trabajo de Fin de Grado tenga para mí un valor especial. En definitiva, es un tema que encaja con lo que estudio, con lo que me interesa y con lo que he vivido como estudiante.

1.3. Metodología

El trabajo combina una revisión cualitativa de la literatura con un enfoque cuantitativo aplicado al *dataset* 'UCM-Acceso' (UniversiDATA, 2025), procedente del portal UniversiDATA, que reúne más de 100.000 registros correspondientes a estudiantes admitidos entre los cursos académicos 2017/18 y 2024/25.

En primer lugar, se realizó una revisión de la literatura sobre analítica de datos en el ámbito educativo y sobre técnicas de aprendizaje no supervisado, con el objetivo de establecer el marco conceptual que sustenta el análisis. A continuación, se llevó a cabo un análisis exploratorio de datos (EDA, por sus siglas en inglés, *Exploratory Data Analysis*) para comprender la estructura del *dataset*, detectar patrones generales y preparar las variables mediante procesos de depuración, transformación y selección.

Posteriormente, se aplicaron técnicas de *clustering* mixto con el fin de identificar perfiles homogéneos de estudiantes. Para ello, se utilizaron dos algoritmos complementarios:

- *k-prototypes*, adecuado para datos mixtos al combinar medias para variables numéricas y modas para variables categóricas.
- *k-medoids*, que permite trabajar directamente sobre una matriz de disimilitud, en este caso, la distancia de Gower, proporcionando una alternativa más robusta frente a valores atípicos.

La calidad de las agrupaciones se evaluó mediante el *silhouette score* (también denominado *Average Silhouette Width*), seleccionando la configuración de variables y parámetros que ofrece una mayor cohesión interna y una mejor separación entre clústeres.

Finalmente, los resultados se analizaron e interpretaron desde una doble perspectiva: por un lado, las implicaciones de los clústeres para proponer estrategias académicas adaptadas a cada perfil; y por otro, las tendencias generales derivadas del EDA que permiten extraer conclusiones útiles para la planificación de recursos universitarios. Todo el proceso analítico se desarrolló en Python, ejecutado en Google Colab, lo que facilitó un flujo de trabajo reproducible y accesible desde la nube.

1.4. Datos utilizados

Los datos empleados en este trabajo proceden de UniversiDATA, el portal colaborativo de datos abiertos impulsado por diversas universidades españolas. En particular, se utilizó el *dataset* 'UCM-Acceso', publicado por la UCM, que recopila información de los estudiantes de nuevo ingreso entre los cursos 2017/18 y 2024/25.

Este *dataset*, anonimizado y de acceso público, incluye variables académicas, demográficas y socioeconómicas que permiten analizar los perfiles del alumnado en el momento de su acceso a la universidad. Su elección se justifica por la amplitud temporal del periodo cubierto, la calidad y homogeneidad de los datos, y la representatividad de la UCM como institución pública con una amplia y variada oferta de estudios y un notable reconocimiento académico.

En los apartados posteriores se describe con mayor detalle la estructura del *dataset* y el proceso de preparación de los datos para su análisis, incluyendo los procedimientos de depuración, transformación y selección de variables.

2. MARCO TEÓRICO

2.1. *Educational Data Mining* y *Learning Analytics*

En el ámbito educativo, las tecnologías de aprendizaje automático (*machine learning*), minería de datos (*data mining*) y el análisis de datos aplicados al aprendizaje han impulsado el surgimiento de dos comunidades científicas principales: *Learning Analytics* (LA) y *Educational Data Mining* (EDM). Ambas comparten objetivos y técnicas, apoyándose en disciplinas como la minería de datos, la psicometría y la modelización estadística, aunque difieren en su enfoque. En el caso de LA, se centra en mayor medida en los sistemas de gestión del aprendizaje y en el análisis de resultados a gran escala, mientras que EDM proviene de la comunidad de sistemas de tutoría inteligente (*intelligent tutoring systems),* que emplean inteligencia artificial para ofrecer una orientación personalizada, adaptando en tiempo real las actividades y la retroalimentación al progreso y necesidades del estudiante, replicando parcialmente la interacción uno a uno entre profesor y alumno. EDM, por tanto, trabaja con datos de cognición a pequeña escala. (Chen et al., 2020).

En este sentido, EDM desarrolla métodos y aplica técnicas derivadas de la estadística, el *machine learning* y el *data mining* para analizar información recogida durante la enseñanza y el aprendizaje, con el objetivo de contrastar teorías educativas y generar evidencias que orienten y mejoren la práctica docente. Por su parte, *Learning Analytics* aplica técnicas procedentes de disciplinas más amplias como la informática, la sociología, la psicología y las ciencias de la información, para analizar datos obtenidos en la administración educativa, los servicios, la enseñanza y el aprendizaje, generando aplicaciones que influyen directamente en la práctica educativa (Bienkowski et al., 2012).

Una de las definiciones más citadas en la literatura, propuesta por Siemens & Baker (2012), concibe el *Learning Analytics* como el uso de datos inteligentes producidos por los propios estudiantes, junto con modelos de análisis, para descubrir información y conexiones sociales, así como para predecir y orientar el aprendizaje. Posteriormente, en la primera *Learning Analytics and Knowledge Conference* (LAK 2011) se adoptó una definición más formal, entendiéndolo como la medición, recopilación, análisis e informe de datos sobre los alumnos y sus contextos, con el propósito de comprender y optimizar

tanto el aprendizaje como los entornos en los que este ocurre (Lang et al., 2022).

En contraste, EDM se consolidó a mediados de la década de 2000 como una disciplina orientada al desarrollo de métodos computacionales para identificar patrones en datos educativos. Se caracteriza por el análisis a nivel micro de las interacciones de los estudiantes con los sistemas de aprendizaje (como clics, tiempos de respuesta o secuencias de resolución de problemas) con el objetivo de modelizar procesos de aprendizaje y ofrecer retroalimentación personalizada (Baker & Yacef, 2009).

Siemens & Baker (2012) señalan cinco diferencias clave entre *Educational Data Mining* (EDM) y *Learning Analytics* (LA):

- **Descubrimiento vs. juicio humano**: EDM se orienta al descubrimiento automatizado de patrones, mientras que LA pone énfasis en el factor humano para la interpretación de los datos y la toma de decisiones.
- **Reducción vs. holismo**: EDM aborda los sistemas educativos de manera reduccionista, dividiéndolos en componentes específicos, mientras que LA adopta una visión holística que busca comprender el sistema de aprendizaje en su conjunto.
- **Orígenes distintos**: EDM procede del *software* educativo y de los *intelligent tutoring systems,* mientras que LA surge de ámbitos como la web semántica, la predicción de resultados o la gestión institucional.
- **Adaptación vs. empoderamiento**: EDM se centra en la adaptación automática de los sistemas al estudiante, mientras que LA busca empoderar a docentes y alumnos en la toma de decisiones educativas.
- **Técnicas y modelos**: EDM recurre con mayor frecuencia a métodos como la clasificación, el análisis de conglomerados (*clustering*), el modelo bayesiano o la minería de relaciones (*relationship mining*), mientras que LA se apoya más en el análisis de redes sociales, el análisis de sentimiento (*sentiment analysis*), el análisis del discurso (*discourse analysis*) y la predicción del rendimiento académico.

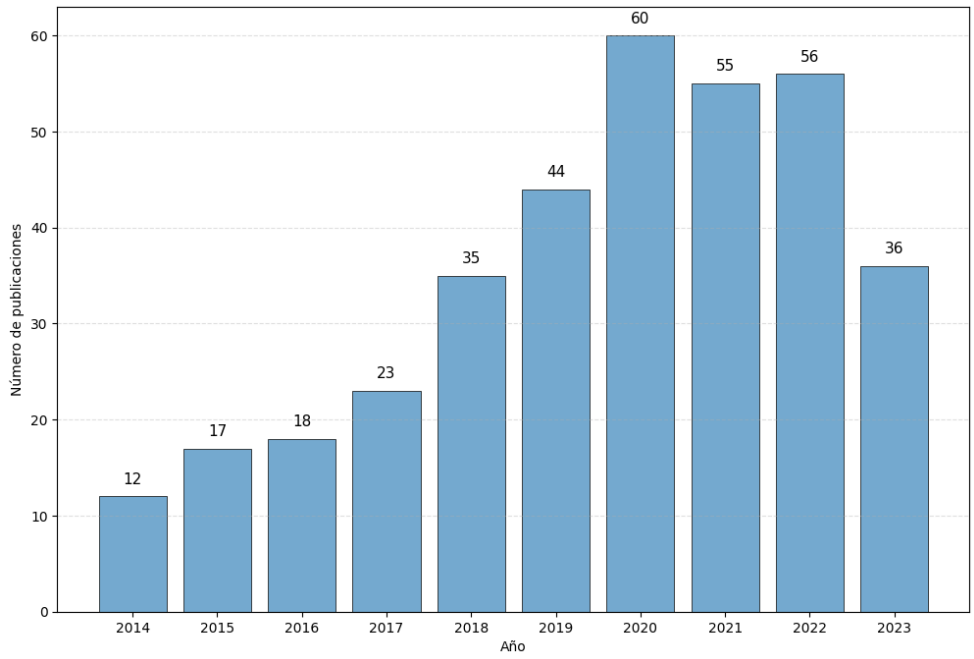
A pesar de estas diferencias, EDM y LA comparten el mismo objetivo: mejorar la calidad de la educación mediante el análisis de grandes volúmenes de datos para extraer

información útil para los distintos actores del proceso educativo. La estadística avanzada, el *machine learning* y el *data mining* son ejemplos de técnicas que ya han sido adoptadas en sectores como la industria, las finanzas o la sanidad para optimizar decisiones a partir de datos históricos, y cuya incorporación en el ámbito educativo refleja una tendencia similar (Calvet Liñán & Juan Pérez, 2015).

Con el fin de actualizar la evolución de la producción científica en los ámbitos de *Learning Analytics* y *Educational Data Mining*, se han considerado los resultados del estudio bibliométrico de Kurday & Vladova (2025), que analiza 356 publicaciones en el periodo 2014-2023. Este análisis resulta especialmente relevante, ya que refleja la consolidación de ambas disciplinas y su creciente presencia en la investigación educativa reciente.

Como se aprecia en la Figura 1, el número de publicaciones académicas sobre aplicaciones de LA y EDM creció de manera sostenida entre 2014 y 2020, año en el que alcanzó un máximo de 60 artículos. A partir de entonces, el volumen se ha mantenido en niveles elevados, con un ligero descenso en 2023, aunque esta reducción se explica probablemente por la fecha de cierre de la recogida de datos (2 de diciembre de 2023), por lo que la cifra final podría ser superior.

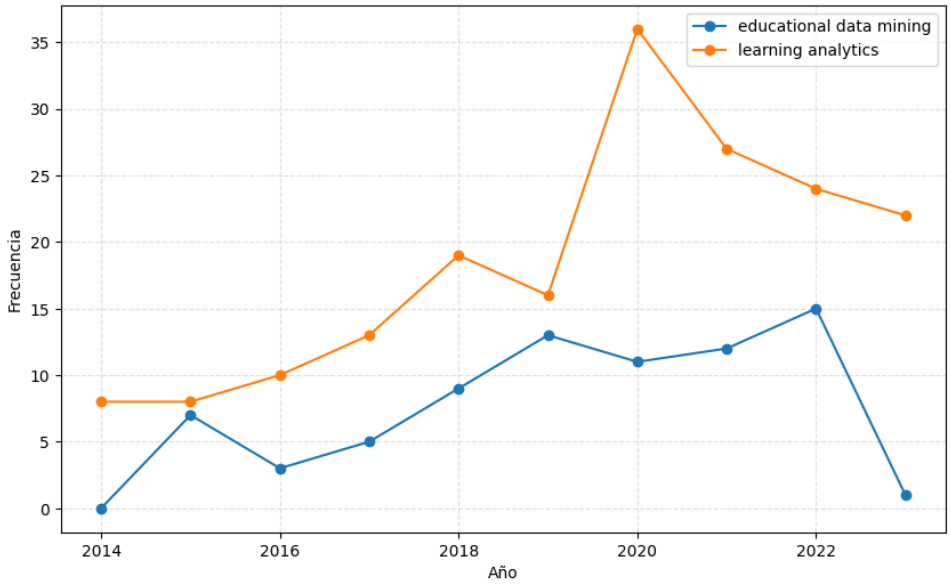
Figura 1: Evolución del número de publicaciones sobre aplicaciones de LA y EDM



Fuente: Elaboración propia a partir de Kurday & Vladova (2025)

Por su parte, la Figura 2 muestra la frecuencia con la que los términos *Learning Analytics* y *Educational Data Mining* aparecen en los artículos revisados. Los resultados indican una mayor presencia del concepto *Learning Analytics*, lo que refleja su papel predominante en los últimos años dentro de la literatura especializada, mientras que EDM mantiene una representación más acotada, aunque significativa.

Figura 2: Frecuencia de los términos *Educational Data Mining* y *Learning Analytics* en publicaciones



Fuente: Elaboración propia a partir de Kurday & Vladova (2025)

En conjunto, estos resultados confirman la madurez alcanzada por estas disciplinas y su creciente relevancia en la investigación educativa, reforzando la necesidad de incorporar sus enfoques en los procesos de análisis y planificación académica.

En resumen, *Educational Data Mining* y *Learning Analytics* han evolucionado como campos interrelacionados que, pese a sus diferencias de origen y enfoque, hoy se entienden como complementarios. EDM aporta el componente metodológico para descubrir patrones y generar modelos, mientras que LA ofrece la perspectiva aplicada para traducir esos hallazgos en decisiones pedagógicas e institucionales. Esta complementariedad resulta clave para el presente trabajo, que combina técnicas de minería de datos con una orientación hacia la mejora de la planificación académica en la universidad.

2.2. Analítica de datos en el sector educativo

En los últimos años, especialmente en la educación superior, el sector educativo ha evolucionado hacia una cultura basada en datos (*data-driven*) para la toma de decisiones. Las instituciones educativas generan una amplia variedad de datos a través de sus sistemas de información, como resultados de exámenes, tasas de asistencia, datos sociodemográficos obtenidos de la matrícula o registros de uso de plataformas virtuales. Estos datos, correctamente analizados, constituyen recursos valiosos para mejorar los procesos educativos y de gestión. Sin embargo, tradicionalmente gran parte de esta información quedaba almacenada sin un aprovechamiento real desperdiciando un gran potencial de mejora (Aristizabal F., 2016).

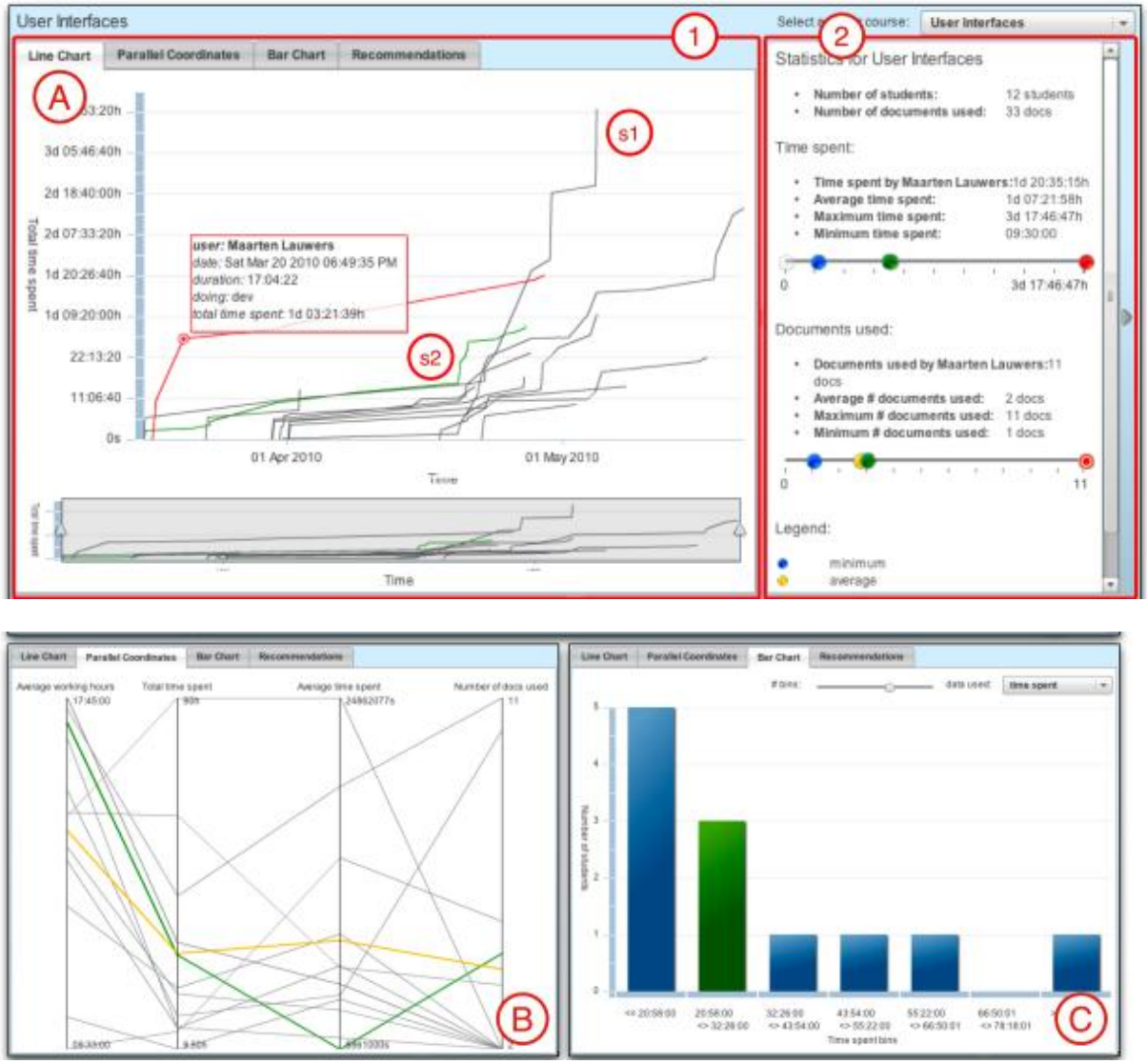
La analítica educativa busca precisamente convertir esos datos en conocimiento aplicable. Al implementar procesos de análisis de datos educativos, las instituciones pueden tomar decisiones informadas y diseñar planes de acción basados en evidencia objetiva, en lugar de hacerlo únicamente por intuición o experiencia previa. De esta manera, la analítica de datos se convierte en una herramienta estratégica para la gestión educativa, permitiendo, por ejemplo, identificar necesidades de apoyo académico, evaluar la efectividad de programas formativos o redistribuir recursos de forma más eficiente (Aristizabal F., 2016).

Además de su papel estratégico en la gestión o dirección de la institución, la analítica de datos tiene un impacto directo en el aula. Una de las herramientas más representativas son los tableros de control (*dashboards*), que permiten a docentes, estudiantes y gestores visualizar de manera clara información relevante sobre asistencia, calificaciones, tendencias académicas o progreso a lo largo del tiempo. Estos paneles analíticos facilitan la detección temprana de estudiantes en riesgo y constituyen un apoyo valioso para la toma de decisiones pedagógicas y administrativas (Aristizabal F., 2016).

Un ejemplo representativo de este tipo de herramientas es el *Student Activity Monitor* (SAM), diseñado para facilitar tanto el seguimiento personal del aprendizaje por parte del alumnado como la supervisión por parte del profesorado. Como se aprecia en la Figura 3, el panel combina diferentes visualizaciones: líneas que reflejan la evolución del tiempo de dedicación de cada estudiante, estadísticas globales del curso (tiempo empleado y documentos consultados) y coordenadas paralelas para comparar patrones de

comportamiento a través de distintas dimensiones, como el número de documentos utilizados, el tiempo promedio por documento o el momento del día en que se suele estudiar. Gracias a esta información, es posible detectar perfiles de trabajo distintos, por ejemplo, estudiantes que concentran la actividad en pocos días frente a quienes mantienen un ritmo más constante y ofrecer recomendaciones personalizadas. Aunque se trata de un panel complejo, los estudios de evaluación demostraron que los estudiantes lo percibieron como una herramienta clara, útil y organizada (Duval, 2011).

