UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escuela Técnica Superior de Ingeniería Informática
Universidad Politécnica de Valencia

Análisis del abandono universitario a partir de datos académicos y de actividad

TRABAJO FIN DE GRADO

Grado en Ciencia de Datos

Autor: Pablo Parrilla Cañadas
Tutor: Andrea Conchado Peiró
Cotutor: José Vicente Benlloch Dualde

Curso 2025-2026

Resum

????

Paraules clau: ????, ?????????, ????, ??????????????

Resumen

????

Palabras clave: ?????, ???, ??????????????

Abstract

????

Key words: ?????, ????? ?????, ?????????????

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

1.1 Motivación

La educación es uno de los pilares básicos de nuestra sociedad, siendo fundamental para el cambio social y económico y para el avance tanto personal como colectivo. La importancia de esta materia es clave a lo largo de nuestra vida. Por lo tanto, cuando muchas personas dejan sus estudios universitarios, es un gran fallo del sistema que trae consecuencias tanto en el ámbito personal como en las instituciones.

Hoy en día, las carreras tecnológicas, como la Ingeniería Informática, son claves para el avance de la sociedad. Sin embargo, la alta tasa de abandono en estas carreras es preocupante y es necesario prestar atención: hay que entender por qué los estudiantes se abandonan las titulaciones universitarias y cómo evitarlo.

La ciencia de datos es una nueva forma de abordar este problema. Aunque no se ha explorado mucho en el ámbito de la educación, tiene el potencial de cambiar la forma en que se toman decisiones para beneficiar a los estudiantes y a la sociedad. Esta disciplina puede ser una herramienta valiosa para ayudar a alcanzar los objetivos, mediante el análisis de datos de los estudiantes y técnicas para prevenir el abandono.

1.2 Objetivos

Este estudio tiene como finalidad aplicar herramientas de análisis de datos al estudio del abandono universitario para ofrecer una visión estratégica del abandono en titulaciones de informática, un enfoque todavía poco explorado en el ámbito educativo que puede ser innovador.

Se pretende encontrar el perfiles y comportamientos de los estudiantes que abandonan sus estudios, detectar patrones, y construir modelos predictivos que puedan anticiparse a la retirada de los estudios universitarios.

Además del enfoque técnico, este Trabajo de Fin de Grado aspira a tener un impacto real en el ámbito de la educación, traduciéndose en resultados concretos que que pueda servir de base para elaborar políticas y medidas por parte de las instituciones universitarias.

1.3 Estructura de la memoria

La memoria de este trabajo se organizará en las siguientes secciones:

- **Capítulo 1 – Introducción**: en esta sección, se contexualiza el problema del abandono universitario a la vez que se expone la motivación del trabajo y se formulan los objetivos generales y específicos.
- **Capítulo 2 – Marco teórico y antecedentes**: se desarrolla un marco conceptual sobre el abandono y se recoge una base teórica necesaria para contextualizar el trabajo.
- **Capítulo 3 – Descripción del conjunto de datos y análisis exploratorio**: se detalla y explica el conjunto de datos, así como todo el procesado de datos y su exploración inicial.
- **Capítulo 4 – Caracterización del alumnado que abandona**: se analizan las características de los estudiantes a partir de variables académcias, sociodemográficas y de rendimiento con el objetivo de encontrar perfiles asociados al cede de la actividad académica.
- **Capítulo 5 – Segmentación del alumnado mediante técnicas de agrupamiento**: se utilizan diferentes técnicas de agrupamiento no supervisado que permiten identificar grupos de estudiantes con patrones similarespara identificar perfiles de riesgo.
- **Capítulo 6 – Modelado predictivo del riesgo de abandono**: se desarrollan modelos de aprendizaje automático y se evalúan con el fin de estimar si un alumno tiene más probabilidad de abandonar a partir de datos históricos de los estudiantes.
- **Capítulo 7 - Creación de un conjunto de datos para docencia** : en este capítulo se trabajará en la creación de un dataset para la asignatura *Análisis de Datos en Educación* del grado de Ciencia de datos para contribuir a seguir mejorando la educación y ofrecer nuevos recursos a docentes y universitarios.
- **Capítulo 8 – Conclusiones y recomendaciones**: se resumen las principales conclusiones del estudio, así como se incluyen propuestas de planes de mejora, continuación del trabajo y se reflexiona sobre las limitaciones del análisis.

1.4 Metodología de trabajo

Este Trabajo de Fin de Grado se ha realizado siguiendo una metodología *incremental y reflexiva*. Se ha desarrollado abordando cada capítulo de manera secuencial, garantizando su adecuado funcionamiento y redacción previamente a avanzar al siguiente capítulo. Sin embargo, varios capítulos utilizan datos modificados anteriormente, por lo que, en varias ocasiones, ha sido necesario revisar y ajustar los capítulos anteriores para poder continuar con los posteriores. De esta manera, el trabajo ha recibido muchas revisiones y correciones, generando un ciclo de mejora continua.

El trabajo, como se ha visto en el apartado anterior, se ha centrado en varias etapas: la carga, limpieza y exploración de los datos, la caracterización del estudiante que abandona y sus perfiles, la anticipación al abandono mediante modelos predictivos y, finalmente, la creación de material docente para su uso en un futuro.

Las herramientas principales utilizadas han sido:

- **R**, como lenguaje para el análisis, transformación y modelado de los datos.
- **RMarkdown y LaTeX**, para integrar el código con la documentación y presentación del trabajo de forma clara y reproducible.
- **GitHub**, como repositorio público del proyecto, para el control de versiones y el almacenamiento de scripts y documentos.

A lo largo del trabajo, se ha tratado de mantener buenas prácticas que contribuyan a la sencillez, claridad y legibilidad del proyecto:

- Separación entre los diferentes datos que servían como *input* y *output* entre diferentes etapas y capítulos.
- Uso de rutas relativas que permitan la lectura y ejecución del proyecto en cualqueir entorno
- Modularización del código en diferentes capítulos y uso de nombres descriptivos.
- Archivos de código legibles mediante comentarios y una adecuada estructuración.

Finalmente, el enfoque se ha centrado en la utilidad práctica del proyecto en el contexto educativo e institucional, identificando claramente conclusiones encontradas y generando una base de datos, permitiendo su explotación futura por el profesorado y el alumnado.

CAPÍTULO 2

Marco teórico y antecedentes

2.1 Factores asociados al abandono en la educación superior

2.2 Comportamiento digital y su influencia académica

2.3 Estado del arte en predicción del abandono

2.4 Técnicas de análisis utilizadas en estudios previos

CAPÍTULO 3
Descripción del conjunto de datos y análisis exploratorio

3.1 Origen y anonimización de los datos

Para el desarrollo de este trabajo, es clave el conjunto de datos del que disponemos. Este dataset tiene como origen el Área de Sistemas Informáticos y Comunicaciones (ASIC) de la UPV, el cuál se ha obtenido a través del cotutor de este trabajo, José Vicente Benlloch Dualde.

El dataset del que estamos hablando es una recopilación de los datos sociodemográficos, académicos y de actividad en las plataformas digitales de todos los estudiantes de la UPV que se han matriculado para el curso 2024-2025. Cada fila corresponde a una asignatura en la que se matriculó cada alumno. Se recibe este dataset con el nombre "dataset_2024.csv".