Figura 3: Ejemplo de *dashboard* educativo (*Student Activity Monitor*, SAM)



Fuente: Duval (2011)

Un aspecto ilustrativo de esta tendencia lo aporta un estudio realizado con gestores de universidades chilenas. La mayoría de los participantes manifestó contar con conocimientos y prácticas relacionadas con la recopilación y utilización de datos para la mejora educativa. Asimismo, el estudio destacó especialmente la importancia del perfil de ingreso de los alumnos, es decir, las características académicas y personales con las que acceden a la universidad, junto con sus primeras calificaciones, como factores clave para identificar de manera temprana a los estudiantes en riesgo. Este tipo de información permite diseñar intervenciones preventivas desde el inicio de la vida universitaria, como tutorías o cursos de nivel, en línea con la lógica de las alertas tempranas en educación superior (Villarroel-Henríquez & Gallardo-Aguayo, 2025).

Además, los resultados del estudio muestran la frecuencia con la que los gestores emplean los datos en distintos ámbitos de la gestión educativa, como se recoge en la Tabla 1.

Tabla 1: Frecuencia de la gestión y uso de datos

Indicadores	Media	Porcentaje de respuestas positivas (6-10)
Uso de datos para analizar el nivel de aprendizaje logrado	7,77	86,5%
Uso de datos para diagnosticar o prevenir situaciones educativas	7,22	80%
Uso de datos para proponer soluciones a dificultades detectadas	7,48	83,3%
Uso de datos para determinar logro de competencias del perfil de egreso	7,12	80%
Uso de datos para predecir el futuro desempeño del egresado	6,09	56,7%
Uso de datos para analizar procesos de enseñanza-aprendizaje	7,45	79,9%

Fuente: Elaboración propia a partir de Villarroel-Henríquez & Gallardo-Aguayo (2025)

Como se aprecia en la Tabla 1, el uso de datos se concentra principalmente en el análisis del nivel de aprendizaje logrado, la propuesta de soluciones a dificultades detectadas y la mejora de los procesos de enseñanza-aprendizaje, todos ellos con medias superiores a 7

y cerca del 80% de respuestas positivas. En contraste, el uso de datos con fines predictivos, como anticipar el desempeño futuro de los egresados, es el menos frecuente, lo que evidencia que este ámbito aún está poco desarrollado en las instituciones encuestadas.

Investigaciones recientes han mostrado que la analítica de datos también se aplica en entornos de aprendizaje no presencial. Por ejemplo, Hao et al. (2022) revisan diversos modelos de predicción del rendimiento en cursos masivos abiertos *online* (MOOCs), orientados a identificar de manera temprana a estudiantes en riesgo de bajo desempeño o abandono y a facilitar mecanismos de retroalimentación personalizada.

Pese a los avances descritos, la implementación de procesos de analítica de datos en el ámbito educativo no está exenta de dificultades. Gestores universitarios señalan obstáculos como la falta de sistematización en la recolección de datos, la escasez de tiempo y recursos financieros para analizarlos en profundidad, la ausencia de colaboración institucional y la resistencia al cambio. También se identifican limitaciones técnicas, como la actualización insuficiente de la información, la carencia de modelos predictivos y la falta de personal especializado para gestionar los datos. Estas barreras explican por qué, a pesar del avance hacia una cultura basada en datos, el aprovechamiento de la analítica en educación superior sigue siendo desigual y enfrenta importantes desafíos (Villarroel-Henríquez & Gallardo-Aguayo, 2025).

2.3. Problemática de la planificación académica en la universidad

La planificación académica en el contexto universitario consiste en organizar de antemano la oferta de asignaturas, la asignación de grupos y horarios, y los recursos docentes y físicos necesarios para cada periodo académico (semestre o curso lectivo). Implica, esencialmente, decidir qué asignaturas se ofrecerán, cuántos grupos de cada una, con cuántos estudiantes, qué profesores las impartirán y en qué aulas o franjas horarias. Esta tarea, que a primera vista podría parecer rutinaria, enfrenta en la práctica múltiples desafíos y limitaciones que la convierten en un problema complejo de gestión.

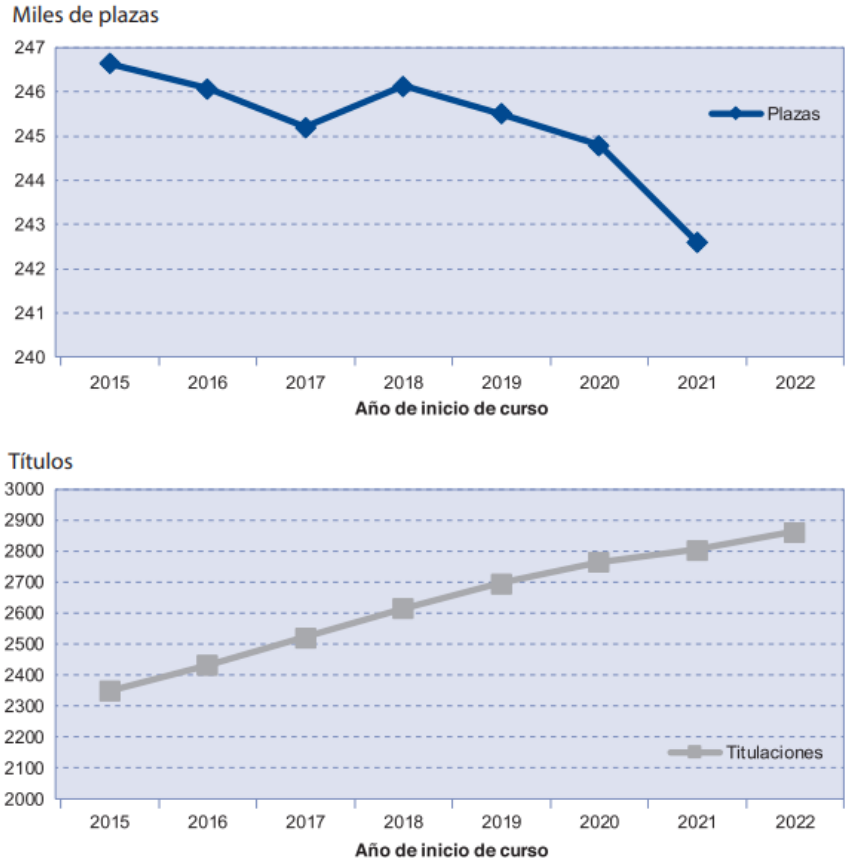
Uno de los principales retos es la incertidumbre en la demanda: la cantidad y el perfil de estudiantes matriculados puede variar significativamente de un curso a otro, dificultando ajustar la oferta a las necesidades reales. Por ejemplo, si se infraestiman las

matriculaciones en una asignatura obligatoria, puede ocurrir que se oferten menos grupos de los necesarios, generando clases sobresaturadas y estudiantes sin posibilidad de matricularse. Por el contrario, sobreestimar la demanda puede llevar a crear grupos que luego quedan con baja matrícula, lo que supone un desperdicio de recursos (horas de docencia y aulas infrautilizadas). Este delicado equilibrio entre oferta y demanda estudiantil está en el centro de la problemática de la planificación académica (Hurtado et al., 2021).

A lo anterior se suma la creciente diversidad en el perfil de los estudiantes que acceden a la universidad. En instituciones con una gran oferta académica como la UCM, es habitual que los alumnos de nuevo ingreso provengan de distintos sistemas educativos y contextos socioeconómicos, lo que se traduce en niveles de preparación desiguales y expectativas diversas. Esta heterogeneidad constituye un desafío para la planificación académica, pues no todos los estudiantes afrontan las asignaturas de primer curso en igualdad de condiciones. Como señalan Cabrera et al. (2006), las diferencias en el bagaje académico y social del alumnado están estrechamente relacionadas con mayores dificultades de integración en la vida universitaria y, en consecuencia, con un incremento del riesgo de abandono. Por ello, los autores subrayan la necesidad de que las instituciones desarrollen programas de apoyo específicos, tutorías y cursos de nivel que atiendan esta diversidad y contribuyan a mejorar la permanencia y el éxito académico.

Además, la problemática de la planificación académica debe analizarse también en el conjunto del sistema universitario. En el caso español, se ha evidenciado un desajuste creciente entre la oferta de titulaciones y la demanda de los estudiantes. Tal como se muestra en la Figura 4, el número de titulaciones presenciales ofertadas en las universidades públicas en España ha aumentado de forma continua en los últimos años, mientras que el número de plazas de nuevo ingreso disponibles ha disminuido. Esto significa que, aunque existen más opciones de estudio, el sistema acoge a un menor número de estudiantes, generando una paradoja entre la diversificación de la oferta y la reducción de la capacidad total. Esta divergencia refleja un desequilibrio estructural que dificulta el acceso de los estudiantes, genera ineficiencias en la utilización de recursos y limita la capacidad de respuesta del sistema universitario (Lacuesta et al., 2024).

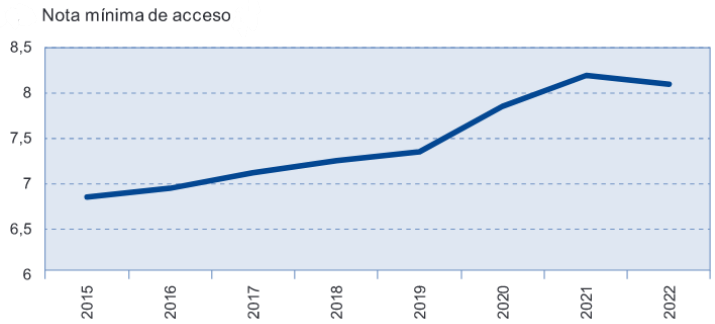
Figura 4: Evolución del número de plazas y titulaciones en universidades públicas presenciales



Fuente: Lacuesta et al. (2024)

A este fenómeno se suma el incremento sostenido de las notas mínimas de acceso a la universidad pública, reflejo de la elevada demanda concentrada en determinadas titulaciones. Como se observa en la Figura 5, la nota media mínima de acceso ha aumentado de manera continua en los últimos cursos académicos, lo que confirma la creciente presión sobre las titulaciones más demandadas. Según Lacuesta et al. (2024), estas carreras, que requieren expedientes más altos, son precisamente las que ofrecen mejores oportunidades laborales y salariales a los egresados. Este hecho pone de relieve una doble problemática: por un lado, las titulaciones con mayor demanda y mejores salidas profesionales restringen su acceso a un grupo reducido de estudiantes; por otro, se refuerzan las desigualdades de acceso y se dificulta la equidad en el sistema universitario. La ausencia de mecanismos ágiles para acompasar oferta y demanda impide optimizar recursos y limita la función de la universidad como instrumento de movilidad social.

Figura 5: Incremento de la nota mínima de acceso a la universidad pública presencial



Fuente: Lacuesta et al. (2024)

En definitiva, los problemas identificados en la planificación académica, tanto a nivel institucional como en el sistema universitario en su conjunto, ponen de relieve la necesidad de incorporar herramientas de análisis de datos propias del ámbito de *Business Analytics* que permitan anticipar tendencias y segmentar la demanda de los estudiantes. En este contexto, técnicas como el *clustering* se presentan como un recurso especialmente útil para identificar perfiles diferenciados de alumnado y facilitar una planificación académica más eficiente, cuestión que se desarrolla en el siguiente apartado.

2.4. Fundamentos y técnicas de *clustering*

El *machine learning* se clasifica habitualmente en tres grandes grupos en función del tipo de algoritmos empleados: supervisados, semi-supervisados y no supervisados. En los algoritmos supervisados se utilizan datos etiquetados, es decir, conjuntos de datos en los que cada observación incluye la respuesta o categoría correcta. Estos datos permiten entrenar modelos capaces de predecir la etiqueta de nuevas observaciones. Los algoritmos semi-supervisados combinan datos etiquetados con datos no etiquetados, entendidos estos últimos como observaciones que carecen de una categoría o valor objetivo asociado. El uso conjunto de ambos tipos de datos mejora el rendimiento del modelo cuando la cantidad de datos etiquetados es limitada. Por su parte, los algoritmos no supervisados trabajan exclusivamente con datos no etiquetados y buscan descubrir estructuras o patrones ocultos en ellos, agrupando los elementos en función de su similitud. Dentro de este último grupo se encuentra el *clustering*, técnica que constituye el objeto de este apartado (Russo et al., 2016).

El *clustering* constituye una de las técnicas más representativas del aprendizaje no supervisado, cuyo objetivo principal es agrupar observaciones de forma que las que pertenecen a un mismo grupo presenten características más similares entre sí que respecto a los de otros grupos. Se trata de un método exploratorio ampliamente utilizado en disciplinas como la minería de datos, el reconocimiento de patrones, la bioinformática y, en los últimos años, en el ámbito de la analítica de negocio (Madhulatha, 2012). Desde un punto de vista formal, esta técnica permite descubrir estructuras ocultas en los datos sin necesidad de categorías previamente definidas, recurriendo a medidas de distancia o similitud para evaluar el grado de cercanía entre observaciones.

Existen diversos enfoques metodológicos de *clustering*, cada uno con características propias. Entre los más utilizados se encuentran los algoritmos jerárquicos, que generan agrupamientos de manera sucesiva hasta construir dendrogramas, y los algoritmos particionales, como *k-means*, que establecen desde el inicio un número fijo de clústeres y los ajustan de manera sucesiva mediante iteraciones, recalculando los centroides hasta que las agrupaciones se estabilizan (Madhulatha, 2012).

No obstante, estos métodos presentan limitaciones cuando el *dataset* combina variables numéricas y categóricas. En particular, *k-means* se basa en la distancia euclídea y exige transformar todas las variables a formato numérico, lo que puede distorsionar la estructura real de los datos y generar agrupaciones poco representativas.

Dado que el *dataset* utilizado en este trabajo contiene tanto variables numéricas como categóricas, se emplean técnicas de *clustering* adecuadas para datos mixtos. En primer lugar, se utiliza el algoritmo *k-prototypes*, que extiende *k-means* incorporando una medida de disimilitud específica para variables categóricas y un parámetro de ponderación (γ) que equilibra la influencia relativa de variables numéricas y cualitativas. Además, se aplica el algoritmo *k-medoids*, que permite trabajar directamente sobre una matriz de disimilitud, en este caso, la distancia de Gower, y proporciona una alternativa más robusta frente a valores atípicos.

En todos estos algoritmos particionales, *k* representa el número de clústeres o grupos que se desean identificar en los datos. Este parámetro actúa como hiperparámetro del modelo y debe seleccionarse antes de ejecutar el algoritmo, puesto que su valor no está definido

a priori. La elección de k es un paso crucial, ya que determina la granularidad de la segmentación y se establece mediante técnicas específicas para estimar el número óptimo de clústeres.

2.4.1. Medidas de distancia

Los algoritmos de *clustering* particional se basan en la definición de una medida de disimilitud que permita cuantificar cuán diferentes son dos observaciones. En *datasets* mixtos, que incluyen simultáneamente variables numéricas y categóricas, esta cuestión es especialmente relevante, dado que el cálculo de distancias debe integrar tipos de información heterogéneos sin distorsionar la estructura real de los datos.

En este trabajo se emplean dos medidas de disimilitud complementarias: la utilizada por el algoritmo *k-prototypes* y, por otro lado, la distancia de Gower, empleada junto al algoritmo *k-medoids*. A continuación, se describen ambas.

Disimilitud utilizada en *k-prototypes*

El algoritmo *k-prototypes* combina dos tipos de medidas de disimilitud para poder trabajar con *datasets* mixtos, compuestos simultáneamente por variables numéricas y categóricas. La distancia total entre dos observaciones X y Y se obtiene mediante la función propuesta por Huang (1998):

$$d(X,Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

donde:

- p es el número de variables numéricas
- m es el número total de variables del *dataset*
- el primer término corresponde a la distancia euclídea cuadrática aplicada a las variables numéricas
- el segundo término incorpora la disimilitud para variables categóricas, definida mediante:

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

- Esta medida categórica, conocida como *simple matching dissimilarity*, contabiliza los desacuerdos entre categorías. Asigna una penalización de 1 a cada discrepancia entre categorías y 0 cuando coinciden; por tanto, cuantas menos discrepancias existan, mayor es la similitud entre las observaciones (Kaufman & Rousseeuw, 1990).
- La constante γ actúa como parámetro de ponderación que ajusta la contribución relativa de las partes numérica y categórica, evitando que una de ellas domine el cálculo de la distancia (Huang, 1998).

En conjunto, esta expresión permite que *k-prototypes* integre coherentemente información tanto cuantitativa como cualitativa, lo que lo convierte en un algoritmo especialmente adecuado para *datasets* mixtos como el empleado en este trabajo.

Distancia de Gower utilizada con *k-medoids*

La segunda medida utilizada en este trabajo es la distancia de Gower, ampliamente reconocida como el método estándar para medir similitud o disimilitud entre observaciones que contienen variables de distinta naturaleza (numéricas, categóricas u ordinales) (D'Orazio, 2024). Su funcionamiento se basa en calcular una similitud parcial por variable y, posteriormente, obtener una similitud global mediante una media ponderada.

- **Variables categóricas**

  Cuando la variable *k* es categórica, la similitud entre las observaciones *i* y *j* se define de forma muy sencilla:

  - $s_{ijk} = 1$ si ambas categorías son iguales
  - $s_{ijk} = 0$ si son distintas

- **Variables numéricas**

  En el caso de variables numéricas, la similitud se calcula como:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R(k)}$$

  donde $R(k)$es el rango de la variable k. La normalización mediante el rango garantiza que la similitud tome valores entre 0 y 1 y evita que las variables con magnitudes mayores dominen la medida de distancia.

- **Similitud global**

  Una vez obtenida la similitud para cada variable, la similitud global entre dos observaciones i y j con p variables se obtiene mediante:

$$S_{ij} = \frac{\sum_{k=1}^{p} w_{ijk}\ s_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

  donde w_{ijk} es el peso asignado a la variable *k.* En la práctica habitual, todas las variables reciben el mismo peso, por lo que el valor por defecto es w_{ijk} = 1.

  Según la definición original de Gower (1971, citado en D'Orazio, 2024), la distancia de Gower es simplemente el complemento de la similitud global: $D_{ij} = 1 - S_{ij}$. Esta expresión transforma la similitud en una matriz de disimilitud, plenamente compatible con algoritmos particionales basados en distancias.

  Como señalan Bektas & Schumann (2019), la distancia de Gower es especialmente adecuada para *datasets* mixtos, ya que produce una matriz de disimilitud simétrica y acotada entre 0 y 1, que puede emplearse directamente en el algoritmo *k-medoids* para la segmentación con variables heterogéneas.

2.4.2. Algoritmo *k-prototypes*

El algoritmo *k-prototypes*, introducido por Huang (1998), es una extensión natural de *k-means* y *k-modes* para *datasets* mixtos, combinando el uso de medias para variables numéricas y modas para variables categóricas. Esta integración permite representar cada clúster mediante un prototipo híbrido y calcular la disimilitud de forma coherente entre

observaciones con ambos tipos de atributos.

Para describir el funcionamiento del algoritmo, se sigue el procedimiento presentado por Madhuri et al. (2014), quienes implementan *k-prototypes* combinando *incremental k-means* y *modified k-modes* para mejorar la eficiencia del proceso de particionado.

Fundamento del algoritmo

K-prototypes minimiza una función de coste que combina:

- la suma de errores cuadráticos para variables numéricas, y
- la suma de desacuerdos para variables categóricas, ponderados mediante γ

Cada prototipo de clúster está formado por:

- la media de las observaciones asignadas para los atributos numéricos
- la moda de las observaciones asignadas para los atributos categóricos

Procedimiento operativo

Según Madhuri et al. (2014) el procedimiento del algoritmo puede resumirse en las siguientes etapas:

1. Seleccionar un *dataset* que contenga simultáneamente variables numéricas y categóricas, condición necesaria para aplicar *k-prototypes*.
2. Elegir las variables concretas que formarán parte del proceso de *clustering*, en función de los objetivos analíticos.
3. Determinar el número de clústeres k que se desea generar.
4. Inicializar los prototipos (centroides) de los clústeres:
    a. Para las variables numéricas, se emplean las medias iniciales (*como en k-means*).
    b. Para las variables categóricas, se utilizan las modas (como en *k-modes*).
5. Para cada observación, calcular su disimilitud respecto a cada clúster utilizando:
    a. Distancia euclídea para las variables numéricas

b. *Simple matching dissimilarity* para las variables categóricas

6. Para cada observación, sumar la parte numérica y la parte categórica de la disimilitud (ponderada por γ) y asignarla al clúster cuyo prototipo minimice la distancia total.
7. Repetir los pasos 5 y 6 hasta que no se produzcan cambios en las asignaciones.
8. Calcular el coste total del *clustering*, que combina: la suma de los errores cuadráticos para las variables numéricas, y la suma de desacuerdos para las categóricas.

Este procedimiento convierte a *k-prototypes* en una solución eficaz para el análisis de *datasets* mixtos, integrando de forma coherente información numérica y categórica a través de un esquema unificado de distancia.

2.4.3. Algoritmo *k-medoids*

El algoritmo *k-medoids* es un método de *clustering* particional relacionado con *k-means*, pero más robusto frente a valores atípicos. A diferencia de *k-means*, que utiliza centroides calculados como medias, *k-medoids* representa cada clúster mediante un *medoid*, es decir, una observación real cuya disimilitud promedio con el resto de elementos del clúster es mínima (Kassambara, 2017). Esto permite trabajar con cualquier matriz de disimilitud, como la definida previamente en el apartado 2.4.1.

El procedimiento más habitual para implementar *k-medoids* es el algoritmo PAM (*Partitioning Around Medoids*), propuesto por Kaufman & Rousseeuw (1990). Su funcionamiento puede resumirse en tres pasos:

1. Selección inicial de *medoids*: se eligen *k* observaciones como representantes iniciales.
2. Asignación: cada observación se asigna al *medoid* más cercano según la disimilitud definida.
3. Optimización (fase SWAP): se evalúan posibles intercambios entre *medoids* y no-*medoids*; un intercambio se acepta únicamente si reduce la suma total de disimilitudes. El proceso continúa hasta que ningún cambio mejora el coste.

El uso de *medoids* hace que el algoritmo sea menos sensible al ruido y a los *outliers* y que los clústeres resultantes sean más interpretables, dado que cada grupo está representado por una observación real.

2.4.4. Selección del número óptimo de clústeres

La elección del número óptimo de clústeres constituye un paso fundamental en cualquier análisis de *clustering*, ya que determina la estructura y la interpretabilidad de los grupos resultantes. Aunque no existe un criterio universal para seleccionar el valor adecuado de *k*, la literatura propone varios métodos que permiten orientar esta decisión. Entre los más utilizados se encuentran el método del codo (*Elbow Method*) y el índice de silueta (*Average Silhouette Width*), ambos ampliamente empleados en algoritmos particionales (Kodinariya & Makwana, 2013).

Método del codo (*Elbow Method*)

En su formulación clásica para *k-means*, el método del codo se basa en analizar la evolución de la heterogeneidad intra-clúster mediante el *Within-Cluster Sum of Squares* (WCSS). En el caso del algoritmo *k-prototypes*, esta métrica se sustituye por la función de coste propia del algoritmo, que combina: el error cuadrático de las variables numéricas, y los desacuerdos categóricos ponderados.

Representar gráficamente esta función de coste para distintos valores de *k* permite identificar un punto de inflexión a partir del cual incrementar el número de clústeres no produce mejoras significativas. Ese punto se interpreta como un candidato razonable para el número óptimo de clústeres.

Método *Average Silhouette Width*

El índice de silueta evalúa la calidad del agrupamiento midiendo, para cada observación, su grado de proximidad al clúster asignado en comparación con otros posibles clústeres. Sus valores oscilan entre -1 y 1:

- Valores cercanos a 1 indican que la observación está bien asignada.
- Valores próximos a 0 sugieren que se encuentra entre dos clústeres.
- Valores negativos reflejan una asignación probablemente errónea.

Para encontrar el número óptimo de clústeres, se calcula la media del índice de silueta para cada valor de k. El valor de k que maximiza este promedio se considera el número óptimo de clústeres.

Ambos métodos buscan equilibrar la homogeneidad interna y la separación entre grupos, aunque no siempre conducen al mismo valor óptimo. Por ello, la decisión final debe apoyarse también en la interpretabilidad de los clústeres y en los objetivos específicos del análisis, especialmente en métodos aplicados a datos mixtos como *k-prototypes* y *k-medoids.*

2.5. Aplicaciones del *clustering* en educación y gestión universitaria

El *clustering* se ha consolidado en los últimos años como una herramienta clave en la analítica educativa, al permitir descubrir patrones ocultos en grandes volúmenes de datos relacionados con el rendimiento, los hábitos de estudio o las competencias del alumnado. Su aplicación resulta especialmente útil en el ámbito universitario, donde la diversidad de perfiles estudiantiles plantea un reto creciente para la planificación académica y la mejora de la experiencia formativa. Mediante la segmentación de estudiantes en grupos homogéneos, las universidades pueden anticipar necesidades, optimizar recursos y diseñar estrategias pedagógicas más ajustadas a la realidad de cada colectivo.

Un uso fundamental del *clustering* en educación ha sido la segmentación de estudiantes según sus patrones de aprendizaje y rendimiento, con el fin de mejorar los resultados académicos. Por ejemplo, un estudio reciente realizado durante la pandemia de la COVID-19 analizó las trayectorias de aprendizaje de 396 estudiantes del Grado de Ingeniería Química de la Universidad Autónoma de Madrid (UAM) mediante la plataforma adaptativa e-valUAM (Subirats et al., 2023). Aplicando algoritmos de *k-means* y de *clustering* jerárquico, los autores identificaron tres perfiles principales de estudiantes: (1) aquellos que estudian de forma continua a lo largo del semestre, (2) quienes concentraban el esfuerzo en los días previos a los exámenes, y (3) un grupo de bajo rendimiento sostenido durante el curso.

La Tabla 2 resume de forma adaptada los perfiles identificados, mostrando la nota final media y las principales características de cada grupo.

Tabla 2: Perfiles de estudiantes identificados mediante *clustering* en la UAM

Perfil de estudiante	Patrón de estudio	Nota examen final	Características destacadas
Constante	Estudio constante a lo largo del semestre	8,78	Mayor número de intentos y mejores resultados
De última hora	Estudio concentrado justo antes de los exámenes	7,91	Menor constancia, repuntes en las entregas
De bajo rendimiento	Bajo rendimiento sostenido	6,57	Pocos intentos, menor dedicación global

Fuente: Elaboración propia a partir de Subirats et al. (2023)

Esta segmentación no solo permitió predecir con antelación el desempeño académico, sino también identificar prácticas inadecuadas, como picos de estudio de última hora o posibles casos de copia, para poder corregirlas a tiempo. De hecho, el grupo de estudio continuo obtuvo las mejores calificaciones finales, confirmando que empezar a preparar las materias con al menos tres meses de antelación conduce a mayores probabilidades de éxito. Por otro lado, el clúster de bajo rendimiento presentó la tasa más alta de fracaso y riesgo de abandono.

Estos hallazgos dieron lugar a recomendaciones prácticas, como implementar mecanismos para fomentar hábitos de estudio constantes (por ejemplo, más evaluaciones periódicas, retroalimentación continua o técnicas de gamificación) con el objetivo de animar a los estudiantes rezagados a adoptar estrategias de estudio progresivas. En definitiva, mediante el *clustering* las universidades han logrado detectar a tiempo grupos vulnerables y actuar sobre ellos con programas de apoyo y seguimiento, reduciendo el riesgo de suspenso o abandono (Subirats et al., 2023).

Más allá de los hábitos de estudio, el *clustering* también ha demostrado su utilidad en el análisis de competencias transversales, donde permite identificar perfiles diferenciados y personalizar la enseñanza. Un claro ejemplo es la identificación de perfiles de pensamiento crítico entre estudiantes de universidades ubicadas en diferentes comunidades autónomas españolas, llevada a cabo por investigadores de la UCM (Vendrell-Morancho et al., 2024). En dicho estudio (con más de 5.000 estudiantes

encuestados) se midió el nivel de pensamiento crítico mediante un instrumento validado con escala tipo Likert, y posteriormente se aplicó *k-means* para agrupar a los estudiantes según sus puntuaciones. El análisis clúster reveló tres grupos bien diferenciados en esta competencia cognitiva: un clúster de pensamiento crítico alto, otro de nivel moderadamente alto y un tercer grupo de nivel medio.

Además, se observaron patrones de distribución asociados a factores sociodemográficos (por ejemplo, ciertas carreras o cursos académicos concentraban más estudiantes de pensamiento crítico elevado). Un resultado destacado fue la correlación positiva entre el pensamiento crítico y el rendimiento académico (medido por las calificaciones). Es decir, los estudiantes categorizados en el clúster de pensamiento crítico más alto tendían a obtener mejores notas, lo que sugiere que esta habilidad transversal influye en el éxito académico.

Gracias a esta segmentación, los autores abogan por programas formativos adaptados a cada perfil, de modo que se puedan reforzar las destrezas de pensamiento crítico especialmente en los grupos intermedios o rezagados. Este ejemplo evidencia cómo el *clustering* permite personalizar planes de estudio y recursos de apoyo según las necesidades detectadas: por ejemplo, talleres específicos para el grupo de nivel medio, mentorías para estimular a los de nivel moderado, o actividades de profundización para los más aventajados. De esta manera, la analítica educativa basada en *clustering* contribuye a mejorar competencias clave del alumnado, aumentando las probabilidades de éxito durante su trayectoria universitaria (Vendrell-Morancho et al., 2024).

En conjunto, los estudios de Subirats et al. (2023) y Vendrell-Morancho et al. (2024) muestran cómo el *clustering* puede emplearse tanto para analizar hábitos de estudio como para identificar competencias transversales, ofreciendo un abanico de aplicaciones que van desde la predicción del rendimiento hasta la personalización de la enseñanza. Estos resultados, obtenidos en el contexto universitario español, respaldan la utilidad del *clustering* como herramienta de apoyo estratégico en la gestión universitaria, lo que a su vez fundamenta su aplicación en este trabajo para la identificación de perfiles de estudiantes en la UCM.

3. ANÁLISIS DE DATOS Y SEGMENTACIÓN

3.1. Fuente de datos y descripción del *dataset*

Como se indicó en el apartado 1.4. Datos utilizados, el análisis empírico de este Trabajo de Fin de Grado se basa en información procedente del *dataset* 'UCM-Acceso', disponible en el portal UniversiDATA, iniciativa de datos abiertos impulsada por universidades públicas españolas. Este portal facilita la reutilización de información estadística universitaria, proporcionando datos anonimizados, estandarizados y comparables entre instituciones, con el fin de fomentar la investigación basada en datos.

El *dataset* utilizado contiene información detallada sobre los estudiantes que se matriculan por primera vez en titulaciones oficiales de Grado en la UCM. El periodo cubierto abarca los cursos académicos 2017/18 a 2024/25, lo que permite analizar la evolución del acceso universitario durante ocho años consecutivos, incluyendo la etapa de la pandemia de la COVID-19, factor clave para este estudio.

La granularidad del *dataset* (matrícula/centro/estudio/curso académico) implica que cada registro corresponde a un estudiante que se matricula por primera vez en un grado y centro concretos en un curso académico específico. El *dataset* original integra 78 variables agrupadas en cuatro grandes bloques:

- Académicas: curso de acceso, nota de admisión, rama de titulación, vía de acceso y especialidad formativa previa, entre otras.
- Sociodemográficas: género, edad, nacionalidad, procedencia territorial y naturaleza del centro de secundaria, entre otras.
- Socioeconómicas y familiares: nivel educativo y ocupación de progenitores, pertenencia a familia numerosa, entre otras.

Esta amplitud de variables hace posible realizar un análisis multidimensional de los perfiles de acceso, especialmente útil para la aplicación de técnicas de segmentación mediante *clustering*, con el fin de identificar grupos homogéneos de estudiantes según sus características.

Los datos se distribuyen en ocho ficheros independientes (uno por curso) en formato

Excel (.xlsx), con una estructura de columnas homogénea que facilita su consolidación en una único *dataset*. No obstante, fue necesario un proceso exhaustivo de unificación, limpieza y transformación para garantizar su calidad (descrito en el apartado 3.4), ya que existían valores ausentes y categorías residuales con bajo nivel de información.

En conclusión, 'UCM-Acceso' constituye una fuente fiable, representativa y de alta calidad para estudiar el acceso universitario y su evolución reciente en la UCM, lo que proporciona una base sólida para aplicar técnicas de análisis de datos propias del ámbito de *Business Analytics* orientadas a la identificación y análisis de perfiles de estudiantes.

3.2. Análisis exploratorio del *dataset* (EDA)

En este apartado se desarrolla un análisis exploratorio del *dataset* proporcionado por UniversiDATA, con el objetivo de comprender su estructura inicial, identificar posibles problemas de calidad y examinar las características generales del alumnado que accede a estudios oficiales de Grado en la UCM durante el periodo 2017/18 - 2024/25 (ambos cursos incluidos).

El *dataset* inicial está compuesto por un total de 120.303 registros, cada uno correspondiente a una matrícula de nuevo ingreso. Incluye variables de distinta naturaleza:

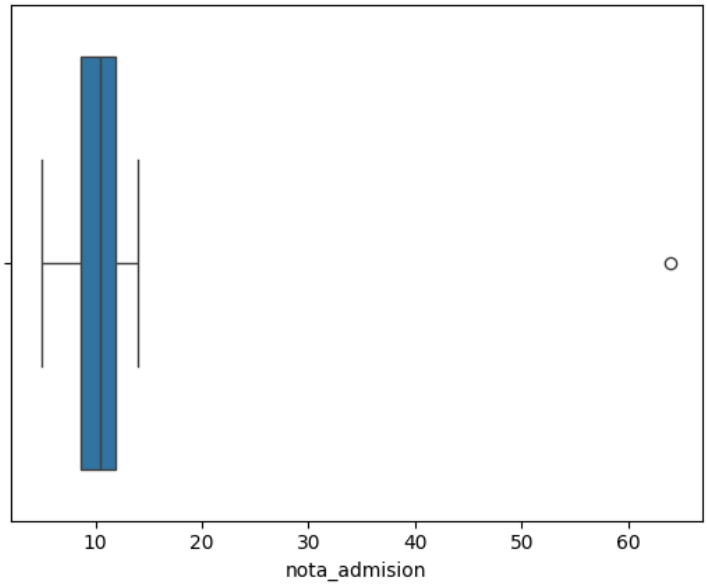
- Numéricas, como la nota de admisión (nota_admision).
- Categóricas nominales, como el género del estudiante (cod_genero) o la titulación universitaria (cod_titulacion).
- Categóricas ordinales, como el nivel de estudios de los progenitores (cod_nivel_estudios_padre, cod_nivel_estudios_madre).

Esta diversidad motiva la necesidad de una exploración preliminar que permita garantizar su correcta preparación en etapas posteriores del análisis.

La variable nota de admisión (nota_admision) constituye un indicador clave del rendimiento académico previo al acceso a la universidad. La distribución se concentra mayoritariamente entre 5 y 14 puntos, en línea con el sistema de calificación oficial en España. Como se observa en la Figura 6, la mediana se sitúa en torno a los 10 puntos, reflejando una tendencia ligeramente desplazada hacia valores altos. No obstante, se

detecta un valor atípico (*outlier*) igual a 64, claramente fuera del rango posible. Este caso se interpreta como un error de registro o una distorsión derivada de la anonimización aplicada al *dataset.* Dado el carácter exploratorio de esta fase, el valor se mantiene sin modificar en este punto del estudio, siendo su tratamiento diferido al apartado 3.4.

Figura 6: Boxplot de la variable nota de admisión (nota_admision) y detección de *outlier*

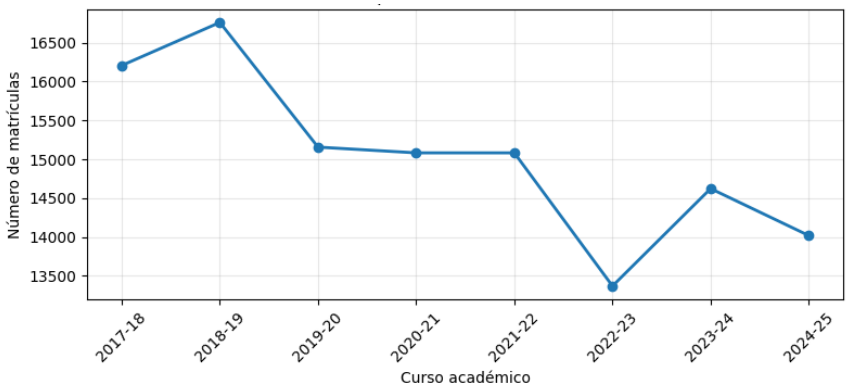


Fuente: Elaboración propia

El análisis descriptivo también revela la existencia de variables con una distribución muy concentrada en una única categoría. Por ejemplo, más del 80% del alumnado se matricula por primera vez directamente en primer curso (ind_se_matricula_en_primero) y más del 90% accede por primera vez al Sistema Universitario Español (ind_accede_al_SUE_por_1a_vez). Estas características sugieren que dichas variables, entre otras, podrían aportar una baja capacidad discriminativa en el análisis de segmentación, aunque su valoración definitiva se realizará posteriormente en el apartado 3.4.

Por otro lado, la Figura 7 muestra la evolución temporal del número de matrículas durante el periodo analizado. Se observa un máximo en el curso 2018/19 (aproximadamente 16.800 estudiantes), seguido de un descenso progresivo que alcanza su mínimo en 2022/23 (aproximadamente 13.400 estudiantes), para posteriormente recuperarse parcialmente en los cursos 2023/24 y 2024/25.

Figura 7: Evolución temporal del número de matrículas de nuevo ingreso en la UCM (2017/18 – 2024/25)



Fuente: Elaboración propia

Esta variabilidad en el acceso universitario motiva la clasificación temporal posterior en tres periodos (Pre-COVID, COVID y Post-COVID) con el fin de estudiar si la pandemia generó modificaciones en el perfil de acceso universitario. Esta estructura se desarrolla específicamente en el apartado 3.3.

Aunque podría plantearse si la evolución del número de matrículas está relacionada con la natalidad, en este caso no es así. Las cohortes que acceden a la universidad en los cursos analizados (aproximadamente 1999 - 2006) nacieron en un periodo en el que la natalidad en España aumentó de forma continuada (INE. Instituto Nacional de Estadística, 2025). Por tanto, la disminución de matrículas observada en algunos cursos no puede explicarse por un menor número de nacimientos, lo que refuerza la pertinencia de analizar separadamente los periodos Pre-COVID, COVID y Post-COVID.

Finalmente, el análisis preliminar de valores nulos revela diferencias significativas en la calidad del dato según la variable considerada. Algunas variables presentan registros completos, como el género (cod_genero) o la condición de familia numerosa (cod_familia_numerosa), mientras que otras muestran una ausencia importante de información.

En particular, las variables relacionadas con el país donde se cursaron los estudios de acceso (cod_pais_fin_estudio_acceso y derivadas) presentan un 100% de valores nulos, lo que confirma que no aportan información útil al análisis y anticipa su eliminación posterior.

Asimismo, la variable que representa la especialidad de estudios previos a la universidad (cod_especialidad_acceso) presenta porcentajes de ausencia considerables (aproximadamente 22,6%), probablemente derivadas de una falta de registro sistemático.

Por otro lado, variables sociodemográficas como la ocupación o nivel educativo de los progenitores muestran un número moderado de ausencias (entre 7% y 17%), atribuibles a omisiones en la declaración y falta de estandarización del registro.

Estos resultados evidencian la necesidad de definir estrategias de tratamiento de valores faltantes y de selección de variables antes de aplicar cualquier técnica de *clustering*. Dichos procedimientos se detallan en profundidad en el apartado 3.4.