Adicionalmente, se reciben tres conjuntos de datos más con los que trabajar:

- **"asignaturas_2024.csv"**, el cual contiene su código hasheado, el nombre de la asignatura, su código numérico y su identificador en PoliformaT.
- **"titulaciones_2024.csv"**, el cual contiene el código hasheado de la titulación, su número identificativo y su nombre.
- **"estudiantes_2024.csv"**, el cual contiene el código anonimizado del DNI del alumno, la nacionalidad, la fecha de nacimiento, el sexo, año que se dio de alta en la universidad y la provincia de origen.

Como se ha visto describiendo el conjunto "estudiantes_2024.csv", los datos originalmente contenía datos sensibles de los diferentes estudiantes. Por ello, se ha anonimizado debidamente el dataset para evitar la identificación de los diferentes estudiantes de la UPV:

- Todos los Documentos Nacionales de Identidad han sido sustituidos por un código hash, así como las asignaturas y los estudios correspondientes.
- Los valores de la variable "nacionalitat", que contenía los países de origen de los alumnos, fueron sustituidos por "E" de española y "XXX" de extranjera.

- Los valores de la variable "prov_origen" han sido sustituidos por los valores "ALICANTE", "CASTELLÓN" y "VALENCIA", en caso de pertenecer a la Comunitat Valenciana. En otro caso, toma el valor "ESPANYA".
- En la variable "data_nac", todos los años previos a 1996 tomarán este mismo valor.

3.2 Estructura del conjunto de datos

"dataset_2024.csv" contiene una gran cantidad de variables, llegando a haber 136 columnas. Mediante el lenguaje de programación R, la primera acción es concatenar las variables existentes en los otros tres archivos, de manera que podemos centralizar todas las operaciones de formateado. Una vez hecho esto, el dataset tiene 143 variables.

Con todas las variables en un solo dataset, se puede proceder a la descripción de las variables. Desde el Área de Sistemas Informáticos y Comunicaciones, junto con los conjuntos de datos ya descritos, también se recibió un archivo .README, que sirve de gran ayuda para entender las distinas variables de las que disponemos. A continuación, se ofrece una tabla resumen de las diferentes variables y su descripción:

3.2.1. Datos identificativos

Variable	Tipo	Descripción
dni_hash	character	Hash del DNI del alumno.
asi_hash	character	Hash del código de asignatura.
tit_hash	character	Hash del código de titulación.
titnom	character	Nombre de la titulación.
asinom	character	Nombre de la asignatura.

Tabla 3.1: Variables identificativas

3.2.2. Datos sociodemográficos

Variable	Tipo	Descripción
nacionalitat	character	Código del país de nacionalidad del alumno.
data_nac	integer	Año de nacimiento del estudiante.
sexe	character	Sexo del estudiante (M: mujer, V: varón).
alta_universitat	integer	Año de alta en la UPV.
prov_origen	character	Provincia del domicilio familiar.
anyo_ingreso	integer	Año de primer ingreso en la titulación.
tipo_ingreso	character	Tipo de ingreso en la UPV (traslado, preinscripción...).
nota10	double	Nota de acceso sobre 10 (fase obligatoria de la EVAU).
nota14	double	Nota de acceso sobre 14 (incluye fase voluntaria).
campus	character	Campus donde se imparte la asignatura.
estudios_p	integer	Nivel de estudios del padre.
estudios_m	integer	Nivel de estudios de la madre.
dedicacion	character	Tipo de dedicación (TC: tiempo completo, TP: parcial).
desplazado	integer	Si estudia en provincia distinta a la de origen.

Variable	Tipo	Descripción
discapacidad	double	Indica si el estudiante tiene alguna discapacidad.
becado	integer	Si ha recibido beca (0: no, 1: UPV, 2: otra entidad).

Tabla 3.2: Variables demográficas

3.2.3. Datos académicos

Variable	Tipo	Descripción
preferencia_seleccion	integer	Preferencia con la que eligió el grado en la preinscripción universitaria.
baja_fecha	character	Fecha en la que se dio de baja o abandonó la titulación.
caca	integer	Curso académico de los datos. Igual para todos.
grupos_por_tipocredito	character	Descripción agrupada de créditos matriculados por tipo.
matricula_activa	integer	Indica si el estudiante tenía matrícula activa en ese curso. Su valor es 1 para todas las filas.
nota_asig	numeric	Nota obtenida en la asignatura (0–10 o NP).
cod_centro	character	Código del centro donde se imparte la titulación.
curso_mas_bajo	integer	Curso más bajo matriculado en el año.
curso_mas_alto	integer	Curso más alto matriculado en el año.
cred_matx	character	Créditos totales matriculados en el curso "x".
cred_sup_normal	character	Créditos superados en asignaturas obligatorias y básicas.
cred_sup_espec	character	Créditos superados en asignaturas optativas.
cred_sup	character	Total de créditos superados ese año.
cred_mat_normal	character	Créditos matriculados en asignaturas normales (no prácticas).
cred_mat_movilidad	character	Créditos matriculados a través de programas de movilidad.
cred_ptes_acta	character	Créditos pendientes de calificación en acta.
cred_mat_practicas	character	Créditos de prácticas externas matriculados.
cred_mat_sem_a	character	Créditos matriculados en el cuatrimestre A.
cred_mat_sem_b	character	Créditos matriculados en el cuatrimestre B.
cred_mat_anu	character	Créditos de asignaturas anuales matriculados.
cred_mat_total	character	Total de créditos matriculados ese curso.
cred_sup_sem_a	character	Créditos superados en el cuatrimestre A.
cred_sup_sem_b	character	Créditos superados en el cuatrimestre B.
cred_sup_anu	character	Créditos superados en asignaturas anuales.
cred_sup_total	character	Total de créditos superados ese curso.
rendimiento_cuat_a	character	Rendimiento en cuatrimestre A (sup./mat.).
rendimiento_cuat_b	character	Rendimiento en cuatrimestre B (sup./mat.). Nulo para todas las filas.
rendimiento_total	character	Rendimiento global del curso (sup./mat.).
exento_npp	integer	Asignaturas con NP no computadas en rendimiento.
anyo_inicio_estudios	integer	Año en que comenzó la titulación.
es_retitulado	double	Si ya poseía otra titulación universitaria.
es_adaptado	double	Si procede de un plan anterior adaptado.
cred_sup_xo	character	Créditos superados en un año concreto.
practicas	character	Créditos de prácticas superados.
Actividades	character	Créditos superados en actividades universitarias.
Ajuste	double	Ajuste administrativo de créditos.
cred_sup_tit	character	Créditos totales superados de la titulación.
cred_pend_sup_tit	character	Créditos pendientes para titularse.
impagado_curso_mat	double	Si hubo impago de la matrícula ese curso.
asig1	character	Número de asignaturas superadas ese curso.
pract1	double	Créditos de prácticas superadas ese curso.
activ1	character	Actividades superadas ese curso.
total1	double	Créditos superados ese curso.
ajuste1	integer	Ajuste de créditos ese curso.
rend_total_ultimo	character	Rendimiento total en el último curso activo.
rend_total_penultimo	character	Rendimiento del curso anterior al último.
rend_total_antepenultimo	character	Rendimiento del curso dos años antes.

Tabla 3.3: Variables académicas

3.2.4. Datos de Poliformat y recursos

Variable	Tipo	Descripción
pft_events_2024_x	double	Número de eventos registrados en Poliformat ese mes.