3.3. Análisis del impacto de la COVID-19 en el acceso universitario

Para analizar rigurosamente las diferencias del perfil estudiantil entre los periodos Pre-COVID, COVID y Post-COVID, se emplea el conjunto de variables ya transformadas y consolidadas en el *dataset* final, entre ellas el nivel socioeconómico familiar (nivel_socieco_familia), la procedencia geográfica del centro de estudios previos (centro_en_madrid) y la rama de la titulación (rama_titulacion).

El proceso completo de depuración y construcción de estas variables se describe en el apartado 3.4.

Dado que el objetivo del análisis es estudiar el acceso a la universidad, la clasificación temporal se ha realizado atendiendo al momento en el que el estudiante inicia sus estudios (matrícula en septiembre). Con este criterio:

- El curso 2019/20 se clasifica como Pre-COVID, ya que los estudiantes accedieron en septiembre de 2019, antes del inicio de la pandemia y de cualquier alteración en los procesos de acceso o docencia.
- Los cursos 2020/21 y 2021/22 se consideran COVID porque, aunque la pandemia comenzó en marzo de 2020, el acceso universitario seguía condicionado por la excepcionalidad académica: evaluaciones adaptadas, efectos acumulados del confinamiento y, especialmente, la implantación generalizada de la docencia online o bimodal durante el curso 2021/22 en muchas universidades españolas.

- A partir del curso 2022/23, y especialmente en 2023/24 y 2024/25, la docencia y los procesos de acceso recuperan condiciones estables y plenamente presenciales, por lo que se clasifican como Post-COVID.

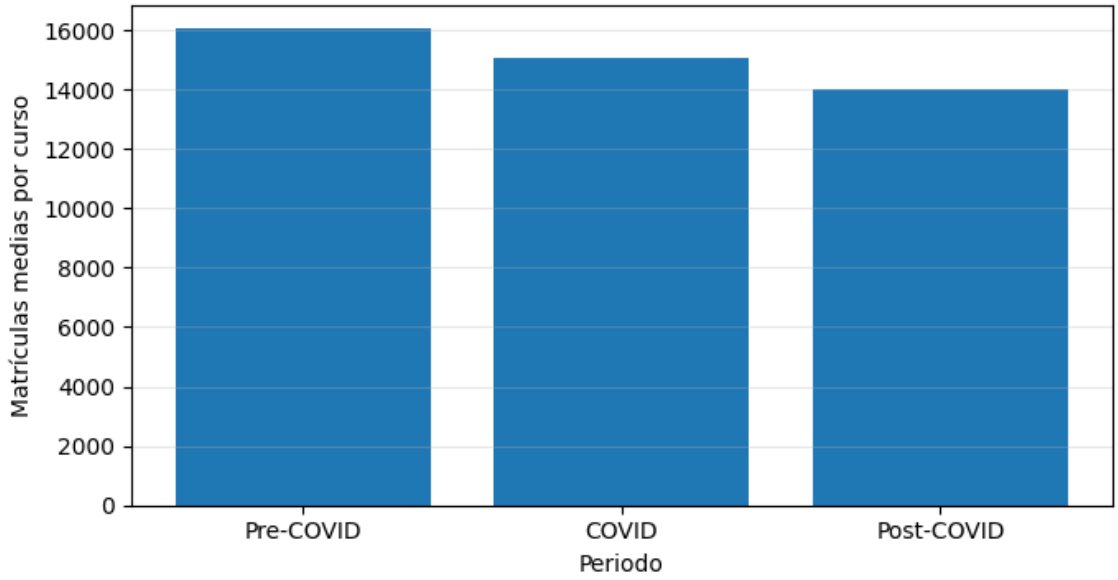
Con este criterio, los cursos quedan clasificados de la siguiente forma:

- Pre-COVID: 2017/18, 2018/19 y 2019/20
- COVID: 2020/21 y 2021/22
- Post-COVID: 2022/23, 2023/24 y 2024/25

3.3.1. Volumen de matrículas de nuevo ingreso

La Figura 8 muestra la evolución de las matrículas medias de nuevo ingreso por curso académico para cada periodo. Los resultados evidencian una disminución sostenida tras el estallido de la pandemia.

Figura 8: Evolución del acceso universitario por periodo COVID en la UCM



Fuente: Elaboración propia

Los resultados evidencian tres patrones diferenciados:

- Periodo Pre-COVID: la matrícula se encontraba en niveles elevados y relativamente estables, con una media de 16.041 estudiantes por curso. El máximo se registra en 2018/19, justo antes del inicio de la pandemia.

- Periodo COVID: se produce una reducción inicial, alcanzando una media de 15.083 estudiantes por curso. Aproximadamente el 6% respecto al periodo Pre-COVID. Esto sugiere un impacto directo de la pandemia sobre la decisión de matricularse, probablemente relacionado con la incertidumbre económica y social, las dificultades asociadas a la enseñanza *online* y las restricciones de movilidad territorial.
- Periodo Post-COVID: la media desciende todavía más hasta aproximadamente 14.004 estudiantes por curso. Aproximadamente -12,7% respecto a Pre-COVID y -7,1% respecto a COVID.

Aunque la demanda de estudios universitarios ha aumentado en los últimos años, especialmente en titulaciones con mejores perspectivas laborales, las universidades públicas presenciales no han logrado ajustar su oferta a este crecimiento. Según el Informe CYD 2024, esta falta de adaptación se debe a procesos burocráticos rígidos, limitaciones económicas y una reducida autonomía para ampliar o modificar titulaciones. Como consecuencia, las notas de corte han aumentado, generando barreras de acceso que afectan especialmente a estudiantes con menor nivel socioeconómico. Una parte de esta demanda desatendida se desplaza hacia las universidades privadas, que han ampliado su oferta y captan una proporción creciente del alumnado. Esta dinámica explica que, pese al incremento de las preinscripciones, la matrícula en las universidades públicas continúe descendiendo en el periodo Post-COVID (Fundación CYD, 2025).

3.3.2. Distribución del alumnado por rama de titulación

El análisis de la distribución del alumnado por rama de titulación revela una tendencia clara en el periodo estudiado. Se observa un creciente desplazamiento hacia titulaciones con mayores exigencias académicas y notas de admisión superiores, como Ingeniería y Arquitectura, Ciencias de la Salud y Ciencias. Estas ramas muestran un aumento progresivo en su peso relativo, especialmente en el periodo Post-COVID, coherente con la mayor demanda social y profesional de perfiles STEM (por sus siglas en inglés, *Science, Technology, Engineering and Mathematics*) y sanitarios.

En contraste, se aprecia una reducción sostenida en la proporción de estudiantes que acceden a titulaciones de Educación, Humanidades y Artes, y Ciencias Sociales y

Jurídicas, que tradicionalmente han concentrado un volumen mayor de alumnado. Aunque estas ramas siguen siendo mayoritarias en términos absolutos, su peso relativo disminuye de manera gradual a lo largo del periodo analizado.

En conjunto, estos resultados sugieren un cambio en las preferencias académicas del estudiantado, orientado hacia titulaciones percibidas como más competitivas y con mayores requisitos de acceso.

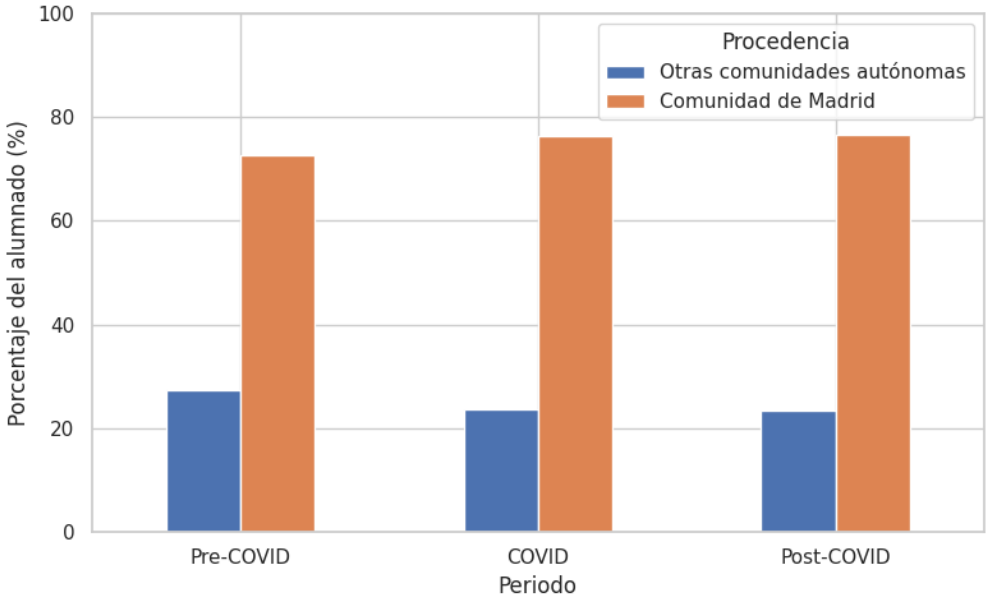
3.3.3. Movilidad interregional

La Figura 9 analiza la movilidad interregional del alumnado, medida a través de la variable procedencia geográfica del centro de estudios previos, que distingue entre estudiantes que cursaron sus estudios previos en la Comunidad de Madrid y aquellos procedentes de otra comunidad autónoma. Para este análisis descriptivo se han considerado únicamente los registros con información válida, con el fin de evitar que los valores nulos distorsionen las proporciones. No obstante, los casos sin información permanecen en el *dataset* completo y su tratamiento específico se aborda posteriormente en el apartado 3.4.

Los resultados evidencian un patrón claro: la movilidad interregional disminuye de forma apreciable con la llegada de la pandemia (del 27,4% en Pre-COVID al 23,6% en COVID). Este descenso se mantiene en el periodo Post-COVID, donde la proporción de estudiantes procedentes de fuera de la Comunidad de Madrid se sitúa incluso ligeramente por debajo (23,45%).

Por tanto, no se observa una recuperación hacia los niveles previos a la crisis sanitaria. La menor propensión a desplazarse para cursar estudios superiores parece haberse consolidado más allá del impacto inicial.

Figura 9: Movilidad interregional del alumnado según periodo COVID



Fuente: Elaboración propia

3.3.4. Nivel socioeconómico

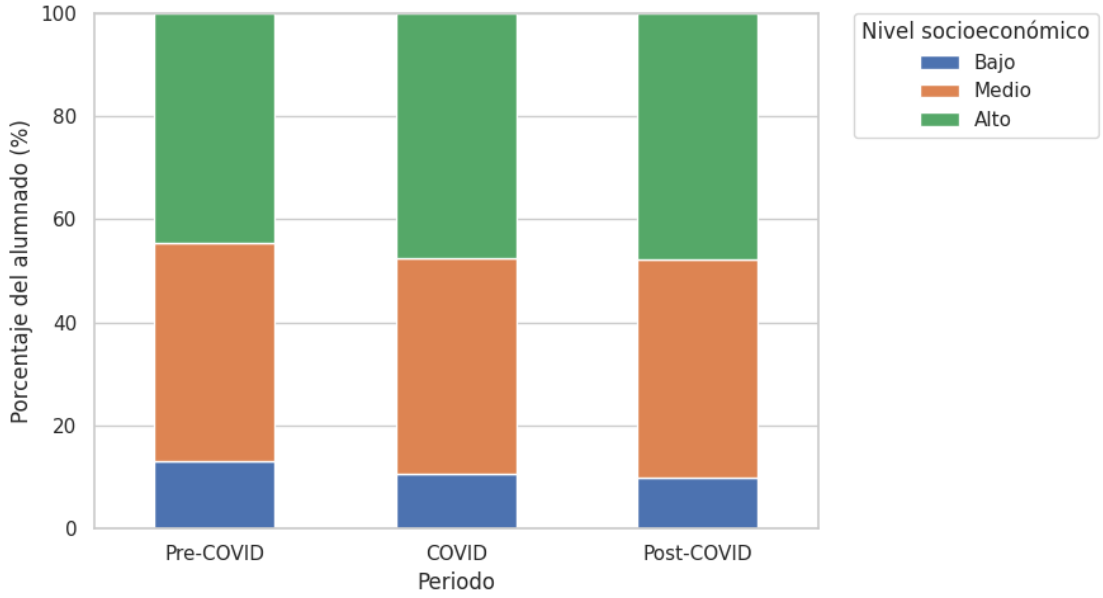
La Figura 10 muestra la evolución del nivel socioeconómico del alumnado a lo largo de los tres periodos analizados (excluyendo temporalmente la categoría "Sin información" para facilitar la visualización). La variable nivel socioeconómico familiar, cuyo proceso de construcción se detalla en el apartado 3.4, se basa exclusivamente en la ocupación de los progenitores, siguiendo el criterio del nivel ocupacional más alto registrado en el hogar.

En el periodo Pre-COVID, la distribución se reparte entre nivel alto (44%), medio (42%) y bajo (13%). Durante la pandemia, aumenta el peso del nivel alto (47,6%) y disminuye el del nivel bajo (10,8%). En el Post-COVID, esta tendencia se consolida: el nivel alto alcanza el 47,9% y el nivel bajo se reduce al 9,9%.

En conjunto, estos resultados sugieren que, tras la pandemia, los estudiantes procedentes de hogares con mayor estatus ocupacional refuerzan su presencia relativa en el acceso universitario, mientras que los perfiles socioeconómicos más vulnerables pierden peso proporcional. Esta evolución coincide con estudios que documentan un ensanchamiento de las brechas educativas y de participación tras la COVID-19 (OECD, 2021).

En particular, la OCDE señala que los estudiantes de origen socioeconómico bajo afrontaron mayores dificultades durante el aprendizaje a distancia, como peores condiciones para estudiar en casa, baja conectividad digital o un menor apoyo familiar. Estas limitaciones afectaron más a su continuidad educativa y a su preparación académica, lo que contribuye a que, en el periodo Post-COVID, su acceso relativo a la universidad se reduzca en comparación con los estudiantes de entornos más favorecidos (OECD, 2021).

Figura 10: Nivel socioeconómico del alumnado por periodo COVID



Fuente: Elaboración propia

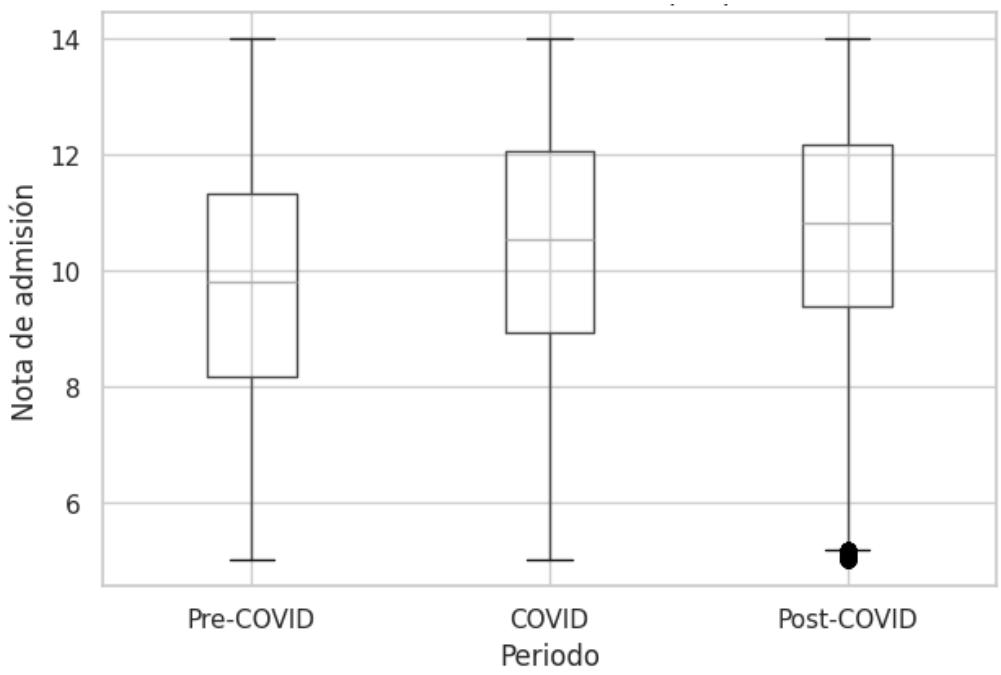
3.3.5. Nota de admisión

La Figura 11 muestra la evolución de la nota de admisión en los tres periodos analizados. Se observa un incremento progresivo desde el periodo Pre-COVID hasta el Post-COVID, lo que indica un cambio sostenido en el nivel de acceso al sistema universitario. En el periodo Post-COVID aparecen numerosos *outliers* asociados a notas de admisión cercanas al mínimo posible (5). Este comportamiento es esperable en distribuciones amplias con un límite inferior fijo y se explica porque la distribución central de las notas se desplaza hacia valores más altos, de modo que las notas bajas pasan a considerarse atípicas. Estos valores no alteran las tendencias generales observadas.

Para comprobar si las diferencias observadas en la nota de admisión entre periodos son estadísticamente significativas, se empleó una prueba ANOVA de un factor. Este método resulta adecuado porque permite contrastar si las medias de un mismo indicador, en este caso, la nota de admisión, son iguales o difieren entre más de dos grupos independientes (Pre-COVID, COVID y Post-COVID). Los resultados obtenidos (F = 2240.36; p-value = 0.000) muestran que las medias no son iguales entre los periodos analizados, lo que confirma la existencia de diferencias significativas en la nota de admisión a lo largo del tiempo.

Este aumento en la nota de admisión no solo puede interpretarse como una mejora en el rendimiento académico previo, sino también como el resultado de factores estructurales. Según el Informe CYD 2024, la creciente demanda de estudios universitarios, combinada con la limitada capacidad de adaptación de las universidades públicas, ha provocado un aumento generalizado de las notas de corte. Este fenómeno refleja una mayor competencia por acceder a determinadas titulaciones y puede estar contribuyendo al incremento observado en la nota de admisión durante el periodo Post-COVID (Fundación CYD, 2025).

Figura 11: Distribución de la nota de admisión por periodo COVID



Fuente: Elaboración propia

3.3.6. Síntesis final

Los resultados muestran que la composición del alumnado que accede a la UCM ha experimentado cambios significativos durante y después de la pandemia, tanto en volumen de matrícula como en movilidad geográfica, nivel socioeconómico y rendimiento académico previo. Estas dinámicas confirman que el perfil de acceso universitario no es estable, sino que responde a factores externos como la COVID-19.

En consecuencia, resulta pertinente aplicar técnicas de *clustering* para analizar si estas transformaciones se reflejan en la aparición de grupos diferenciados de estudiantes, así como para comprender mejor los patrones subyacentes que caracterizan el acceso a la educación superior en este periodo.

3.4. Procesamiento, selección y preparación del *dataset* final para *clustering*

Este apartado describe el proceso seguido para transformar el *dataset* original en una versión depurada, coherente y analíticamente sólida, adecuada para la aplicación de técnicas de segmentación mediante *clustering*. El objetivo es garantizar que las variables utilizadas aporten información relevante, estén libres de inconsistencias y reflejen adecuadamente los patrones académicos, geográficos y socioeconómicos del alumnado.

3.4.1. Depuración inicial del *dataset* y construcción de nuevas variables

Antes de proceder al análisis cuantitativo y al *clustering*, se realizó una depuración sistemática del *dataset* con el fin de garantizar la consistencia interna de los registros y la adecuación de las variables al objeto del estudio.

En primer lugar, se identificaron y eliminaron las filas duplicadas basadas en coincidencias exactas en todas las variables originales. Esta operación permitió evitar sesgos derivados de la repetición de registros y mantener únicamente observaciones únicas correspondientes a estudiantes en su primera matrícula de Grado.

Posteriormente, se generaron varias variables derivadas con el objetivo de capturar dimensiones relevantes para el análisis:

- edad_acceso: calculada a partir del año de nacimiento y el curso académico, permite analizar la edad real de acceso al sistema universitario.

- antiguedad_SUE: indicador de los años transcurridos desde que el estudiante accedió por primera vez al Sistema Universitario Español (SUE).
- nivel_socieco_familia: variable construida exclusivamente a partir de la ocupación de los progenitores, tomando el nivel ocupacional más alto registrado en el hogar.