Variable	Tipo	Descripción
pft_days_logged_2024_x	double	Días del mes en los que el estudiante accedió a Poliformat.
pft_visits_2024_x	double	Número total de visitas a Poliformat durante el mes.
pft_total_minutes_2024_x	character	Minutos totales en Poliformat ese mes.
n_wifi_days_2024_x	double	Número de días con conexión WiFi en el campus.
n_resource_days_2024_x	double	Días en los que accedió a recursos digitales.
resource_events_x	double	Total de acciones en la pestaña 'Recursos" de la asignatura.
pft_assignments_submissions_2024_x	double	Envíos de tareas realizados en Poliformat.

Tabla 3.4: Variables de actividad en Poliformat y recursos digitales.

Nota: El sufijo x indica el número de mes (7 es julio, 8 es agosto... y 1 es enero).

3.3 Definición de la variable "abandono"

El conjunto de datos, tal como está ahora, no nos ofrece ninguna variable que nos indique si un estudiante ha abandonado o no los estudios. Por ello, se ha creado una variable binaria llamada "abandono". Esta variable tiene como función identificar a aquellos estudiantes que han abandonado los estudios académicos.

Para poder encontrar a los alumnos que han abandonado, se ha hecho un recuento de las fechas en las que se ha dado de baja, a partir de la variable "baja_fecha".

Para poder realizarlo, se han agrupado los diferentes estudiantes por su "dni_hash" y se ha procedido al recuento de las asignaturas con una baja registrada.

Si el recuento es de 5 o más asignaturas en las que existe una fecha de desmatriculación, se considerará que ha abandonado la carrera y tomará el valor 1. En caso de que sea menor, se pueden considerar bajas puntuales de algunas materias por diferentes motivos, pero que no necesariamente implica un abandono del título, por lo que tomará el valor 0.

3.4 Preprocesado y limpieza de datos

El conjunto de datos del que disponemos, en formato Dataframe del lenguaje R, aún requiere de limpieza y tratamiento para poder ser utilizado para los fines descritos en este documento.

Inicialmente, existen una gran cantidad de columnas numéricas mal formateadas. Estas se han reconocido como si fueran cadenas de texto por tener una "," como separador. Se han sustituido por ".", lo que ha permitido reformatearlas a números reales (formato "double" en R) para poder ser utilizados. Además, se ha reparado el tipo fecha de la variable "fecha_datos", así como de "baja_fecha", la cual es especialmente relevante.

También se han eliminado las columnas que solo contienen valores nulos o están vacías. Estas han sido:

- `abandono`: Variable indicadora de si el estudiante ha abandonado la titulación, que se encontraba vacía.
- `rendimiento_cuat_b`: Rendimiento académico en el cuatrimestre B. Solo disponemos de datos del primer cuatrimestre, por lo que no tiene sentido mantener esta variable.

- `pft_assignment_submissions_2024_7`: El número de entregas de tareas en Poliformat en julio de 2024 para todas las asignaturas está vacía, al no ser, habitualmente, un periodo de entregas.
- `pft_test_submissions_2024_7`: Número de envíos de test en Poliformat en julio de 2024 para todas las asignaturas está vacía, al no ser, habitualmente, un periodo de entregas.
- `pft_assignment_submissions_2024_8`: Número de entregas de tareas en Poliformat en agosto de 2024 para todas las asignaturas está vacía al no ser un periodo lectivo.
- `pft_test_submissions_2024_8`: Número de envíos de test en Poliformat en agosto de 2024 para todas las asignaturas está vacía, al no ser un periodo lectivo.
- `impagado_curso_mat`: Número de matrículas impagadas.
- `ajuste1`: ajuste de créditos por reconocimientos, adaptaciones...
- `cred_sup_anu`: créditos superados en asignaturas anuales para este curso.
- `cred_mat_5`: créditos matriculados en quinto curso. La carrera de informática solo tiene cuatro años, por lo que está vacío.
- `cred_mat_6`: créditos matriculados en sexto curso. La carrera de informática solo tiene cuatro años, por lo que está vacío, además de que solo tiene sentido en carreras de arquitectura.
- `cred_sup_5o`: créditos superados en sexto curso. La carrera de informática solo tiene cuatro años, por lo que está vacío, además de que solo tiene sentido en carreras de arquitectura.
- `cred_sup_6o`: créditos superados en sexto curso. La carrera de informática solo tiene cuatro años, por lo que está vacío, además de que solo tiene sentido en carreras de arquitectura.

Nota: Aunque se elimine la variable "abandono", se recuperará más tarde conforme al criterio establecido.

Una vez eliminadas estas 6 variables, el conjunto de datos tiene 137 columnas.

Adicionalmente, se han eliminado todas las asignaturas que pertenecían al segundo cuatrimestre. Al no disponer de datos sobre el desempeño de los alumnos en estas asignaturas, solo ocupan espacio de manera innecesaria. Se han recogido desde el plan de estudios de la Escuela Técnica Superior de Ingeniería Informática las asignaturas del primer cuatrimestre para hacer un filtro eliminar las demás. Para asegurarse de que se disponen de todas las asignaturas de las que un alumno se ha desmatriculado, se ha creado, previo a la eliminación de las asignaturas del segundo cuatrimestre, el dataset "abandono" que se comentará más adelante.

Para facilitar el trabajo con todas las variables referidas al entorno de PoliformaT, se ha realizado un filtrado y ha sido reordenadas en el dataset, de manera que se encuentran seguidas y en orden de mes todas variables referidas al mismo aspecto. Es decir, si tomamos como ejemplo los eventos en PoliformaT, "pft_events_2024_x", encontraremos que están seguidas todas las columnas de eventos desde julio hasta enero.

Además, en las variables relacionadas con PoliformaT, se han sustituido todos los valores nulos por 0, ya que realmente un valor nulo en "total_minutes", por ejemplo, en una asignatura corresponde a 0 minutos dedicados. También se ha aplicado esta medida a las variables "desplazado", "discapacitado", "es_adaptado", "es_retitulado", "exento_npp", que se encontraban en la misma situación.

Continuando con este dataset, se han detectado que hay varias variables que tienen datos menores a 0, los cuáles no tienen sentido. Las variables son: "pft_total_minutes_2024_7", "pft_total_minutes_2024_10", "pft_total_minutes_2024_11", "pft_total_minutes_2024_12". Sus valores negativos serán sustituidos por 0.

Finalmente, se ha tomado un subconjunto de los datos, que son los referidos a la carrera de informática. Eliminaremos la variable "campus", ya que no tiene sentido al pertenecer todos al campus de Vera. Lo mismo ocurre con "matricula_activa", pues para todos los estudiantes de informática tiene el valor 1. Finalmente, también se va a prescindir de la variable "nota_asg", ya que 1000 valores (la mitad) son nulos, y solo 50 son mayores que 0.

Adicionalmente, se ha cambiado el tipo de la variable "preferencia_seleccion" a factor, y todos los valores superiores a 4, se han considerado preferencia baja ("Baja").

Al explorar los cuartiles de varias variables, se pueden observar varias de ellas que tienen hasta un tercer cuartil prácticamente 0, para luego tener un valor máximo muy por encima de los valores anteriores. Sin embargo, se ha decidido mantenerlo por el momento, ya que pueden estar contextualizadas. En el análisis exploratorio que hay en el siguiente capítulo se procederá a eliminar o mantener dichas variables con valores muy bajos y datos atípicos.

Los únicos valores extremos superiores que han sido cambiados son los tests de las variables pft_test_submissions_2024_x y resource_events_2024_10, donde se han cambiado sus valores a 30 y 400, respectivamente.

3.5 Creación de nuevas variables y conjuntos de datos

Las variables de las que se disponen actualmente resultan claves y ofrecen una información muy valiosa para lograr los objetivos previstos anteriormente. Sin embargo, todavía resulta posible obtener algo más de información y algunas variables más, las cuáles permitan tener un mejor contexto de la problemática.

En primer lugar, se toman la formación sobre el comportamiento de los alumnos en PoliformaT y se crean variables que resultan de la suma de los diferentes valores para cada asignatura y alumno por cada variable digital, tal como se muestra en la tabla3.5

n_resource_days_2024_7	n_resource_days_2024_8	n_resource_days_2024_9
0	0	1

n_resource_days_2024_10	n_resource_days_2024_11	n_resource_days_2024_12
3	3	3

n_resource_days_2025_1	**n_res_day_asg**
0	**10**

Tabla 3.5: Acceso a recursos digitales por mes y asignatura (dividido en bloques)

Una vez hecha la suma para cada asignatura, se hace la media de las diferentes asignaturas, de manera que solamente tendremos una fila por alumno, obteniendo una serie de variables llamadas "{ámbito_digital}_{año}_{mes}_asg_media".

Por otro lado, también se han creado unas columnas similares, pero en este caso, dichas variables son la suma de las interacciones cada mes de cada campo digital por estudiante, pero no por asignatura. Es posible ver un ejemplo en la tabla 3.6.

dni_hash	**asinom**	**pft_total_minutes_2024_12**
01b51458bc21	Introducción a la informática y a la programación	0.00000
01b51458bc21	Análisis matemático	222.59768
01b51458bc21	Matemática discreta	0.00000
01b51458bc21	Fundamentos Físicos de la Informática	45.14922
01b51458bc21	Fundamentos de computadores	0.00000
01b51458bc21	Tecnología de sistemas de información en la red	45.16452
01b51458bc21	Bases de datos y sistemas de información	0.00000
01b51458bc21	Computación paralela	376.50470
01b51458bc21	English for Computing (B2)	367.27135
01b51458bc21	Diseño y gestión de bases de datos	127.02882
	pft_total_minutes_2024_12_est	**1183.716**

Tabla 3.6: Minutos totales en Poliformat durante diciembre de 2024 por asignatura

Finalmente, se han separado todas las variables en varios Dataframes de R con el objetivo de facilitar el trabajo y de que exista una mayor claridad. Se ha separado en tres conjuntos de datos diferentes: "poliformat", que contiene los datos sobre PoliformaT de cada alumno, pero solamente de los datos artificiales, habiendo una fila por alumno; "sociodemografía", que contiene todos los datos sobre los alumnos y sus características sociales; y "creditos", que contiene toda la información académica de los diferentes estudiantes.

Finalmente, se va a crear un Dataframe solamente con datos de "abandono", apartado de los demás, que recibe el mismo nombre. Este contiene el código del DNI del alumno, el número de asignaturas que ha abandonado, la fecha de la baja y la variable binaria anteriormente comentadada que indica si ha abandonado o no.

3.6 Análisis exploratorio de los datos

Para realizar el análisis, vamos a apoyarnos en gráficas y en funciones de las librerías "ggplot2" y "skim". Todos los datasets contienen datos de 1889 alumnos matriculados en la carrera, siendo uno cada fila. La excepción es el conjunto de la actividad digital, donde una fila es la actividad que ha tenido un alumno en esa asignatura concreta, que tiene 17576 filas.

3.6.1. Análisis del conjunto de datos sociodemográficos

Este dataset tiene 1754 filas y 18 columnas. La información sobre las variables categóricas se puede encontrar en la tabla 3.7, mientras que sobre las numéricas se puede ver en la tabla 3.8.

Variable	NAs	Posibles valores	Valores más frecuentes
nacionalitat	0	E, XXX	E: 1634, XXX: 120
sexe	0	V, M	V: 1478, M: 276
prov_origen	0	COMVAL, ESPANYA	COM: 1508, ESP: 246
tipo_ingreso	0	NAP, NTE, BMA, NLE, NRO, (otros)	NAP: 1617, NTE: 43, BMA: 41, NLE: 30
estudios_p	5	1, 2, 3, 4, 5, 6, Desconocido	5: 756, 4: 594, 3: 323, 6: 57
estudios_m	5	1, 2, 3, 4, 5, 6, Desconocido	5: 867, 4: 606, 3: 244, 2: 18
dedicacion	0	TC, TP	TC: 1727, TP: 27
desplazado	0	0, 1	0: 1279, 1: 475
discapacidad	0	0, 1	0: 1723, 1: 31
becado	0	0, 1, 2	0: 1232, 2: 489, 1: 33
preferencia_seleccion	0	1–10, Baja, Desconocido	1: 1267, 2: 235, 3: 111, Desconocido: 56

Tabla 3.7: Resumen de variables categóricas según `skimr`, sin datos ausentes.

Variable	NAs	Compl.	Media	SD	Mín	P25	P50	P75	Máx
data_nac	0	1.00	2003.23	2.62	1996	2002	2004	2005	2007
alta_universitat	0	1.00	2021.34	3.31	1989	2021	2022	2023	2024
anyo_ingreso	5	0.997	2021.67	2.38	2010	2021	2022	2023	2024
nota10	51	0.97	7.9	0.81	5.15	7.4	7.9	8.43	10
nota14	28	0.99	10.86	1.27	5	10.31	10.9	11.62	13.73

Tabla 3.8: Resumen de variables numéricas según `skimr`.

Observando las variables que son tipo factor, nos damos cuenta de que todas están prácticamente completas, a excepción de 5 valores en el tipo de ingreso a la universidad, en los estudios tanto del padre como de la madre y luego, 67 valores faltantes en preferencia de selección.

Con respecto a las variables numéricas, no se observa nada nada fuera de lo común. Existen valores faltantes en las variables de las notas de selectividad y la preferencia de selección. Mirando el ratio de completitud, vemos que no es una cantidad significativa de valores faltantes, pues el mínimo es de 0.97.

Para una mejor interpretación y un borrado de valores faltantes, se van a sustituir los valores nulos por el valor "Desconocido", para un mejor desempeño en tareas futuras.

Empezando con los resultados obtenidos en la figura 3.1, en losgráficos de caja y bigotes, se pueden observar que existen datos previos de alta de universidad a 1996. Esto no tiene sentido, ya que l¡el valor mínimo es 1996, por lo que se debe hacer un ajuste. Por otro lado, se observa que el año de ingreso y el alta de universidad, a partir de 2010, están muy a la par, siendo prácticamente idénticas. Solo viendo las gráficas, se puede predecir que estarán muy relacionadas.

Los gráficos de densidad de la figura 3.1 de las notas de selectividad de los estudiantes parecen bastante normales, con medias centradas en 7,5 y en el 9,5 para la nota en fase obligatoria y la fase voluntaria, respectivamente.

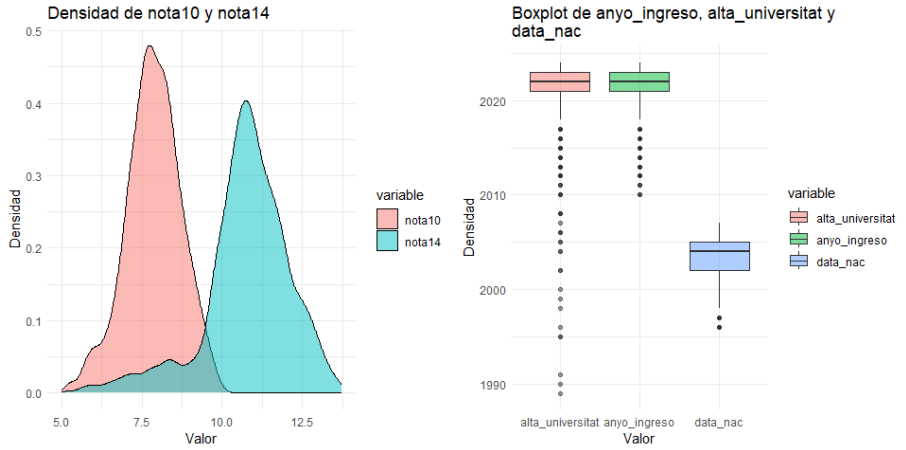


Figura 3.1: Boxplot de las variables sociodemográficas: año de ingreso, año de alta en la universidad y fecha de nacimiento.