Asimismo, se revisaron y recodificaron varias variables con el fin de mejorar su interpretabilidad y homogeneidad:

- centro_en_madrid: la variable original (cod_comunidad_centro_sec) presentaba múltiples códigos de procedencia y un volumen significativo de valores sin registrar. Para su utilización en el análisis, se recodificó en tres categorías: Comunidad de Madrid, Otras comunidades autónomas y Sin información.
- especialidad_rama: se deriva de cod_especialidad_acceso, que contenía más de 40 categorías. Por lo que se creó esta nueva variable que se agrupa en cinco áreas: Artes, Ciencias y Tecnología, Humanidades y CCSS, Formación Profesional y No clasificable.
- estudio_acceso_grupo: se deriva de cod_estudio_acceso, que es una lista de tipos de estudios según la ley educativa bajo la que se cursaron (LOGSE, LOE, LOMCE) o títulos equivalentes de FP. Se decidió crear esta nueva variable que se agrupa en tres categorías: Bachillerato, FP o equivalentes y Sin información.
- rama_titulacion: agrupación de las 99 titulaciones originales (cod_titulacion) en seis grandes ramas: Humanidades y Artes, Ciencias Sociales y Jurídicas, Ciencias de la Salud, Ciencias, Ingeniería y Arquitectura, Educación y Sin información.
- es_internacional: se deriva del nombre del país correspondiente a la nacionalidad del estudiante (des_pais_nacionalidad). Toma el valor 1 si la nacionalidad es extranjera y 0 si la nacionalidad es España.

Estas transformaciones facilitaron el análisis descriptivo, la comparabilidad entre estudiantes y la futura aplicación del *clustering*.

3.4.2. Detección y eliminación de *outliers*

Durante la exploración del *dataset* se identificaron dos tipos de *outliers*:

1. Registros con edades imposibles para el acceso universitario: se identificaron tres observaciones correspondientes a estudiantes nacidos en 2010 o 2019. Estos valores implican edades de entre 0 y 11 años en el momento de la matriculación universitaria, claramente incompatibles con cualquier itinerario formativo razonable, incluso en casos excepcionales de alumnado con altas capacidades. En uno de los registros, además, el año de acceso coincide incluso con el año de nacimiento. Por tanto, se consideraron errores de registro y fueron eliminados del *dataset*.
2. Nota de admisión = 64: este valor se sitúa completamente fuera del rango de calificación permitido (0-14), por lo que se considera un error evidente asociado a procesos de anonimización o codificación y fue eliminado en esta fase.

Estas medidas garantizan que las variables numéricas no distorsionen la estimación de medias y varianzas ni afecten al algoritmo de *clustering*, sensible a *outliers*.

3.4.3. Criterios de selección inicial de variables

El proceso de selección inicial de variables se realizó siguiendo una estrategia combinada basada en criterio experto, evaluación de completitud, capacidad discriminativa y utilidad para el posterior análisis de *clustering*. Este procedimiento permitió reducir la dimensionalidad del *dataset* original y garantizar que las variables retenidas aportaran información relevante y no redundante.

1. Revisión conceptual del diccionario de UniversiDATA (criterio experto): en una primera fase se revisó el diccionario de variables proporcionado por UniversiDATA, eliminando aquellas cuya información era irrelevante para el análisis o no aportaban valor al objetivo del estudio. Se excluyeron especialmente variables administrativas (códigos de universidad, centro o campus), geográficas demasiado detalladas, y descripciones textuales redundantes respecto a sus códigos.
2. Eliminación de variables sin información útil (ausencia total o casi total de registros válidos): a continuación, se emplearon funciones de diagnóstico del *dataset* en Python (como df.info()), que permiten identificar columnas con valores nulos en su totalidad o con un grado extremo de ausencia. Aquellas con 100% de

valores nulos o con un grado extremo de ausencia fueron descartadas, al no resultar recuperables mediante imputación ni útiles para el análisis. Ejemplos típicos: país de finalización de estudios de acceso y sus derivados.

3. Eliminación de variables empleadas únicamente para obtener otras nuevas: determinadas variables fueron utilizadas para generar atributos más informativos, tras lo cual se eliminaron para evitar redundancias. Es el caso, por ejemplo, del año de nacimiento (anio_nacimiento), que se transformó en edad de acceso (edad_acceso), o del año de acceso al SUE (anio_acceso_SUE), a partir del cual se construyó la variable que indica los años transcurridos desde que el estudiante accedió por primera vez al SUE (antiguedad_SUE).
4. Conservación de variables codificadas para garantizar compatibilidad con algoritmos de *clustering:* dado que muchos algoritmos trabajan mejor con valores numéricos, se mantuvieron ciertos códigos categóricos siempre que aportaran información o permitieran generar variables agregadas útiles. Es el caso de los códigos de titulación universitaria (cod_titulacion) o de los códigos de especialidad de acceso de los estudios previos (cod_especialidad_acceso), que contenían un elevado número de categorías. Para mejorar la utilidad analítica, estos códigos se agruparon en categorías más amplicas, generando las variables rama de titulación universitaria (rama_titulacion) y rama de la especialidad cursada en los estudios previos (especialidad_rama).
5. Eliminación de variables redundantes relacionadas con el nivel socieconómico. Asimismo, se decidió no incorporar al *dataset* final las variables relativas al nivel educativo de los progenitores (cod_nivel_estudios_padre y cod_nivel_estudios_madre). Aunque conceptualmente relevantes, su información se solapa en gran medida con la ocupación familiar, ya integrada en la variable de nivel socioeconómico de la familia (nivel_socioeco_familia) que toma como referencia el nivel ocupacional más alto del hogar. Mantener simultáneamente nivel educativo y ocupación habría incrementado la dimensionalidad del modelo sin aportar variación adicional relevante, por lo que se optó por conservar únicamente el indicador derivado de la ocupación, más directamente vinculado al estatus socioeconómico.

6. Además, se decidió no incluir la variable del curso académico de acceso (curso_academico) en el *dataset* final destinado al *clustering*. Aunque resulta fundamental para el análisis temporal desarrollado en el apartado 3.3, esta variable no describe características propias del estudiante, sino el momento en el que tuvo lugar su matrícula. Su inclusión podría inducir agrupaciones artificiales basadas en la cronología, y no en las características académicas, sociodemográficas o socioeconómicas del alumnado, generando clústeres diferenciados por año en lugar de por perfil estudiantil. Por este motivo, la variable del curso académico de acceso (curso_academico) se excluyó del *dataset* final y se empleó únicamente en los análisis temporales previos.

3.4.4. Tratamiento de valores nulos

Antes de aplicar cualquier técnica de *clustering*, se analizó la presencia de valores faltantes en el conjunto de variables preseleccionadas. La Figura 12 resume el conteo y porcentaje de valores nulos por variable mediante un mapa de calor.

Figura 12: Conteo y porcentaje de valores nulos por variable (formato de gradiente)

Variable	Nulos	Porcentaje (%)
cod_especialidad_acceso	27203	22.64
cod_naturaleza_centro_sec	24811	20.65
cod_ocupacion_padre	19779	16.46
cod_estudio_acceso	15835	13.18
cod_ocupacion_madre	14888	12.39
cod_nivel_estudios_padre	10770	8.96
cod_nivel_estudios_madre	8004	6.66
nota_admision	3491	2.91
cod_titulacion	212	0.18

Fuente: Elaboración propia

El análisis revela tres patrones diferenciados:

1. Variables con ausencia elevada de información (superior al 20%): destacan la especialidad de acceso de los estudios previos (cod_especialidad_acceso) y la naturaleza del centro de secundaria (cod_naturaleza_centro_sec). En ambos casos,

los valores ausentes no pueden imputarse de forma válida porque no existe evidencia suficiente para asignarles una categoría realista. Por ello, se optó por recodificar los valores nulos en una categoría informativa adicional ("No clasificable" o "Sin información"), preservando así la integridad del *dataset* sin introducir supuestos artificiales.

2. Variables sociodemográficas con ausencia moderada (entre 6% y 17%): incluyen las variables relacionadas con el nivel educativo y la ocupación de los progenitores.

    - Ocupación madre/padre: se crea una categoría adicional "Sin información", puesto que los valores ausentes representan una falta de respuesta significativa.

    - Nivel educativo madre/padre: se realiza imputación mediante moda, dado que el porcentaje de nulos es bajo y así se evita la creación de una categoría adicional ("Sin información").

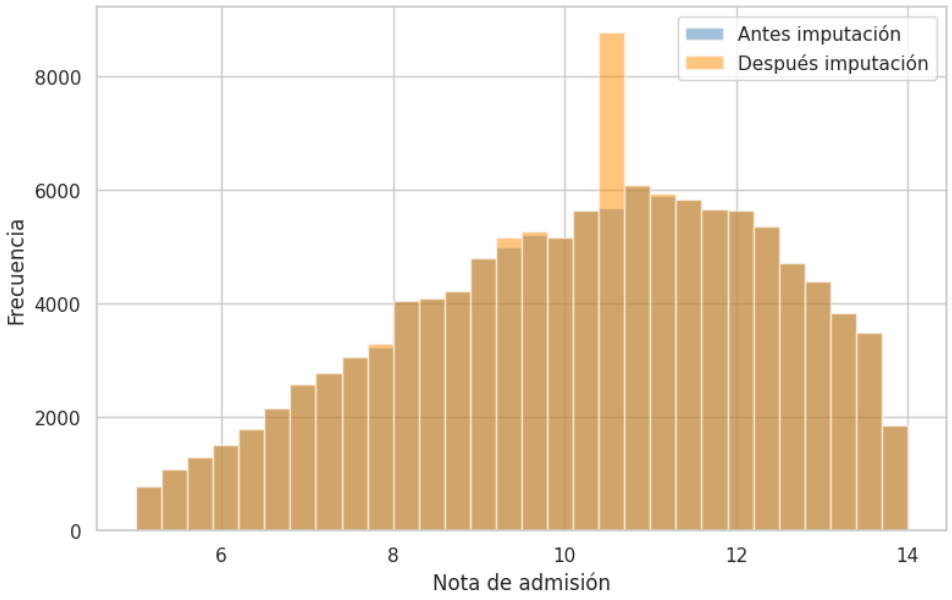
3. Variables casi completas (inferior al 3% de nulos):
    - Nota de admisión (nota_admision): presenta un porcentaje reducido de ausencias. En este caso sí es posible imputar usando información contextual: se asignó la mediana de la nota dentro de cada vía de acceso (cod_forma_admision). La elección de la mediana, en lugar de la media, se justifica porque es más robusta frente a valores extremos, de modo que evita que estos valores distorsionen la imputación y permite mantener diferencias reales entre los distintos itinerarios académicos.

    - Titulación universitaria (cod_titulacion): los registros con nulos fueron eliminados, dado que el volumen era muy reducido y la variable es esencial para la obtención de la variable rama de titulación universitaria (rama_titulacion).

Como se observa en la Figura 13, la comparación entre las distribuciones antes y después de la imputación muestra una forma prácticamente idéntica en términos de mediana, dispersión y asimetría. Únicamente se aprecia una ligera acumulación adicional alrededor de los valores más frecuentes, un efecto esperado al imputar múltiples registros con la

misma mediana. Esta variación mínima no altera la estructura general de la variable ni afecta a la validez del análisis posterior.

Figura 13: Distribución de la nota de admisión (nota_admision) antes y después de imputar nulos



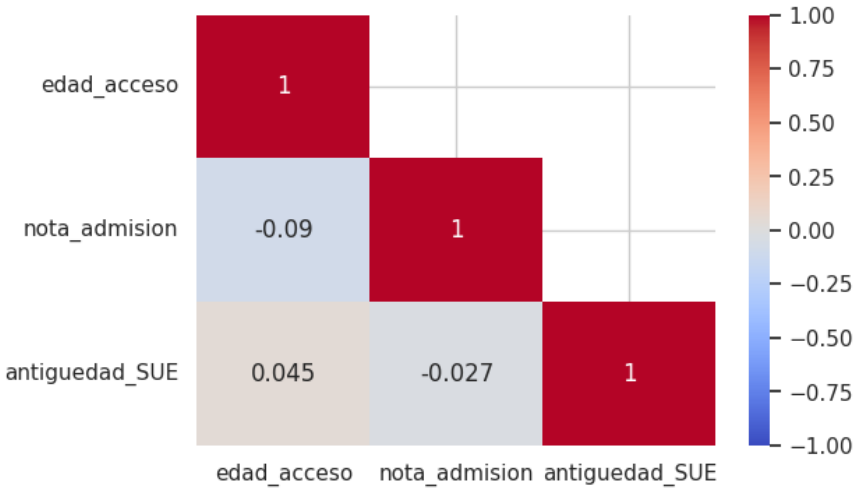
Fuente: Elaboración propia

3.4.5. Análisis de correlación y asociaciones entre variables

Para evaluar la presencia de relaciones fuertes entre las variables del *dataset* final y garantizar que no existan problemas de multicolinealidad antes del *clustering*, se estimaron correlaciones de Pearson para las variables numéricas y asociaciones mediante el coeficiente V de Cramer para las categóricas.

La Figura 14 muestra la matriz de correlaciones entre las variables edad de acceso (edad_acceso), nota de admisión (nota_admision) y años transcurridos desde que el estudiante accedió por primera vez al SUE (antiguedad_SUE). Las correlaciones son muy bajas en todos los casos (coeficientes inferiores a 0,10), lo que indica que no existe multicolinealidad entre estas variables. Este resultado confirma que pueden incorporarse simultáneamente al modelo de *clustering* sin riesgo de redundancia estadística.

Figura 14: Matriz de correlación de Pearson entre variables numéricas



Fuente: Elaboración propia

Para analizar la relación entre las variables categóricas del conjunto de datos se empleó el coeficiente V de Cramer, una medida basada en el estadístico chi-cuadrado y habitualmente utilizada para evaluar la fuerza de asociación entre variables cualitativas. Siguiendo el criterio de interpretación propuesto por Isea et al. (s. f.) los valores de V se clasifican de la siguiente forma:

- valores entre 0 y 0,2 indican ausencia de asociación
- un valor de 0,2 refleja una asociación débil
- valores entre 0,2 y 0,6 indican una asociación moderada
- valores entre 0,6 y 1 representan una asociación fuerte

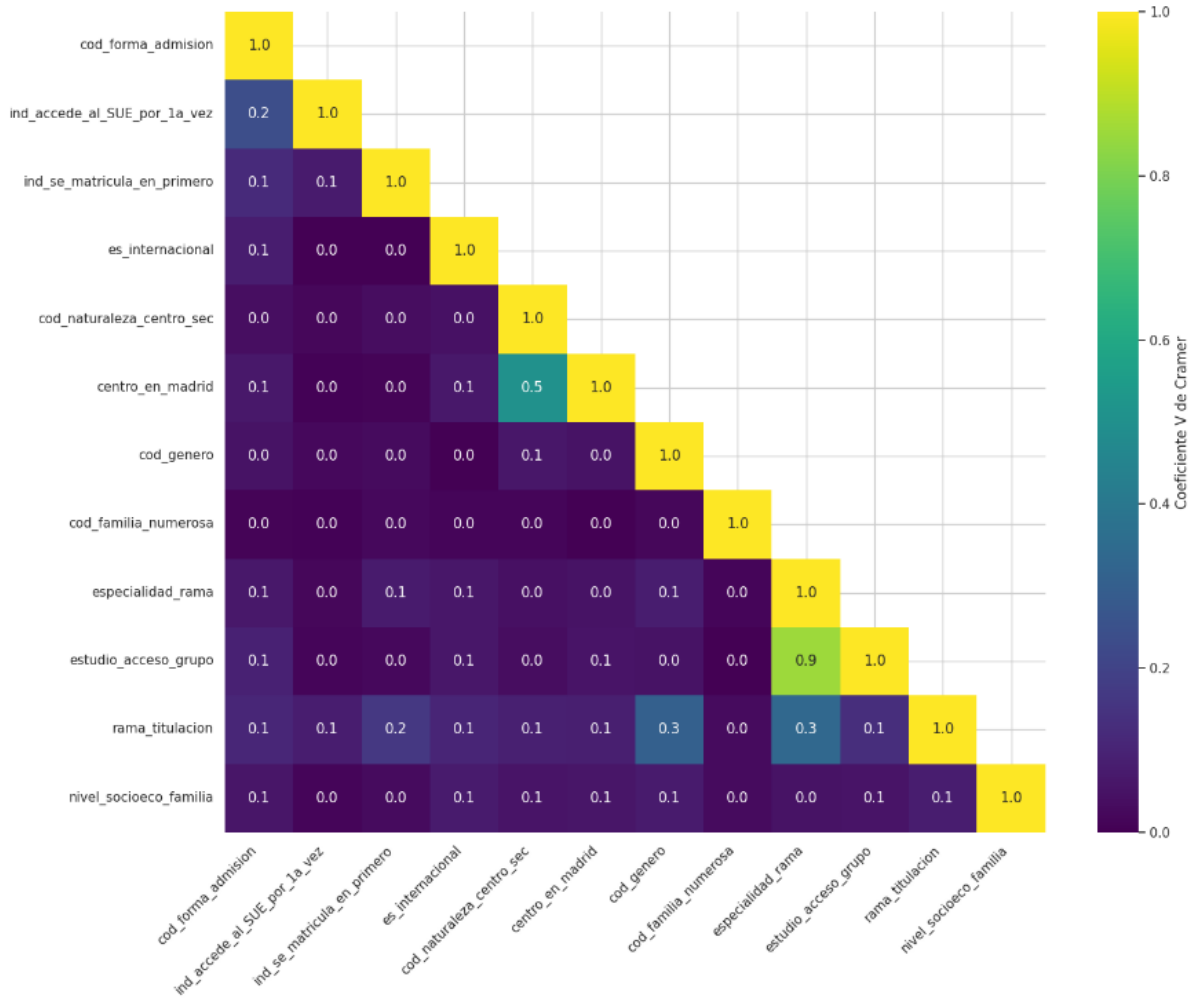
La Figura 15 presenta la matriz de asociaciones mediante el coeficiente V de Cramer para las variables categóricas utilizadas en el estudio. Los principales resultados son los siguientes:

- La mayoría de las asociaciones entre las variables categóricas presentan valores inferiores a 0,2, lo que indica que no existe solapamiento relevante de información entre ellas. Esto implica que cada variable aporta dimensiones distintas del perfil del estudiante, evitando problemas de redundancia en el análisis posterior.

- El análisis identifica varias asociaciones clasificadas como moderadas según el criterio utilizado. Todas ellas son coherentes con patrones esperables y, aunque muestran una relación relevante, no implican duplicidad conceptual entre las variables implicadas. Entre estas asociaciones destacan, por ejemplo, el acceso por primera vez al SUE (ind_accede_al_SUE_por_1ª_vez) frente a la vía de acceso (cod_forma_admision), el género del alumno (cod_genero) frente a la rama de la titulación universitaria (rama_titulacion), la rama de la especialidad cursada en los estudios previos (especialidad_rama) frente a la rama de la titulación universitaria (rama_titulacion), y la naturaleza del centro de secundaria (cod_naturaleza_centro_sec) frente a la procedencia geográfica del centro de secundaria (centro_en_madrid).

- La única asociación fuerte aparece entre el estudio de acceso al grado (estudio_acceso_grupo) y la rama de especialidad cursada en los estudios previos (especialidad_rama) (V = 0,85). Este valor, muy cercano a 1, indica que ambas variables representan esencialmente la misma información: el tipo de estudios previos del estudiante. Esto es coherente con la naturaleza de ambas variables: estudio_acceso_grupo clasifica de forma muy general (Bachillerato, Formación Profesional), mientras que especialidad_rama ofrece una diferenciación más detallada entre itinerarios (Ciencias y Tecnología, Humanidades y Ciencias Sociales, Artes, FP). Dado que especialidad_rama captura con mayor precisión la formación previa del alumnado, se decide conservarla y eliminar estudio_acceso_grupo por su menor capacidad informativa.

Figura 15: Matriz de asociación entre variables categóricas (Coeficiente V de Cramer)



Fuente: Elaboración propia

3.4.6. Identificación de variables con muy baja capacidad discriminativa

En el análisis descriptivo de las variables preseleccionadas para el *dataset* final se evaluó su distribución con el fin de detectar atributos cuya variabilidad fuese insuficiente para aportar información útil en el *clustering*. Se identificaron varias variables cuya distribución se concentra de forma extrema en una única categoría (más del 80% de los casos), lo que reduce de manera significativa su capacidad para diferenciar perfiles: acceso por primera vez al SUE (ind_accede_al_SUE_por_1ª_vez), matrícula en primer curso (ind_se_matricula_en_primero), modalidad de acceso (cod_forma_admision), nacionalidad extranjera o española (es_internacional) y años transcurridos desde el primer acceso al SUE (antiguedad_SUE).

Dado su escaso poder discriminativo y el riesgo de introducir ruido en algoritmos no supervisados, se decidió excluir estas variables del *dataset* final utilizado en la fase de *clustering*. Únicamente se conservaron aquellas variables con suficiente heterogeneidad y contribución informativa para distinguir adecuadamente entre distintos perfiles de estudiantes.

3.4.7. Estandarización de variables numéricas

Las variables numéricas edad de acceso (edad_acceso) y nota de admisión (nota_admision) se estandarizaron mediante StandardScaler, transformándolas a una escala con media 0 y desviación estándar 1. Esta estandarización garantiza que ambas contribuyan de forma equilibrada a los métodos de *clustering* utilizados.

Es importante destacar que la interpretación de los resultados se realiza siempre en la escala original, utilizando la estandarización únicamente como procedimiento técnico para el cálculo de distancias.

3.4.8. Resultado final del *dataset* para *clustering*

De las 78 variables originales del dataset 'UCM-Acceso', y tras aplicar los criterios de relevancia conceptual, calidad del dato y capacidad discriminativa descritos en los apartados anteriores, se seleccionó un subconjunto de 11 variables originales con utilidad potencial para el análisis *clustering*:

- Año de nacimiento (anio_nacimiento)
- Curso académico de acceso (curso_academico)
- Nota de admisión (nota_admision)
- Código de la titulación universitaria (cod_titulacion)
- Género del estudiante (cod_genero)
- Comunidad autónoma del centro de secundaria (cod_comunidad_centro_sec)
- Ocupación de la madre (cod_ocupacion_madre)
- Ocupación del padre (cod_ocupacion_padre)

- Especialidad de estudios previos (cod_especialidad_acceso)
- Naturaleza del centro de secundaria (cod_naturaleza_centro_sec)
- Condición de familia numerosa (cod_familia_numerosa)

Como resultado del proceso completo de depuración, el número total de registros queda establecido en 119.948 observaciones válidas, resultado de eliminar duplicados, registros sin información recuperable y *outliers* claramente no plausibles. Este volumen garantiza una base sólida y representativa para los algoritmos de segmentación.

A partir del subconjunto de 11 variables originales, y tras los procesos de transformación, recodificación, imputación y estandarización descritos en los apartados 3.4.1 - 3.4.7, se obtuvo el conjunto final de 9 variables utilizado en los modelos de *clustering*. Dicho conjunto integra dos variables numéricas estandarizadas y siete variables categóricas, en su mayoría transformadas o agregadas:

- **edad_acceso_scaled**: edad del estudiante en el momento de acceso al Grado, estandarizada.
- **nota_admision_scaled**: nota global de admisión al Grado, estandarizada.
- **rama_titulacion**: rama académica del Grado, derivada de la agrupación de más de 90 códigos iniciales.
- **cod_genero**: género del estudiante recodificado a formato binario (0 = Hombre, 1 = Mujer).
- **centro_en_madrid**: procedencia geográfica del centro de secundaria donde se cursó el último año de los estudios previos.
- **nivel_socioeco_familia**: nivel socioeconómico familiar construido a partir de la ocupación más alta de los progenitores (Sin información, Bajo, Medio, Alto).
- **especialidad_rama**: rama de especialidad de estudios previos a la universidad, agrupando más de 40 especialidades.

- **cod_naturaleza_centro_sec**: naturaleza del centro de secundaria en el que se cursó el estudio de acceso al Grado (Público, Privado, Concertado, Sin información).
- **cod_familia_numerosa**: condición de familia numerosa del estudiante, que se mantiene en su estructura original: 0 = No, 1 = Sí general, 2 = Sí especial.

En conjunto, estas 9 variables ofrecen una representación equilibrada de las dimensiones académicas, sociodemográficas, territoriales y socioeconómicas del alumnado, y constituyen la base final utilizada para la segmentación mediante *k-prototypes* y *k-medoids*.

A partir de estas variables es posible describir un perfil general del alumnado que accede a estudios de Grado en la UCM. En términos demográficos, la distribución por género presenta una mayor proporción de mujeres (63%), un patrón coherente con el predominio de ramas universitarias como Ciencias Sociales y Jurídicas o Educación, tradicionalmente más feminizadas. Desde el punto de vista territorial, aproximadamente una cuarta parte del estudiantado procede de fuera de la Comunidad de Madrid, mientras que el resto ha cursado sus estudios previos en centros madrileños.

La distribución socioeconómica familiar muestra una concentración en niveles medio y alto, mientras que la proporción de estudiantes procedentes de hogares de nivel bajo resulta comparativamente más reducida.

En cuanto a variables académicas, la mayoría del alumnado accede desde itinerarios de Humanidades y Ciencias Sociales o de Ciencias y Tecnología, mientras que las especialidades de Artes y la Formación Profesional representan proporciones menores. Asimismo, la edad de acceso se distribuye principalmente entre los 17 y 19 años, y la nota de admisión presenta valores coherentes con la escala habitual del Sistema Universitario Español.

3.5. Selección de la técnica de *clustering* adecuada

Una vez definido el *dataset* final para el análisis de *clustering*, se evaluaron dos técnicas capaces de trabajar con datos mixtos, es decir, que combinan simultáneamente variables numéricas y categóricas. En particular, se consideraron dos métodos ampliamente

utilizados en este tipo de contextos:

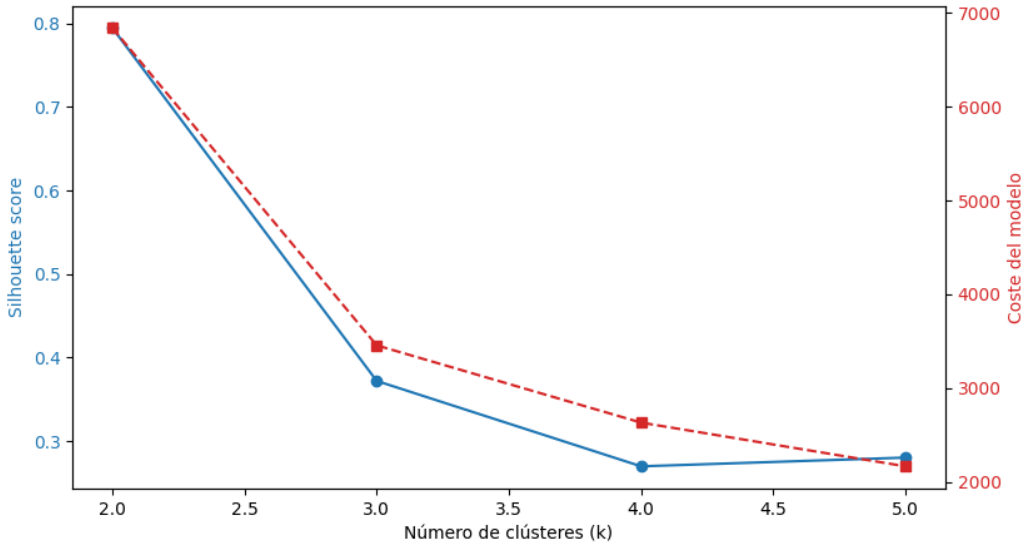
- *k-prototypes*, un algoritmo diseñado específicamente para datos mixtos al integrar los principios de *k-means* (para variables numéricas) y *k-modes* (para variables categóricas).
- *k-medoids* con distancia de Gower, una alternativa robusta que utiliza una métrica de disimilitud compatible con variables de distinta naturaleza y menos sensible a valores atípicos.

Para garantizar una comparación objetiva entre ambas técnicas, se utilizó una misma muestra aleatoria de 5.000 observaciones, sobre la que se calcularon dos métricas complementarias para valores de *k* entre 2 y 5:

1. *Silhouette score*, indicador de cohesión interna y separación entre grupos.
2. Función de coste, equivalente al método del codo (*elbow method*).

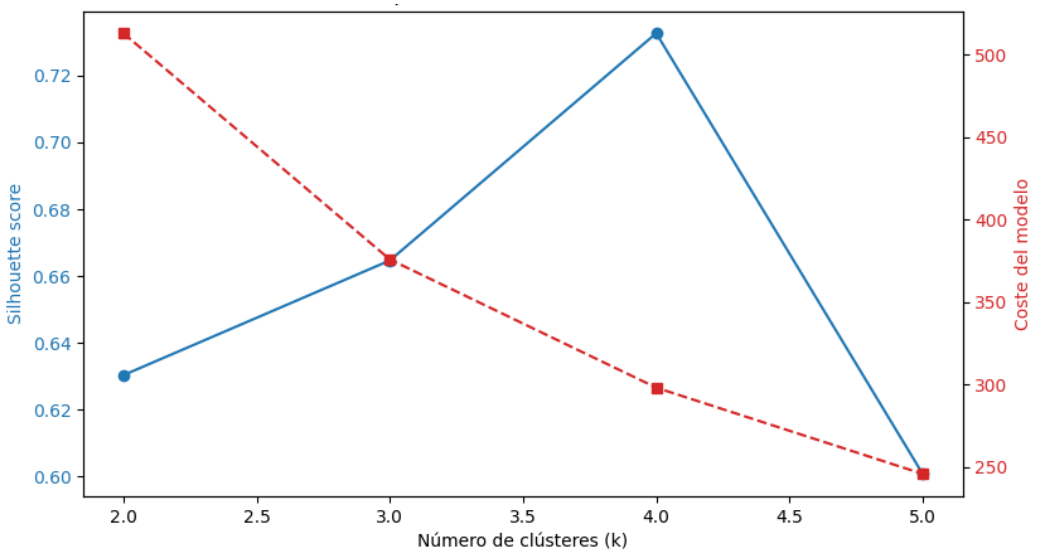
Las Figuras 16 y 17 muestran la evolución conjunta de ambas métricas para cada algoritmo, lo que permitió evaluar visualmente el comportamiento de los modelos al aumentar el número de clústeres.

Figura 16: Evolución del *silhouette score* y la función de coste (*k-prototypes*)



Fuente: Elaboración propia

Figura 17: Evolución del *silhouette score* y la función de coste (*k-medoids* con distancia de Gower)



Fuente: Elaboración propia

Los resultados mostraron que *k-prototypes* alcanzó el mayor *silhouette score* de ambos métodos, alcanzando un valor aproximado de 0,79 para $k = 2$, mientras que *k-medoids* obtuvo un valor máximo ligeramente inferior, en torno a 0,74 para $k = 4$.

A esta diferencia metodológica se sumó una limitación técnica relevante: el cálculo de la matriz de distancias de Gower tiene una complejidad cuadrática (n x n), lo que hizo imposible ejecutar *k-medoids* sobre el *dataset* completo (aproximadamente 120.000 observaciones). De hecho, el entorno de Google Colab agotó la memoria incluso con muestras de 50.000 y 30.000 observaciones, siendo finalmente viable únicamente con 20.000 registros, lo que condiciona su uso como técnica principal.

Por tanto, la combinación de un mejor desempeño en términos de cohesión interna (*silhouette score* más elevado), la posibilidad de ajustarse sobre el *dataset* completo sin restricciones computacionales y la mayor facilidad para interpretar los grupos resultantes (solo dos clústeres bien diferenciados) justificó la elección de *k-prototypes* como técnica principal para la segmentación final del alumnado.

En paralelo, *k-medoids* se mantuvo como un análisis complementario aplicado sobre una muestra reducida.

3.6. Implementación del *clustering*

Este apartado describe el proceso técnico seguido para implementar los modelos de *clustering* aplicados en este estudio. Se desarrollaron dos enfoques (*k-prototypes* y *k-medoids* con distancia de Gower) con el objetivo de evaluar alternativas metodológicas y seleccionar la más adecuada para el *dataset*. Toda la implementación se realizó en Google Colab, lo que condicionó la ejecución de los modelos con mayor demanda computacional, especialmente en el caso de *k-medoids*.

3.6.1. Implementación de *k-prototypes*

Tras seleccionar *k-prototypes* como técnica principal (ver apartado 3.5), se implementó el algoritmo siguiendo un proceso estructurado que abarcó desde la preparación de los datos hasta la evaluación final de la calidad del agrupamiento.

El análisis partió del archivo generado tras el EDA y las transformaciones previas, que contenía únicamente las variables limpias y preparadas para su uso en *clustering*. Una vez importado el CSV, se verificó que cada variable tuviera el tipo correcto, ya que el algoritmo distingue explícitamente entre atributos numéricos (utilizados en la distancia euclídea) y categóricos (empleados en la disimilitud por *simple matching*).

Para garantizar la correcta ejecución, las variables numéricas se transformaron a tipo *float*, es decir, valores numéricos con decimales que permiten operar matemáticamente sobre ellos. Las variables categóricas se transformaron a tipo *string*, que corresponde a texto y permite identificar categorías o modalidades sin realizar operaciones numéricas sobre ellas.

Posteriormente, con el objetivo de determinar el número óptimo de clústeres, se extrajo una muestra aleatoria de 5.000 observaciones, fijando una semilla para garantizar la reproducibilidad. A partir de esta muestra se eliminaron algunas variables que, tras pruebas iniciales, no contribuían a mejorar la calidad del agrupamiento. Durante este proceso se realizaron múltiples iteraciones evaluando el impacto de incluir o excluir determinadas variables, siempre tomando como referencia el *silhouette score*.

Se identificó que la combinación con mejores resultados, y con mayor sentido analítico, incluía únicamente cuatro variables:

- edad de acceso estandarizada (edad_acceso_scaled)
- nota de admisión estandarizada (nota_admision_scaled)
- género del estudiante (cod_genero)
- condición de familia numerosa del estudiante (cod_familia_numerosa)

Este conjunto reducido permitió obtener grupos más homogéneos, modelos más estables y una mejor capacidad interpretativa de los clústeres finales. El resultado es una segmentación coherente y metodológicamente fundamentada, adecuada para el análisis posterior de perfiles estudiantiles.

Una vez definido el conjunto final de variables, se ajustaron modelos *k-prototypes* con valores de *k* entre 2 y 5. Para cada valor de *k*, se calcularon dos métricas complementarias:

- El *silhouette score*
- La función de coste del modelo (equivalente al método del codo)

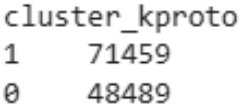
Ambas métricas se representaron en una gráfica de doble eje, permitiendo analizar visualmente el comportamiento del modelo conforme aumentaba el número de clústeres.

Los resultados obtenidos sobre la muestra mostraron que $k = 2$ proporcionaba la mejor combinación: *silhouette score* más elevado (en torno a 0,79), estabilidad del modelo, y grupos interpretables y bien diferenciados.

Una vez seleccionado $k = 2$, se aplicó *k-prototypes* sobre el *dataset* completo.

Tras ejecutar el algoritmo con $k = 2$, se obtuvo la distribución de observaciones entre los dos clústeres resultantes. Esta información se muestra en la Figura 18 y es relevante porque permite comprobar el equilibrio entre grupos y descartar posibles problemas como clústeres vacíos o extremadamente descompensados (lo que podría indicar un mal ajuste del modelo).

Figura 18: Distribución de observaciones por clúster (*k-prototypes*)



Fuente: Elaboración propia

Como se observa, el clúster 1 agrupa aproximadamente un 60% del alumnado, mientras que el clúster 0 concentra el 40% restante. Esta distribución es adecuada y, aunque no perfectamente simétrica, no presenta desequilibrios extremos que dificulten la interpretación o indiquen problemas en la segmentación.

Para evaluar la calidad del agrupamiento:

1. Se transformaron las variables categóricas a códigos numéricos.
2. Se calculó el *silhouette score* global y por clúster.
3. Se generó un gráfico de *silhouette* por clúster, útil para evaluar la cohesión interna.

Este procedimiento permitió validar tanto cuantitativa como visualmente la estructura de los dos clústeres resultantes, confirmando que *k-prototypes* produce una segmentación sólida y coherente para este *dataset*.

3.6.2. Implementación de *k-medoids* (distancia de Gower)

El modelo se aplicó sobre las mismas variables seleccionadas para *k-prototypes*: edad de acceso, nota de admisión, género y condición de familia numerosa. Nuevamente, se garantizó la correcta tipificación: las variables numéricas se transformaron a tipo *float*, y las categóricas a tipo *string*, tal como requiere la función gower_matrix.

Posteriormente, se generó la matriz de distancias de Gower y se aplicó el algoritmo *k-medoids* mediante la librería pyclustering, que permite trabajar directamente con matrices de distancia precomputadas. Debido a la elevada complejidad computacional de este tipo de matrices (dimensión n x n), la implementación se realizó sobre una muestra aleatoria de 20.000 observaciones, ya que tamaños superiores superaban la capacidad de memoria del entorno de ejecución (Google Colab).

Una vez calculada la matriz de distancias, el proceso de *clustering* siguió los pasos habituales del algoritmo PAM: selección inicial de *medoids*, asignación de observaciones según distancia mínima y actualización iterativa del *medoid* de cada clúster hasta convergencia. Tras la ejecución final con $k = 4$ (valor óptimo según el análisis previo), se obtuvieron las etiquetas de clúster para cada observación de la muestra.

Finalmente, para evaluar la calidad del agrupamiento:

1. Se calculó el *silhouette score* global.

2. Se obtuvo el *silhouette score* promedio por clúster, lo que permitió comparar la cohesión relativa de cada grupo.
3. Se generó el gráfico de *silhouette*, que facilita la interpretación visual de la estructura interna del modelo y el grado de solapamiento entre clústeres.

Dado que este modelo solo pudo aplicarse sobre una muestra reducida, sus resultados se interpretan como un análisis complementario, no como una base completamente representativa del conjunto total del alumnado.

3.7. Evaluación de la calidad de las agrupaciones

La calidad de los modelos de *clustering* se evaluó mediante el *silhouette score*, una métrica que permite valorar simultáneamente la cohesión interna de los grupos y la separación entre ellos. Los valores oscilan entre -1 y 1, siendo deseables valores cercanos a 1.

Se analizaron tanto el *silhouette score* global como el *silhouette score* promedio por clúster, complementados con el gráfico de siluetas, que permite una interpretación visual del comportamiento de cada grupo.

3.7.1. Evaluación del modelo *k-prototypes* (*dataset* completo)

Tras seleccionar $k = 2$ como número óptimo de clústeres y aplicar el modelo sobre la totalidad del *dataset* (119.948 registros), se evaluó la cohesión de la segmentación obtenida.

Tal como se muestra en la Figura 19, los resultados fueron los siguientes:

Figura 19: *Silhouette score* global y *silhouette score* medio por clúster ($k = 2$) en el modelo *k-prototypes* aplicado al *dataset* completo

```
--------------------------------------------------
 Silhouette score global: 0.5705
--------------------------------------------------

Silhouette score promedio por clúster:
cluster_kproto
0    0.5441
1    0.5883
```

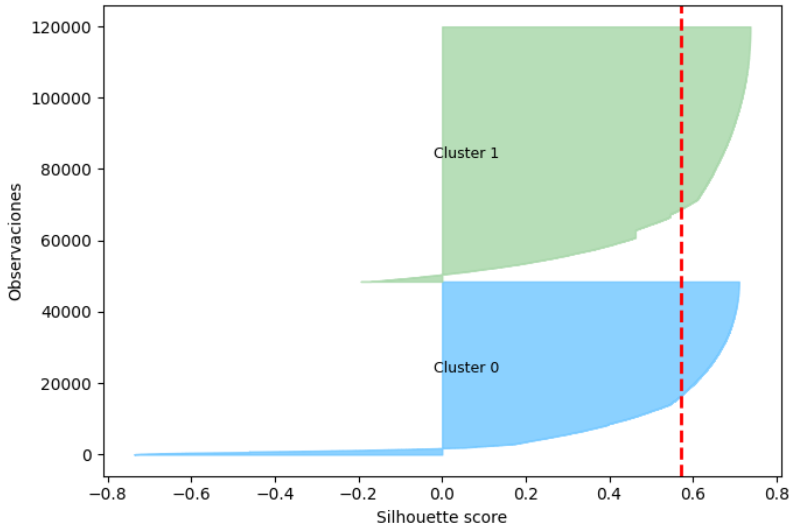
Fuente: Elaboración propia

Los clústeres aparecen numerados como 0 y 1 porque Python utiliza indexación basada en cero, de modo que la enumeración automática de grupos comienza en 0 en lugar de 1.