Pasando a las variables categóricas mostradas en la figura 3.2, se puede ver que el tipo de ingreso más común es "NAP", que significa "Nueva Admisión en Preinscripción", que suele ser el formato más habitual de entrada a la universidad. El resto de tipos de ingreso tienen valores mínimos, indicando que no son muy comunes esas formas de inscribirse.

La mayoría de estudiantes que están en la carrera la cogieron como primera opción. Alrededor de 250 estudiantes la tomaron como segunda opción, muchos menos que en primera. El número sigue descendiendo a medida que bajamos posiciones, siendo mínimo a partir de ser la séptima opción.

Con respecto a las becas, más de 1250 alumnos que se matricularon en la carrera de Informática entraron sin beca, mientras que 500 estudiantes fueron becados por otras entiedades (Ministerio, GVA) y los demás han sido becados por la Universitat Politècnica de València.

Los niveles de estudios del padre y de la madre son muy similares. Predominan los estudios terciarios con el código 5 con cierta superioridad de las madres, seguidos de los secundarios con datos muy parecidos, mientras que en los estudios primarios son los padres quienes tienen datos más altos. Finalmente, está el nivel con el código 6, que significa "No procede", y "Sin estudios" y "Analfabetos" teniendo valores muy bajos en ambos padres.

Cuando se toman en consideración las variables binarios, se ve a primera vista que en las variables "desplazado" y "discapacidad", prácticamente la totalidad de los estudiantes estudian a tiempo completo y no son discapacitados.

Más del 90 % de los matriculados son de nacionalidad española. Un 86 % de los estudiantes procede de la misma comunidad autónoma, coincidiendo con que un 27 % de los estudiantes es desplazado, es decir, procede de fuera de la provincia de Valencia, pero puede ser de la Comunitat Valènciana.

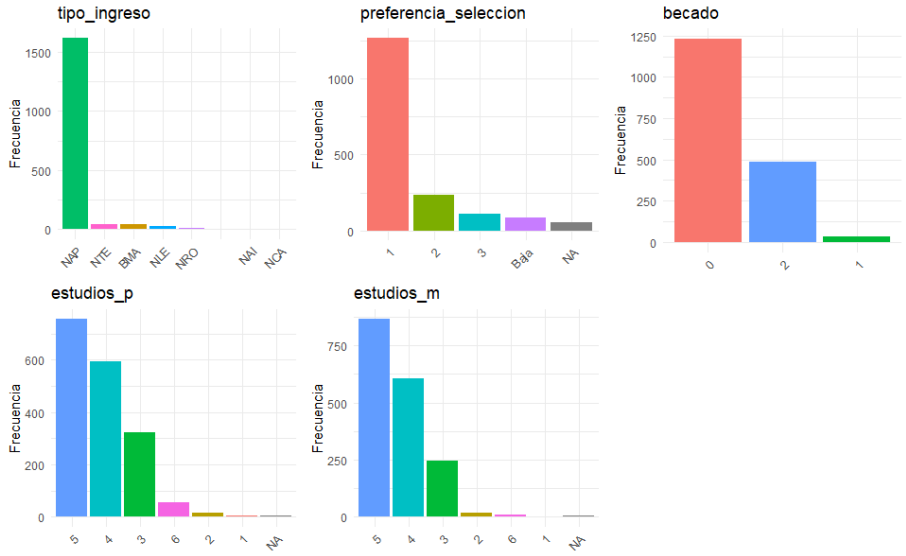


Figura 3.2: Boxplot de las variables binarias sociodemográficas.

Continuando con el estudio, en la figura 3.4 se puede apreciar la matriz de correlaciones de las diferentes variables del conjunto de datos actual, donde se han filtrado correlaciones mayores a 0.3 en valor absoluto. Se puede ver que existen claras correlaciones entre el año de nacimiento, el año de ingreso y el alta en la universidad. Todas tienen valores similares y, por lo general, los estudiantes entran con una edad muy similar (18-20) años a la universidad, con lo que existe una clara correlación entre ellas. Entre la notas de selectividad también hay una clara correlación, dado que la nota14 es nota10 más el resultado de la fase voluntaria del alumno. También parecen estar muy relacionadas las notas de la selectividad con el año de ingreso, de alta a la universidad y de ingreso. Esto puede darnos la pista de que las notas variaron entre los diferentes años. Finalmente, se destaca la correlación entre los estudios del padre y de la madre, los cuáles no tienen correlación relevante con ninguna otra variable. Sin embargo, nos encontramos con que "abandono" está completamente desaparecido.

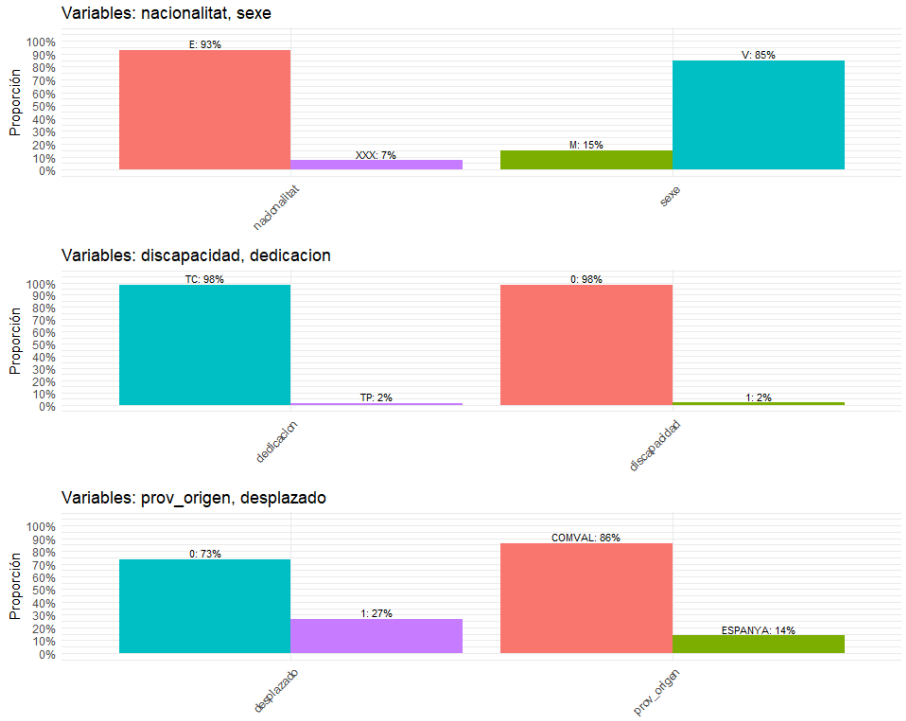


Figura 3.3: Boxplot de las variables binarias sociodemográficas.

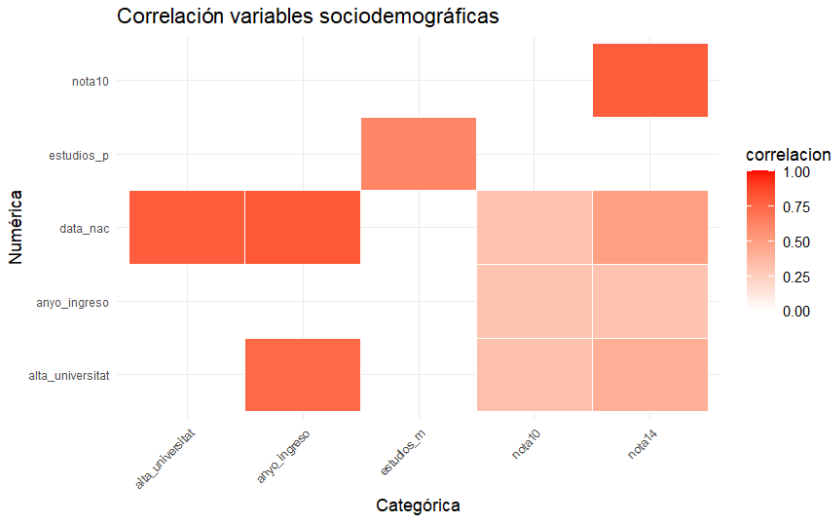


Figura 3.4: Matriz de correlación de las variables sociodemográficas

Dadas las correlaciones vistas en la figura 3.4, resulta relevante utilizar una técnica de reducción de la dimensionalidad. A partir del Análisis de Componentes Principales que se puede ver en la figura 3.5, podemos extraer que, con carácter más general, las variables sociodemográficas están poco correlacionadas. La primera componente solo logra explicar un 10 %. Tomando 10 dimensiones, llegaríamos al 53 % de varianza explicada.

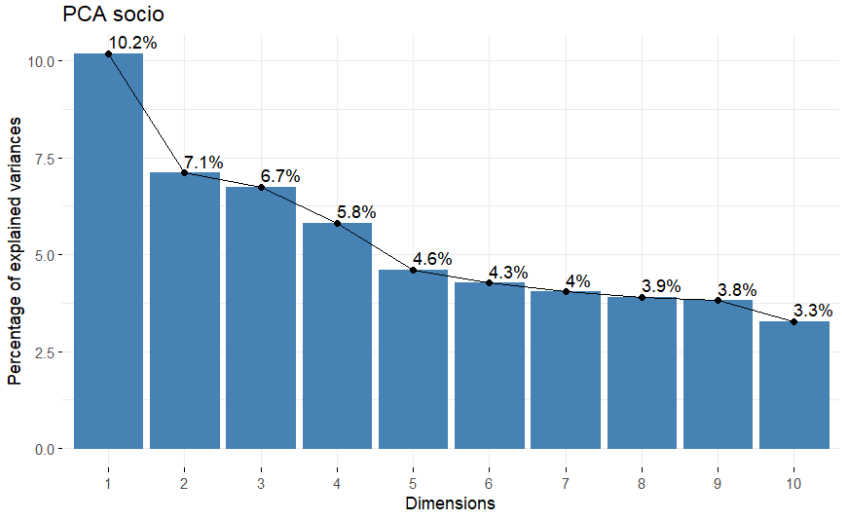


Figura 3.5: PCA de las variables sociodemográficas

3.6.2. Análisis del conjunto de datos académicos

Este dataset tiene 1754 filas y 44 columnas. La información de las variables categóricas se puede ver en la tabla 3.9, mientras que las numéricas se pueden ver en la tabla A.1.

Variable	NAs	Valores	Valores más frecuentes
curso_mas_bajo	0	1-4	1: 596, 2: 552, 4: 349, 3: 257
curso_mas_alto	0	1-4	4: 510, 1: 436, 2: 409, 3: 399
exento_npp	0	0-1	0: 1338, 1: 416
es_retitulado	0	0-1	0: 1752, 1: 2
es_adaptado	0	0-1	0: 1748, 1: 6

Tabla 3.9: Resumen de variables categóricas académicas según `skimr`.

Las variables categóricas, a la vista de su tabla correspondiente, tienen todos sus datos dentro de valores normales. Llama la atención los pocos positivos de "es_retitulado" y "es_adaptado".

Con respecto a las variables numéricas, se pueden ver algunos valores poco comunes que se deben examinar. Por ejemplo, variables como las referidas a los créditos superados presentan hasta terceros cuartiles con valor 0, para luego tener valores máximos altos. Aunque, al igual que muchas otras variables, tiene valores dentro de los rangos esperados, se pondrá énfasis en dichas variables una vez sean graficadas.

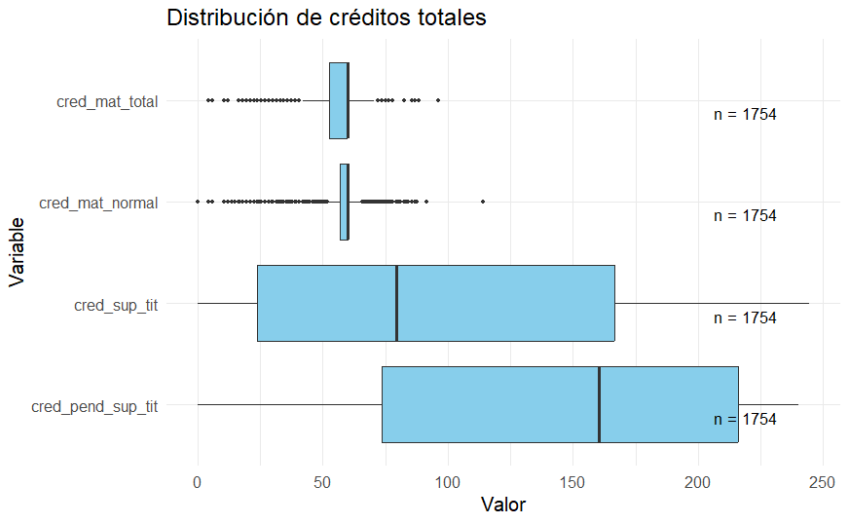


Figura 3.6: Distribución de créditos acumulados en titulaciones.

En la gráfica 3.6 no se observan grandes anomalías. Los estudiantes, habitualmente, se matriculan de 60 créditos por curso, pudiendo haber más o menos. Sin embargo, llaman la atención los valores bajos, ya que el mínimo de créditos que te puedes matricular por curso son 18.

Por otro lado, tanto cred_sup_tit como cred_pend_sup_tit tienen valores bastante normales, y no enfrentan valores atípicos. Ambas variables son complementarias, ya que hasta 240 créditos que se deben superar en la carrera, los que no se hayan superado son los pendientes. Hay algunos casos en los que se han superado hasta 250, probablemente por haber cursado asignaturas o haber obtenido créditos de otras maneras como actividades o prácticas, siendo sobrantes. Por ello, se va a proceder a la eliminación de cred_pend_sup_tit.