Estos valores indican una estructura de agrupamiento moderadamente cohesionada, en la que ambos clústeres presentan consistencia interna aceptable. El clúster 1 muestra, además, una compactación ligeramente superior.

Para complementar esta evaluación, se generó el gráfico de *silhouette* por clúster, que permite visualizar la distribución de los valores individuales de *silhouette* para cada observación. Este gráfico se muestra en la Figura 20.

Figura 20: Gráfico de *silhouette* por clúster (*k-prototypes*)



Fuente: Elaboración propia

En el gráfico observa un número reducido de observaciones con valores de *silhouette* negativos, lo que indica que estos casos se encuentran ligeramente más cerca del clúster alternativo que del propio. No obstante, su proporción es mínima respecto al total, por lo que no afecta de forma significativa a la validez global del agrupamiento.

El clúster 1 presenta valores de *silhouette* más elevados y homogéneos, lo que confirma su mayor cohesión. Por el contrario, el clúster 0 presenta una distribución más dispersa, aunque mantiene un comportamiento estable en la mayoría de observaciones. En conjunto, la estructura del modelo es consistente y refuerza la validez de utilizar dos clústeres como segmentación final del alumnado.

3.7.2. Evaluación del modelo *k-medoids* (muestra de 20.000 observaciones)

Tras ejecutar *k-medoids* con $k = 4$ sobre una muestra aleatoria de 20.000 observaciones, se evaluó la calidad del agrupamiento mediante el *silhouette score*.

Tal como se muestra en la Figura 21, los resultados fueron los siguientes:

Figura 21: *Silhouette score* global y *silhouette score* medio por clúster ($k = 4$) en el modelo *k-medoids*

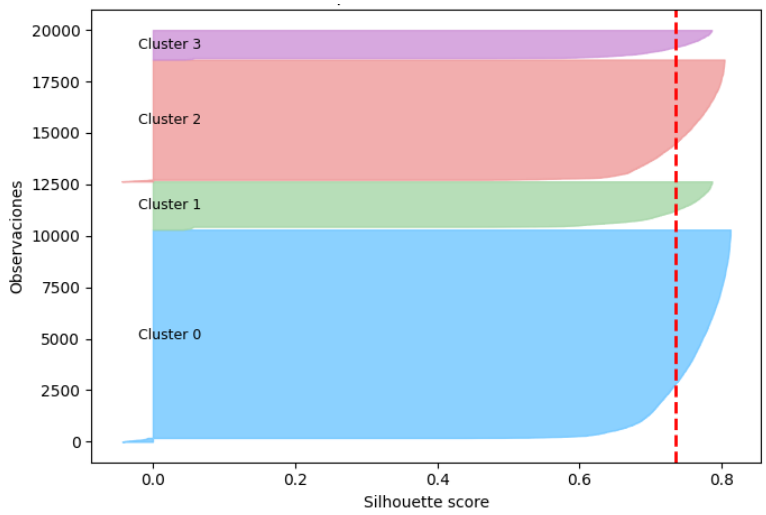
```
Silhouette global: 0.7358
cluster_kmedoids
0    0.746970
1    0.694755
2    0.741714
3    0.698072
```

Fuente: Elaboración propia

El modelo obtuvo un *silhouette score* global de 0,7358, un valor elevado que indica una buena cohesión interna y una separación razonable entre los cuatro clústeres. Asimismo, los resultados por clúster muestran que todos los clústeres presentan valores superiores a 0,69, lo que refleja una estructura relativamente estable y sin grupos excesivamente solapados. Los clústeres 0 y 2 destacan como los más cohesionados.

Al igual que con *k-prototypes*, para complementar esta evaluación, se generó el gráfico de *silhouette* por clúster. Este gráfico se muestra en la Figura 22.

Figura 22: Gráfico de *silhouette* por clúster (*k-medoids* con distancia de Gower)



Fuente: Elaboración propia

El gráfico permite observar visualmente la forma de cada clúster, el grado de solapamiento y la existencia de observaciones con bajo ajuste. A diferencia de *k-prototypes*, aquí apenas aparecen valores negativos, lo que sugiere que la distancia de Gower captura adecuadamente las relaciones entre observaciones en esta submuestra.

Aunque el valor del *silhouette score* es superior al obtenido por *k-prototypes*, este resultado no puede interpretarse como un mejor desempeño del modelo, por dos motivos fundamentales:

1. El modelo se evaluó sobre una muestra reducida. Se aplicó solo a 20.000 observaciones, mientras que *k-prototypes* se ejecutó sobre el *dataset* completo (119.948 registros). En muestras más pequeñas es habitual obtener valores de *silhouette* más altos, ya que la estructura interna tiende a ser más homogénea y la matriz de distancias se calcula con mayor precisión relativa. No implica necesariamente una segmentación mejor en el conjunto completo.
2. La métrica de Gower no reproduce exactamente la estructura sobre el *dataset* completo. Al no poder generar la matriz n x n para todos los estudiantes, la estructura real del alumnado no pudo evaluarse con esta técnica.

En consecuencia, los resultados de *k-medoids* deben considerarse exploratorios, pero no comparables directamente con el modelo principal aplicado sobre el *dataset* completo.

3.8. Análisis descriptivo de los clústeres

Una vez evaluada la calidad de las segmentaciones mediante el *silhouette score*, se realizó un análisis descriptivo de los clústeres identificados. Este apartado resume las características académicas y sociodemográficas de cada grupo, con el objetivo de interpretar la composición del alumnado y detectar patrones diferenciales.

Dado que *k-prototypes* constituye el modelo principal aplicado sobre la totalidad del *dataset*, su interpretación tiene carácter conclusivo. Por su parte, los resultados de *k-medoids*, calculados sobre una muestra reducida de 20.000 observaciones, se presentan como un análisis complementario, pero sin peso definitivo en las conclusiones finales.

3.8.1. Análisis de los clústeres obtenidos mediante *k-prototypes*

El modelo *k-prototypes* con $k = 2$ identificó dos perfiles claramente diferenciados de

estudiantes. Las diferencias se observan tanto en las variables numéricas (edad y nota de acceso) como en las categóricas (características sociodemográficas y académicas).

Variables numéricas

Las diferencias más significativas se observan en la edad de acceso y la nota de admisión (Figura 23):

- El clúster 1 está compuesto por estudiantes que acceden al sistema universitario a edades tempranas (media de 18,73 años) y que presentan un rendimiento significativamente superior, con una nota media de admisión de 11,62 puntos.
- El clúster 0, por el contrario, agrupa a estudiantes que acceden de forma más tardía (20,36 años de media) y que presentan una nota de admisión mucho más baja (8,12 puntos).

Esta brecha constituye la separación estructural más relevante del *dataset*, y es la que explica en mayor medida la división final entre los dos grupos.

Figura 23: Distribución de las variables numéricas por clúster (*k-prototypes*)

cluster_kproto	edad_acceso	nota_admision
0	20.363051	8.122927
1	18.728852	11.620583

Fuente: Elaboración propia

Variables categóricas

El análisis de las variables categóricas permite profundizar en los matices que caracterizan a cada clúster. Tal como se observa en la Figura 24, algunas variables presentan distribuciones muy similares entre grupos, mientras que otras muestran diferencias más estructurales. Es importante señalar que la categoría -99 corresponde a los casos “Sin información”, tal como se definió durante el proceso de recodificación.

En primer lugar, la naturaleza del centro de secundaria presenta distribuciones prácticamente idénticas entre ambos grupos, sin diferencias relevantes entre estudiantes de centros públicos, concertados o privados. Aunque el clúster 1 muestra una ligera mayor presencia de estudiantes provenientes de centros privados, esta variación es reducida y no

desempeña un papel estructural en la formación de los clústeres.

En cuanto a la localización del centro de procedencia, sí se observa una diferencia más marcada: el 73% de los estudiantes del clúster 0 cursó sus estudios previos en centros ubicados en la Comunidad de Madrid, frente al 63% en el clúster 1.

Respecto al género, ambos grupos están compuestos mayoritariamente por mujeres, aunque esta proporción es ligeramente superior en el clúster 1 (65% frente a 60%). No obstante, la variable no parece ser un elemento determinante en la separación entre clústeres.

La condición de familia numerosa muestra distribuciones prácticamente idénticas en ambos grupos, por lo que no contribuye de manera diferencial a la separación entre perfiles.

Sí emergen diferencias más claras en la especialidad de Bachillerato o formación previa. El clúster 0 presenta una mayor presencia de itinerarios de Humanidades y Ciencias Sociales, mientras que el clúster 1 concentra un porcentaje superior de estudiantes provenientes del itinerario de Ciencias, lo cual resulta coherente con las diferencias observadas en nota de admisión, puesto que las carreras de Ciencias suelen tener una nota de admisión más alta.

A nivel universitario, también existen diferencias en la rama de la titulación cursada. En el clúster 0 predominan las titulaciones del ámbito de Ciencias Sociales y Jurídicas, junto con un peso mayor de Educación, Humanidades y Artes. En el clúster 1, aunque Ciencias Sociales continúa siendo la rama más representada, aumenta de forma significativa la presencia de titulaciones vinculadas a Ciencias de la Salud y STEM, lo que nuevamente guarda coherencia con su mayor rendimiento previo.

Finalmente, el nivel socioeconómico familiar refleja una distribución ligeramente diferenciada: el clúster 0 concentra un mayor número de estudiantes de nivel socioeconómico medio, mientras que en el clúster 1 aumenta la proporción de estudiantes pertenecientes a niveles socioeconómicos altos. Esta tendencia no es lo suficientemente marcada como para considerarse un factor determinante, pero sí apunta a un patrón moderado en la distribución de ambos grupos.

Figura 24: Distribución de las variables categóricas por clúster (*k-prototypes*)

Distribución de 'cod_naturaleza_centro_sec' por clúster:

cluster_kproto	cod_naturaleza_centro_sec	
0	1	45.28
	2	28.14
	-99	23.27
	3	3.32
1	1	46.22
	2	32.06
	-99	18.85
	3	2.87

Distribución de 'centro_en_madrid' por clúster:

cluster_kproto	centro_en_madrid	
0	1	73.07
	0	15.92
	-99	11.01
1	1	62.76
	0	26.76
	-99	10.48

Distribución de 'cod_genero' por clúster:

cluster_kproto	cod_genero	
0	1	60.02
	0	39.98
1	1	65.17
	0	34.83

Distribución de 'cod_familia_numerosa' por clúster:

cluster_kproto	cod_familia_numerosa	
0	1	80.31
	2	17.11
	3	2.58
1	1	78.67
	2	18.94
	3	2.39

Distribución de 'especialidad_rama' por clúster:

cluster_kproto	especialidad_rama	
0	3	45.38
	2	22.66
	-99	20.66
	4	8.29
	1	3.01
1	2	36.71
	3	32.08
	-99	23.97
	4	4.21
	1	3.03

Distribución de 'rama_titulacion' por clúster:

cluster_kproto	rama_titulacion	
0	2	41.01
	6	26.75
	1	18.94
	3	5.70
	5	4.44
	4	3.16
1	2	35.81
	3	20.11
	6	15.64
	1	11.58
	4	10.46
	5	6.40

Distribución de 'nivel_socioeco_familia' por clúster:

cluster_kproto	nivel_socioeco_familia	
0	2	43.86
	3	39.07
	1	13.16
	-99	3.91
1	3	49.02
	2	38.72
	1	9.65
	-99	2.61

Fuente: Elaboración propia

3.8.2. Análisis de los clústeres obtenidos mediante *k-medoids*

El análisis realizado con *k-medoids* sobre una muestra de 20.000 estudiantes ha permitido identificar cuatro clústeres adicionales. Dado que esta técnica no pudo aplicarse sobre el conjunto completo del *dataset*, la interpretación de estos grupos debe considerarse exploratoria. Por este motivo, y al tratarse de una aproximación aplicada únicamente a una muestra reducida, no se incluyen representaciones visuales detalladas de las distribuciones por clúster.

Variables numéricas

En términos de variables numéricas, los cuatro clústeres muestran medias de edad de acceso muy similares, oscilando entre 19,22 y 19,64 años, así como notas de admisión en torno a los 10 puntos. Estas diferencias menos pronunciadas sugieren que, en esta muestra parcial, la distancia de Gower detecta variabilidad más fina y menos dependiente del rendimiento académico previo.

Variables categóricas

En cambio, la variable género se muestra especialmente estructurante para este modelo. Los clústeres 0 y 1 están compuestos exclusivamente por mujeres, mientras que los clústeres 2 y 3 incluyen únicamente hombres. Este patrón indica que la métrica de Gower otorga un peso significativo a la proximidad entre categorías, captando con claridad esta dimensión dentro de la muestra.

La condición de familia numerosa introduce diferencias adicionales: los clústeres 0, 1 y 2 presentan mayor presencia de estudiantes que no pertenecen a familia numerosa, mientras que el clúster 3 solo tiene estudiantes que pertenecen a familia numerosa, principalmente a la condición de familia numerosa general.

A su vez, la especialidad de Bachillerato o formación previa contribuye a diversificar los perfiles: el clúster 0 presenta mayor peso de itinerarios de Humanidades y Ciencias Sociales, mientras que el clúster 3 concentra más estudiantes del ámbito de Ciencias. Los clústeres 1 y 2 muestran distribuciones intermedias.

El nivel socioeconómico familiar presenta diferencias suaves, aunque se aprecia un mayor peso del nivel 3 (alto) en los clústeres 1, 2 y 3, mientras que el clúster 0 concentra más estudiantes de nivel medio.

Finalmente, variables como la localización del centro de secundaria, la naturaleza del centro y rama de titulaciones universitarias presentan distribuciones muy similares entre los cuatro grupos, lo que indica que no desempeñan un papel diferenciador en esta segmentación exploratoria.

4. RESULTADOS E IMPLICACIONES PRÁCTICAS

4.1. Perfiles de estudiantes identificados

La aplicación del modelo *k-prototypes* con $k = 2$ sobre el conjunto completo de estudiantes permitió identificar dos perfiles claramente diferenciados, cuya caracterización combina variables académicas, formativas y sociodemográficas. Estos perfiles constituyen la base interpretativa del comportamiento del alumnado y sirven como fundamento para las recomendaciones de gestión académica expuestas en el apartado siguiente.

Perfil 1 – Estudiantes de acceso temprano y alto rendimiento

Este grupo, que muestra la mayor cohesión interna según el *silhouette score*, está caracterizado por un conjunto de rasgos que apuntan a trayectorias académicas lineales y un rendimiento previo elevado.

En términos académicos, los estudiantes de este perfil acceden a la universidad a edades tempranas (media de 18,73 años) y presentan notas de admisión notablemente altas (11,62 puntos de media). Este rendimiento previo elevado se corresponde con trayectorias formativas más estables durante la etapa preuniversitaria, generalmente asociadas a progresiones educativas sin interrupciones.

Desde el punto de vista formativo, este perfil presenta un mayor peso de estudiantes procedentes de itinerarios de Ciencias y Tecnología en Bachillerato, un patrón coherente con sus notas medias más altas. Aunque las titulaciones universitarias de la rama de Ciencias Sociales y Jurídicas siguen siendo mayoritarias en términos absolutos, este perfil muestra una presencia relativamente superior de grados vinculados a Ciencias de la Salud y a áreas STEM en comparación con el otro perfil.

A nivel sociodemográfico, este grupo está ligeramente más feminizado (65% mujeres) y presenta una mayor proporción de estudiantes pertenecientes a niveles socioeconómicos altos. En cuanto a la procedencia geográfica, el 63% realizó sus estudios previos en centros ubicados en la Comunidad de Madrid, una proporción menor que en el perfil 0, lo que sugiere un patrón territorial algo más distribuido.

En conjunto, este perfil representa al estudiante “tradicional”: acceso directo tras

Bachillerato, rendimiento destacado y progresión académica relativamente homogénea.

Perfil 0 – Estudiantes de acceso más tardío y menor rendimiento previo

El segundo grupo presenta una trayectoria académica significativamente distinta. Los estudiantes acceden a la universidad a edades más elevadas (20,36 años de media) y cuentan con notas de admisión considerablemente más bajas (8,12 puntos). Esto sugiere rutas formativas menos lineales, posibles cambios de itinerario, periodos de preparación adicionales o decisiones vocacionales tomadas más tarde.

En cuanto a su trayectoria previa, este perfil concentra un mayor número de estudiantes provenientes de itinerarios de Humanidades y Ciencias Sociales, lo que contribuye a explicar, en parte, la ausencia de notas de admisión tan elevadas como en el perfil 1. Dentro de la Universidad, predominan las titulaciones de Ciencias Sociales y Jurídicas, seguida de grados vinculados a Educación, Humanidades y Artes.

Desde el punto de vista sociodemográfico, el perfil se distribuye mayoritariamente entre niveles socioeconómicos medios. Aunque las diferencias con respecto al perfil de alto rendimiento no son extremas, sí se aprecia una ligera mayor concentración de hogares con menor nivel socioeconómico. Por otra parte, este grupo presenta un porcentaje significativamente superior de estudiantes cuya formación previa se desarrolló en centros ubicados en la Comunidad de Madrid (73%).

El perfil 0 representa, por tanto, a estudiantes con una trayectoria más heterogénea, con mayor dispersión académica previa y potencialmente con necesidades diferenciadas en términos de acompañamiento y apoyo institucional.

En conjunto, la segmentación obtenida mediante *k-prototypes* pone de manifiesto la existencia de dos perfiles claramente diferenciados dentro del alumnado: un grupo mayoritariamente joven, con alto rendimiento previo y trayectorias académicas más lineales, frente a otro de acceso más tardío, menor nota de admisión y mayor heterogeneidad formativa. Esta distinción constituye la base para el diseño de estrategias de apoyo académico y planificación institucional que permitan responder de forma más ajustada a las necesidades de cada perfil, tal y como se desarrolla en el apartado 4.2.

4.2. Recomendaciones para la gestión académica basadas en los clústeres

A partir de los dos perfiles identificados mediante el modelo *k-prototypes*, se proponen diversas estrategias orientadas a mejorar la adaptación, el rendimiento y la permanencia del alumnado. Las recomendaciones se articulan de manera diferenciada para cada grupo, con el fin de responder a sus características específicas, pero también incluyen una serie de medidas transversales aplicables al conjunto de estudiantes.

4.2.1. Estudiantes de acceso temprano y alto rendimiento (Grupo 1)

El primer grupo está compuesto por estudiantes que acceden a la universidad a edades tempranas y con notas de admisión muy elevadas. Su trayectoria refleja un progreso académico lineal y un desempeño previo consistente. Además, muestran una mayor orientación hacia itinerarios científicos y titulaciones STEM o de Ciencias de la Salud. Este tipo de alumnado suele mostrar alta motivación y capacidad de trabajo autónomo.

Para aprovechar plenamente su potencial y favorecer su desarrollo, las universidades pueden implementar estrategias orientadas a la excelencia académica y a la ampliación de oportunidades formativas. Entre las medidas más adecuadas destacan los programas de mentoría avanzada, la participación en actividades de investigación temprana y la oferta de seminarios especializados o cursos de profundización en áreas de interés. Estas iniciativas permiten reforzar su progresión académica y fomentar la retención del talento.

Asimismo, dado que este perfil presenta mayor presencia de estudiantes provenientes de centros de secundaria fuera de la Comunidad de Madrid, puede resultar beneficioso reforzar la orientación inicial sobre opciones de alojamiento, residencias universitarias o colegios mayores, especialmente para quienes se incorporan por primera vez a la ciudad. Adicionalmente, la implementación de becas de excelencia u otros incentivos al rendimiento contribuiría a motivar a los estudiantes a seguir obteniendo buenos resultados en la universidad.