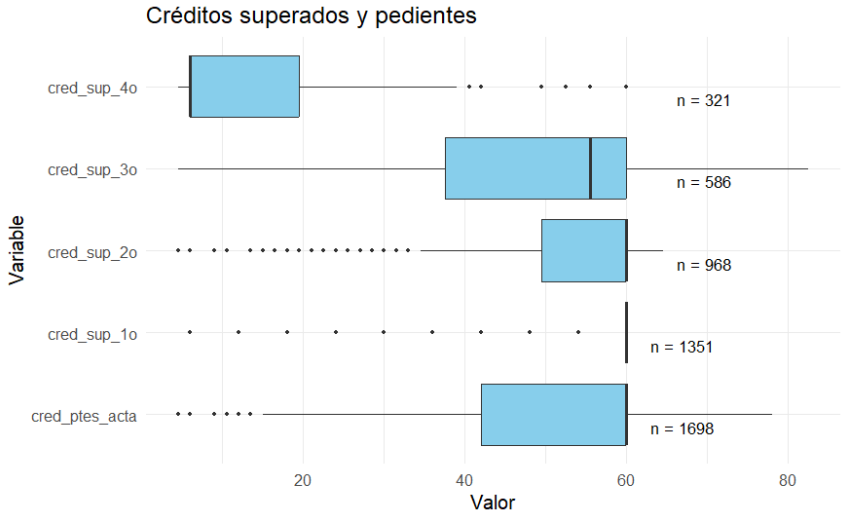


Figura 3.7: Créditos totales por curso o fase académica.

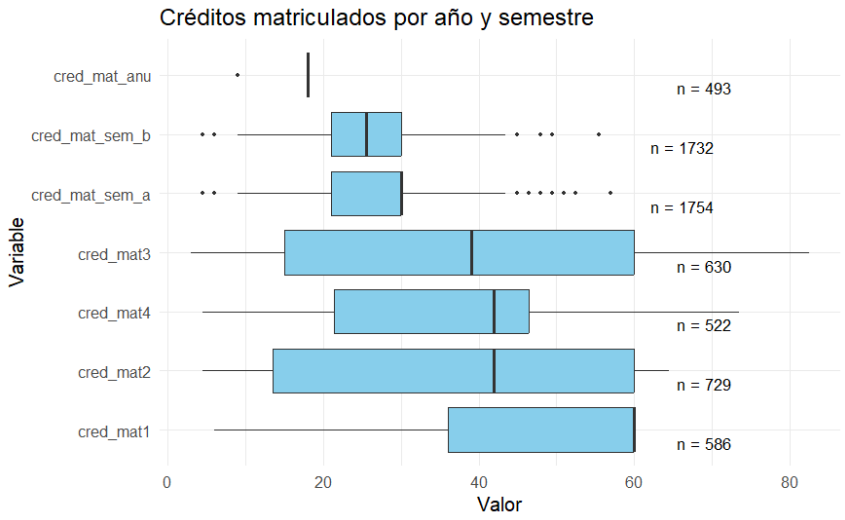


Figura 3.8: Créditos matriculados por año y semestre.

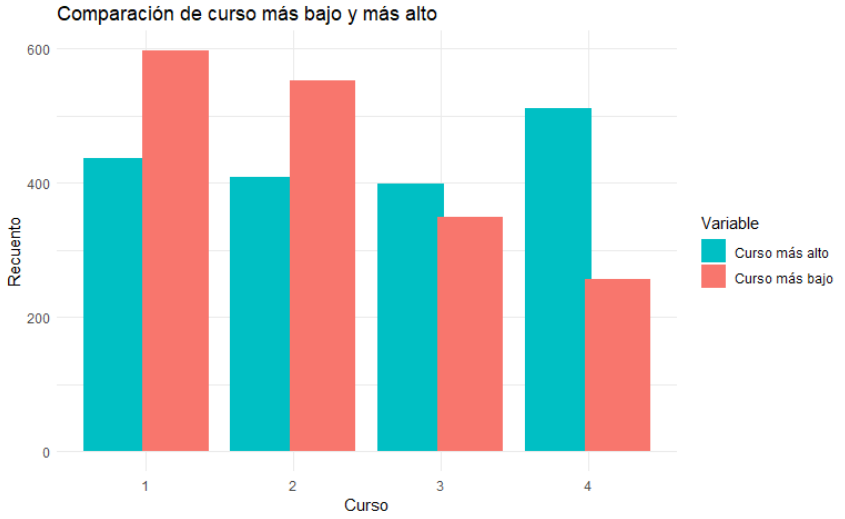


Figura 3.9: Comparación de curso más bajo y más alto.

Viendo la gráfica 3.7 y la gráfica 3.8, se pueden apreciar diversos factores. En primer lugar, los estudiantes de informática no superan todos los créditos de los que se matriculan. En primer curso, donde la mediana se encuentra en 60 créditos (las 10 asignaturas de primero) casi todos parecen superar todos los créditos, a excepción de algunos, que son esos valores atípicos. En segunda, ya se puede ver como el rango es más amplio, dejando entrever que los alumnos se matriculan de un número mucho más variable, al igual que en tercero. Esto indica que muchos han suspendido alguna asignatura, pero que siguen adelante con la carrera. Las asignaturas que pudieran haberles quedado son el primer cuartil de "cred_mat1", y los valores atípicos de segundo dependen de las asignaturas que no hayan logrado aprobar. De manera similar ocurre con tercero de carrera, con un rango de créditos superados más amplio a la baja, pero similar en matriculaciones. En cuarto de carrera hay menos creditos superados. Esto puede indicar que muchos alumnos que cursan asignaturas de cuarto estando no logran sacarlas adelante, teniendo que quedarse otro año más solo con asignaturas de ese mismo curso, aumentando el rango a la baja en ambas gráficas.

Finalmente, los aspirantes a terminar la carrera se matriculan, generalmente, del mismo número de asignaturas cada año: principalmente 60, pero lo habitual es que no sea menos de 45. Aunque solo se disponga del número de datos y no sea el de alumnos en ese curso (un mismo alumno se puede matricular en dos cursos), es posible tener un esbozo de cuántos estudiantes están matriculados en cada año. A la vista de la figuro 3.8, parece haber un número similar de estudiantes en cada año.

Por otro lado, también se puede ver en la gráfica 3.7 que hay una gran cantidad de créditos pendientes del acta, pues en el momento en que se recogieron los datos aún no se había publicado las actas del primer cuatrimestre. Además, observando los créditos matriculados en ambos cuatrimestres en la gráfica 3.8, se puede ver que intentan balancear los créditos entre ambos, estando alrededor de 25 y 30 generalmente.

Siguiendo con créditos superados y matriculados, la figura 3.10 deja ver que siguen faltando muchos datos al no estar las calificaciones subidades, asó como un número bajo de estudiantes matriculados en prácticas (alumnos de cuarto, principalmente) y de movilidad. Además, los créditos superados en total y el primer cuatrimestre son idénticos (lógico, ya que no disponemos de datos del segundo, a excepción de un par de datos), por lo que convendría eliminar una de las dos, que será la variable total.

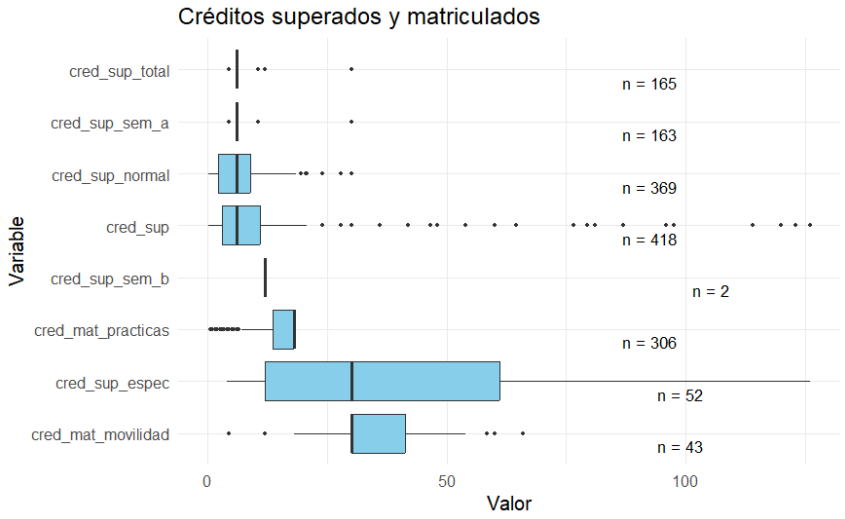


Figura 3.10: Créditos por movilidad, prácticas y características especiales.