4.2.2. Estudiantes de acceso más tardío y menor rendimiento previo (Grupo 0)

Este grupo está formado por estudiantes que acceden a la universidad a edades más avanzadas y que presentan una nota de admisión significativamente inferior a la del perfil

anterior. Sus trayectorias educativas tienden a ser más heterogéneas, con posibles cambios de itinerario o periodos de preparación adicionales, lo que se traduce en necesidades académicas diferenciadas.

A la luz de estas características, se proponen tres líneas de actuación prioritarias:

1. Programas de refuerzo académico en primer curso: especialmente recomendables en asignaturas con elevada tasa de repetición o con un componente matemático-cuantitativo relevante (frecuentes en grados de ADE, Economía o titulaciones STEM). Estos cursos permitirían nivelar competencias iniciales y reducir la probabilidad de abandono temprano.
2. Tutorías académicas individualizadas al inicio del grado: un seguimiento más estrecho durante el primer semestre ayudaría a identificar dificultades tempranas y a orientar al estudiante hacia hábitos de estudio eficaces, reduciendo la incertidumbre asociada a trayectorias educativas no lineales.
3. Flexibilización en la organización académica. Dado que este grupo presenta una edad media más elevada, puede ser más probable que compatibilicen estudios con otras responsabilidades. La oferta de turnos de tarde, modalidades híbridas o recursos docentes accesibles en formato virtual facilitaría su continuidad y rendimiento académico.

En conjunto, estas medidas buscan favorecer la adaptación al entorno universitario, reducir las barreras derivadas de trayectorias previas heterogéneas y mejorar la probabilidad de éxito académico durante las primeras etapas del grado.

4.2.3. Recomendaciones transversales para toda la comunidad estudiantil

Además de las estrategias específicas para cada perfil, existen diversas medidas de carácter general que pueden reforzar la experiencia académica del conjunto del alumnado:

- Programas de bienestar psicológico, esenciales para mejorar el rendimiento y la adaptación.
- Seguimiento académico basado en datos, integrando analítica de aprendizaje para identificar patrones de riesgo.

- Comunicación personalizada según perfil, adaptando campañas informativas a las características de cada grupo.
- Evaluación continua de la planificación docente, asegurando que la oferta formativa es coherente con las necesidades reales del alumnado.

Gracias al uso de *clustering*, la universidad puede diseñar estrategias basadas en evidencia para personalizar la atención, mejorar el rendimiento global y optimizar la planificación académica. Estos resultados permiten orientar la toma de decisiones institucionales hacia modelos educativos más adaptativos y centrados en el estudiante.

4.3. Implicaciones para la planificación de recursos universitarios

El análisis exploratorio del *dataset* ha permitido identificar varias tendencias estructurales en la composición y evolución del alumnado universitario durante el periodo analizado. Estas tendencias tienen implicaciones directas para la planificación de recursos, la organización docente y la toma de decisiones estratégicas de la institución. A continuación, se sintetizan los patrones observados y sus principales implicaciones para la gestión universitaria.

4.3.1. Descenso progresivo de las matriculaciones

Los datos muestran una reducción gradual del número total de matrículas desde el curso 2018/19, con un mínimo en 2022/23. Aunque el curso 2023/24 presenta una ligera recuperación, la tendencia general es descendente.

Esta evolución sugiere la necesidad de ajustar el dimensionamiento de grupos, evitar sobredimensionar la oferta docente y revisar la planificación de infraestructuras. La reducción de estudiantes en determinadas titulaciones puede exigir una redistribución de recursos hacia áreas con mayor demanda o un rediseño del mapa de titulaciones.

4.3.2. Crecimiento de la demanda en titulaciones STEM y Ciencias de la Salud

La proporción de estudiantes en Ciencias, Ciencias de la Salud e Ingeniería y Arquitectura aumenta de forma continuada en el periodo Post-COVID, mientras que las ramas de Humanidades, Artes, Ciencias Sociales y Jurídicas y Educación experimentan descensos moderados.

Las titulaciones STEM y de Salud requieren recursos docentes y materiales más intensivos. El aumento de la demanda en estas áreas obliga a planificar inversiones en laboratorios, equipamiento tecnológico y profesorado especializado, así como a revisar la capacidad de acceso de los grados con alta presión de demanda. De manera complementaria, la disminución de estudiantes en otras ramas puede requerir revisar la viabilidad de titulaciones con descenso sostenido de demanda para evitar infrautilización de recursos.

4.3.3. Incremento de la proporción de estudiantes procedentes de centros de secundaria ubicados en la Comunidad de Madrid

La proporción de alumnado cuyo último centro educativo se encontraba en la Comunidad de Madrid pasa del 72,6% en Pre-COVID al 76,5% en Post-COVID.

La institución se consolida como un centro universitario predominantemente local, lo que sugiere la conveniencia de reforzar los vínculos con institutos de la Comunidad de Madrid mediante programas de orientación preuniversitaria, actividades de captación y convenios formativos. A la vez, la reducción relativa de estudiantes de otras comunidades autónomas puede afectar a la demanda de alojamiento universitario, como residencias y colegios mayores y a la necesidad de programas de acogida específicos.

4.3.4. Incremento del nivel socioeconómico del alumnado

El nivel socioeconómico alto (3) aumenta su peso relativo en los periodos COVID y Post-COVID, mientras que la proporción de estudiantes de nivel socioeconómico bajo (1) disminuye.

Este cambio requiere revisar las políticas de becas y ayudas al estudio para garantizar la equidad en el acceso, especialmente para estudiantes con menos recursos. Asimismo, un alumnado con mayor capacidad económica puede influir en la demanda de servicios como programas internacionales, alojamiento en residencias o actividades extracurriculares, aspectos que deben considerarse en la planificación institucional.

4.3.5. Aumento generalizado de la nota de admisión

La nota media de acceso aumenta de manera consistente, pasando de 9,73 en Pre-COVID a 10,62 en Post-COVID.

Un mayor nivel académico previo puede requerir una revisión del grado de exigencia de los primeros cursos y la implementación de programas de excelencia o itinerarios avanzados para estudiantes de alto rendimiento. Al mismo tiempo, es esencial garantizar medidas de apoyo para estudiantes con notas de acceso más bajas, a fin de evitar brechas de rendimiento y promover su permanencia.

Las tendencias detectadas en el EDA ofrecen una visión clara de cómo está evolucionando el perfil del alumnado universitario. El incremento del interés por titulaciones STEM y de Salud, la concentración territorial del alumnado en la Comunidad de Madrid, la elevación del nivel socioeconómico alto y el aumento del rendimiento previo constituyen señales relevantes para la planificación futura. Integradas con los resultados del *clustering*, estas conclusiones permiten orientar la toma de decisiones hacia una asignación de recursos más eficiente, una oferta docente mejor adaptada y un diseño de servicios universitarios más acorde con las características reales del estudiantado.

5. CONCLUSIONES

5.1. Principales conclusiones

El presente trabajo demuestra que la aplicación de técnicas de análisis de datos propias del ámbito de *Business Analytics* puede aportar un valor significativo a la comprensión del alumnado universitario y a la toma de decisiones institucionales. Más allá de describir patrones, el análisis realizado permite transformar un volumen elevado de datos heterogéneos en información útil para orientar la planificación académica.

En primer lugar, el estudio exploratorio identificó tendencias estructurales que condicionan el contexto actual de la universidad: un aumento de la demanda en titulaciones científicas y sanitarias, la llegada de estudiantes con un rendimiento previo más elevado, una composición socioeconómica progresivamente más alta y una mayor concentración de alumnado procedente de centros madrileños. Estas dinámicas reflejan un entorno cada vez más exigente y heterogéneo, en el que comprender la evolución del alumnado resulta esencial para anticipar necesidades y definir prioridades institucionales.

La segmentación mediante *k-prototypes* constituye la aportación central del trabajo. La identificación de dos perfiles diferenciados, uno de acceso temprano y alto rendimiento, y otro caracterizado por trayectorias más tardías y menores notas de admisión, proporciona un marco claro para adaptar estrategias académicas de manera más precisa. Esta clasificación permite superar enfoques generalistas y avanzar hacia una atención más ajustada a las necesidades reales de cada grupo de estudiantes.

Las recomendaciones derivadas del análisis muestran cómo esta segmentación puede traducirse en actuaciones concretas: becas de excelencia y oportunidades formativas avanzadas para los estudiantes con mayor potencial; y, para quienes presentan trayectorias más complejas, medidas de apoyo temprano, tutorías individualizadas y mayor flexibilidad horaria. Estas propuestas, junto con actuaciones transversales en ámbitos como el bienestar psicológico, la orientación académica o la analítica de aprendizaje, refuerzan la utilidad práctica del modelo.

Finalmente, el análisis exploratorio ofrece una perspectiva complementaria de carácter estratégico. Las tendencias detectadas, en especial el crecimiento de la demanda en

titulaciones STEM y relacionadas con la salud y los cambios en el perfil socioeconómico y geográfico del alumnado, aportan información relevante para la planificación de recursos docentes, la adaptación de infraestructuras y la definición de políticas de apoyo institucional. Esta visión macro contribuye a contextualizar la segmentación obtenida y a comprender cómo evoluciona el sistema universitario en su conjunto.

En conjunto, el trabajo pone de manifiesto cómo una aproximación analítica rigurosa puede convertirse en una herramienta estratégica para la educación superior. Su valor añadido reside no solo en la segmentación obtenida, sino en demostrar cómo un enfoque basado en datos puede contribuir a modelos de gestión académica más adaptativos, personalizados y alineados con las necesidades reales de los estudiantes y de la institución.

5.2. Limitaciones del estudio

A pesar de los resultados obtenidos y del valor añadido que aporta el uso de técnicas analíticas avanzadas aplicadas a datos educativos, el estudio presenta una serie de limitaciones que conviene tener en cuenta a la hora de interpretar sus conclusiones.

En primer lugar, el uso del algoritmo *k-medoids* con distancia de Gower estuvo condicionado por restricciones computacionales. La construcción de la matriz completa de distancias (de tamaño n x n) requiere un volumen elevado de memoria RAM, lo que impidió aplicar esta técnica al conjunto completo de 119.948 matrículas. Esto obligó a trabajar con una muestra aleatoria de 20.000 observaciones, por lo que los resultados de *k-medoids* deben interpretarse como exploratorios y no como una segmentación alternativa comparable a la obtenida mediante *k-prototypes*.

En segundo lugar, aunque la base de datos utilizada cuenta con un tamaño considerable y proporciona información sociodemográfica y académica relevante, presenta limitaciones propias de los datos anonimizados. La ausencia de un identificador individual impide integrar distintas bases de datos de la universidad y realizar análisis longitudinales o multidimensionales que incorporen, por ejemplo, información sobre rendimiento a lo largo del grado, participación en actividades universitarias, situación económica detallada o indicadores de bienestar estudiantil. Estas dimensiones, que no se encuentran disponibles en UniversiDATA, podrían aportar una caracterización más completa de los

perfiles del alumnado y mejorar la precisión de las recomendaciones propuestas.

En conjunto, estas limitaciones no invalidan los resultados obtenidos, pero sí acotan su alcance. Futuras investigaciones que cuenten con mayor capacidad computacional o con bases de datos más integradas podrían ampliar el análisis realizado, profundizando en la segmentación del alumnado y en su aplicación práctica a la gestión universitaria.

5.3. Futuras líneas de investigación

El presente trabajo abre diversas posibilidades de ampliación y profundización que permitirían avanzar hacia una comprensión más completa y dinámica del alumnado universitario. Entre las líneas de investigación más prometedoras destacan las siguientes:

En primer lugar, sería de gran interés desarrollar modelos de predicción basados en técnicas de *machine learning* que permitan anticipar la evolución de la demanda por ramas de conocimiento y titulaciones concretas. La incorporación de modelos de series temporales o algoritmos supervisados (como *Random Forest* o *Gradient Boosting*) contribuiría a proyectar escenarios futuros y apoyar la planificación estratégica de oferta académica y recursos docentes.

En segundo lugar, la ampliación del análisis hacia modelos predictivos de rendimiento y riesgo de abandono constituiría un avance relevante desde la perspectiva de la gestión universitaria. Algoritmos supervisados aplicados a datos académicos longitudinales permitirían identificar tempranamente a los estudiantes con mayor probabilidad de enfrentar dificultades, facilitando intervenciones preventivas más eficaces.

Asimismo, sería pertinente extender el enfoque de *clustering* a subconjuntos específicos del alumnado, por ejemplo, segmentaciones diferenciadas por rama de estudio, tipo de titulación o modalidad académica, o incluso analizar la evolución de los perfiles a lo largo del tiempo. Estos enfoques permitirían capturar con mayor detalle la heterogeneidad interna dentro de cada ámbito disciplinar y estudiar trayectorias académicas longitudinales.

Otra línea de investigación especialmente relevante consistiría en ampliar el análisis comparativo a diferentes niveles. En primer lugar, podría aplicarse el mismo procedimiento analítico a otras universidades para examinar similitudes y diferencias en la composición y evolución del alumnado. En segundo lugar, resultaría de interés realizar

comparaciones sistemáticas entre universidades públicas y privadas, lo que permitiría evaluar divergencias en perfiles de acceso, rendimiento y composición socioeconómica. Por último, sería pertinente analizar diferencias territoriales mediante comparaciones entre universidades de distintas comunidades autónomas, incorporando así la dimensión regional en el estudio del acceso y la diversidad del alumnado.

Finalmente, futuras investigaciones podrían beneficiarse de un enriquecimiento del *dataset* mediante la integración de fuentes externas o institucionales adicionales. La disponibilidad de identificadores individuales permitiría combinar la información académica con datos sobre participación universitaria, bienestar, motivación o situación socioeconómica, ampliando notablemente el alcance interpretativo y la capacidad de personalización de las estrategias diseñadas.

En conjunto, estas líneas de investigación apuntan a reforzar la aplicación de técnicas analíticas avanzadas en el ámbito universitario, potenciando el desarrollo de modelos de gestión más predictivos, adaptativos y centrados en las necesidades reales del alumnado.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Carlos Martín de los Santos, estudiante de ADE y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Segmentación de estudiantes universitarios mediante *clustering* para la planificación académica", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
3. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
4. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 03/12/2025

Firma: Carlos Martín de los Santos Antoñanzas

6. BIBLIOGRAFÍA

Aristizabal F., J. A. (2016). Analítica de datos de aprendizaje (ADA) y gestión educativa. *Revista Gestión de la Educación*, *6*(2), 149-168. https://doi.org/10.15517/rge.v1i2.25499

Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3-16. https://doi.org/10.5281/zenodo.3554657

Bektas, A., & Schumann, R. (2019). How to Optimize Gower Distance Weights for the k-Medoids Clustering Algorithm to Obtain Mobility Profiles of the Swiss Population. *6th Swiss Conference on Data Science (SDS 2019)*, 51-56. https://doi.org/10.1109/SDS.2019.000-8

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief*. Office of Educational Technology, US Department of Education. https://eric.ed.gov/?id=ED611199

Cabrera, L., Bethencourt, J. T., Álvarez Pérez, P., & González Afonso, M. (2006). El problema del abandono de los estudios universitarios. *RELIEVE*, *12*(2), 171-203. https://doi.org/10.7203/relieve.12.2.4226

Calvet Liñán, L., & Juan Pérez, Á. A. (2015). Educational Data Mining and Learning Analytics: Differences, similarities, and time evolution. *RUSC. Universities and Knowledge Society Journal*, *12*(3), 98-112. https://doi.org/10.7238/rusc.v12i3.2515

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, *8*, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

D'Orazio, M. (2024). *Gower's similarity coefficients with automatic weight selection*. Italian National Institute of Statistics - Istat. http://arxiv.org/abs/2401.17041

Duval, E. (2011). Attention please! Learning analytics for visualization and recommendation. *Proceedings of LAK 11: 1st International Conference on Learning Analytics and Knowledge*, 9-17. https://doi.org/10.1145/2090116.2090118

Fundación CYD. (2025). *Informe CYD 2024*. https://www.fundacioncyd.org/publicaciones-cyd/informe-cyd-2024/

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857-871. https://doi.org/10.2307/2528823

Hao, J., Gan, J., & Zhu, L. (2022). MOOC performance prediction and personal performance improvement via Bayesian network. *Education and Information*

Technologies, *27*(5), 7303-7326. https://doi.org/10.1007/s10639-022-10926-8

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *2*, 283-304. https://doi.org/10.1023/A:1009769707641

Hurtado, S., Suaza, C., Aguilar, J., & Suescún, E. (2021). Ciclo Autonómico de Análisis de Datos aplicado en la Planificación Académica de instituciones educativas. *RISTI: Revista Ibérica de Sistemas e Tecnologias de Informaçao*, (Extra 41), 193-206.

INE. Instituto Nacional de Estadística. (2025). *MNP: Estadística de nacimientos*. https://www.ine.es/consul/serie.do?d=true&s=MNP162&c=2&

Isea, R., Ojeda, V., Fernandez, J., Gutierrez, A., & Salazar, V. (s. f.). Coeficiente V de Cramer (V). *Universidad Central de Venezuela - Facultad de Humanidades y Educación*. https://es.scribd.com/document/624898606/Coeficiente-V-de-Cramer-interpretacion

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis*. DOI: 10.2307/2532178

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6), 90-95.

Kurday, M., & Vladova, G. (2025). Learning analytics and educational data mining applications: Bibliometric and ChatGPT-based analysis of research publications from 2014 to 2023. *Journal of Computers in Education*, 1-36. https://doi.org/10.1007/s40692-025-00362-1

Lacuesta, A., Martínez-Matute, M., Sainz, J., & Sanz, I. (2024). Desajustes entre demanda y oferta de titulaciones en las universidades públicas presenciales. *Papeles de Economía Española*, (180), 98-113.

Lang, C., Siemens, G., Wise, A. F., Gašević, D., & Merceron, A. (2022). *The Handbook of Learning Analytics* (2.ª ed.). SOLAR. https://doi.org/10.18608/hla22

Madhulatha, T. S. (2012). An Overview on Clustering Methods. *IOSR Journal of Engineering*, *2*(4), 719-725. https://doi.org/10.48550/arXiv.1205.1117

Madhuri, R., Murty, M. R., Murthy, J. V. R., Reddy, P. V. G. D. P., & Satapathy, S. C. (2014). Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms. *ICT and Critical Infrastructure: Proceedings of the 48th Annual*

Convention of Computer Society of India- Vol II, 137-144. https://doi.org/10.1007/978-3-319-03095-1_15

OECD. (2021). *The State of Global Education: 18 Months into the Pandemic*. OECD Publishing. https://doi.org/10.1787/1a23bb23-en

Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., & Tessore, J. P. (2016). Tratamiento Masivo de Datos Utilizando Técnicas de Machine Learning. *XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina)*, 131-134. https://sedici.unlp.edu.ar/handle/10915/52838

Siemens, G., & Baker, R. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *Proceedings of LAK 12: 2nd International Conference on Learning Analytics and Knowledge*, 252-254. https://doi.org/10.1145/2330601.2330661

Subirats, L., Palacios Corral, A., Pérez-Ruiz, S., Fort, S., & Gómez Moñivas, S. (2023). Temporal analysis of academic performance in higher education before, during and after COVID-19 confinement using artificial intelligence. *PLOS ONE*, *18*(2), e0282306. https://doi.org/10.1371/journal.pone.0282306

UniversiDATA. (2025). *UCM-Acceso | UniversiDATA*. https://www.universidata.es/datasets/ucm-acceso

Vendrell-Morancho, M., Rodríguez-Mantilla, J. M., & Fernández-Díaz, M. J. (2024). Identificación de Perfiles de Pensamiento Crítico entre el Estudiantado Universitario Español: Un Análisis de Conglomerados con el Método K-Medias. *RELIEVE*, *30*(2). https://doi.org/10.30827/relieve.v30i2.28208

Villarroel-Henríquez, V., & Gallardo-Aguayo, C. (2025). Gestión del Proceso Educativo utilizando Analíticas del Aprendizaje y Big Data en la Universidad. *Revista de Investigación en Educación*, *23*(1), 76-92. https://doi.org/10.35869/reined.v23i1.6110