A excepción de los créditos superados de manera especial ("cred_sup_espec"), que han tenido valores muy altos este curso (posiblemente sean cambios de carrera o de universidad), los créditos superados tienen valores muy bajos. La justificación es la ausencia de la mayoría de notas para el cuatrimestre en el momento de la recogida de datos. Por lo tanto, estas variables tienen pocos datos y poca variabilidad. Además, los créditos superados para el cuatrimestre b son, 2 datos, por lo que esta variable se eliminará.

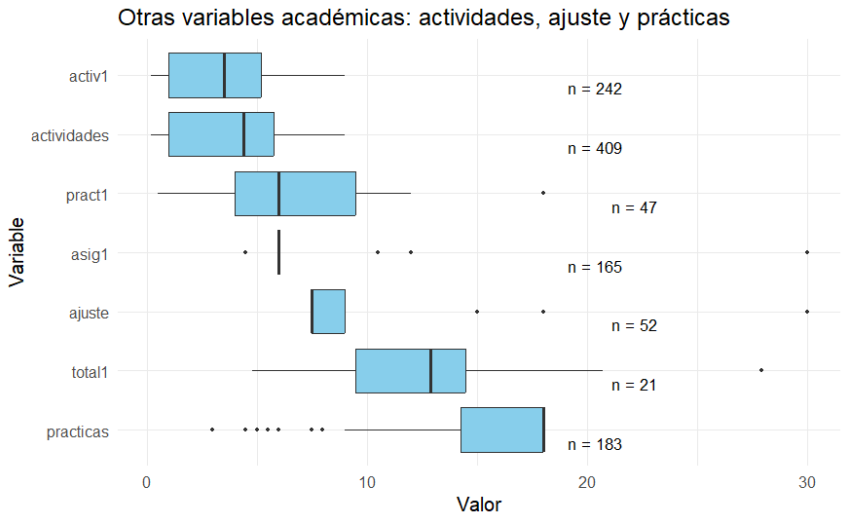


Figura 3.11: Otras variables académicas: actividades, ajuste y rendimiento.

PREGUNTAR PARA COMENTAR MEJOR MÁS ADELANTE.

Observando la gráfica 3.11, se puede observar que las variables "practicas" y "asig1" son prácticamnte nulas. Esto indica que no muchos estudiantes aún no han superado las prácticas y muchas asignaturas. Esto tiene sentido, que los datos fueron recogidos previamente a que las notas del primer cuatrimestre fueran publicadas, de manera que las notas de los alumnos están ausentes. Además, siendo el primer cuatri, no muchos alumnos son los que han cursado prácticas a la vista de la gráfica.

Siguiendo con la misma gráfica, se puede apreciar que muchos alumnos han superado muchos créditos a través de actividades. Desconocemos el tipo de actividades, pero estas pueden ser mediante colaboraciones en entidades de la universidad (Generación Espontánea o Delegación de Alumnos, por ejemplo) o por

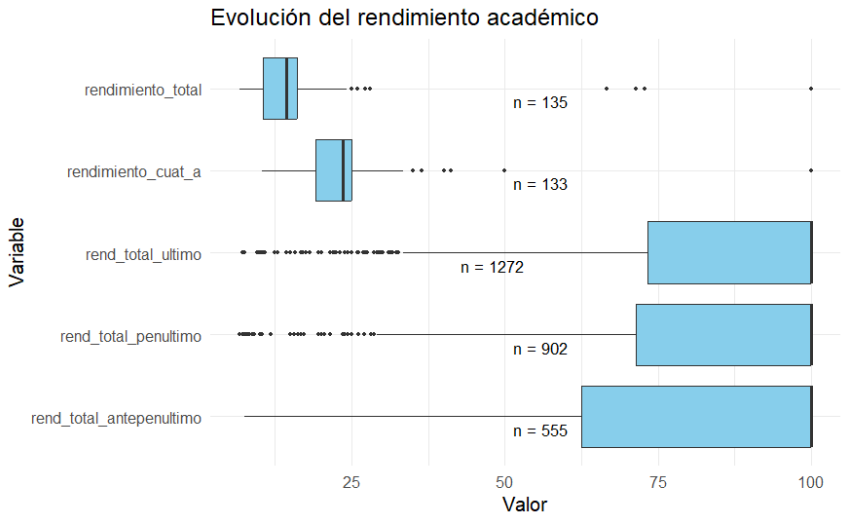


Figura 3.12: Evolución del rendimiento académico: total, por cuatrimestre y últimos cursos.

Observando la figura 3.12, se puede ver que los rendimientos de los estudiantes de entre hace tres años y el año pasado son similares, con una notable mejora. El número de matriculaciones crece con cuánto menos es el histórico, lo cual tiene sentido, pues los estudiantes abandonan la universidad al terminar y no se disponen de esos datos, aunque

no se disponen de todos. Sin embargo, los rendimientos total y de este curso no son significativos, al disponer solamente de un 10 % de los datos, por motivos anteriormente comentados.

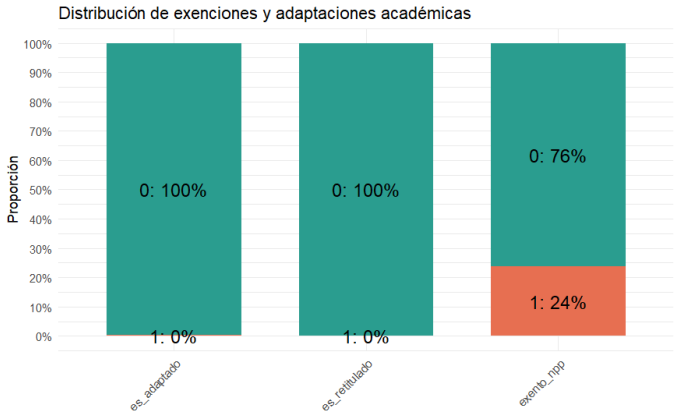


Figura 3.13: Distribución de exenciones y adaptaciones académicas.

También se puede observar en la figura 3.13 que la práctica totalidad de los estudiantes no son adaptados y, aunque los cálculos reflejen un 100 %, existen tres estudiantes retitulados, y solo el 24 % de lo estudiantes está a punto de terminar la carrera (es decir, prácticamente son los estudiantes de cuarto de carrera).

Continuando con el estudio, en la figura 3.14 se puede apreciar la matriz de correlaciones de las diferentes variables del conjunto de datos actual, habiéndose filtrado correlaciones mayores a 0.3. Se puede ver que las variables de rendimiento están relacionadas con las variables de créditos superados, lo cual es esperable, ya que el rendimiento viene de los créditos superados y matriculados. También se puede ver una notable relación entre curso más alto y los créditos matriculados, indicando que el nivel del curso está relacionado con el volumen de la matrícula. Con respecto a abandono, este se encuentra con correlaciones bastante bajas, pero existen. Se puede apreciar correlación con variables de rendimiento y de créditos superados y pendientes de acta.

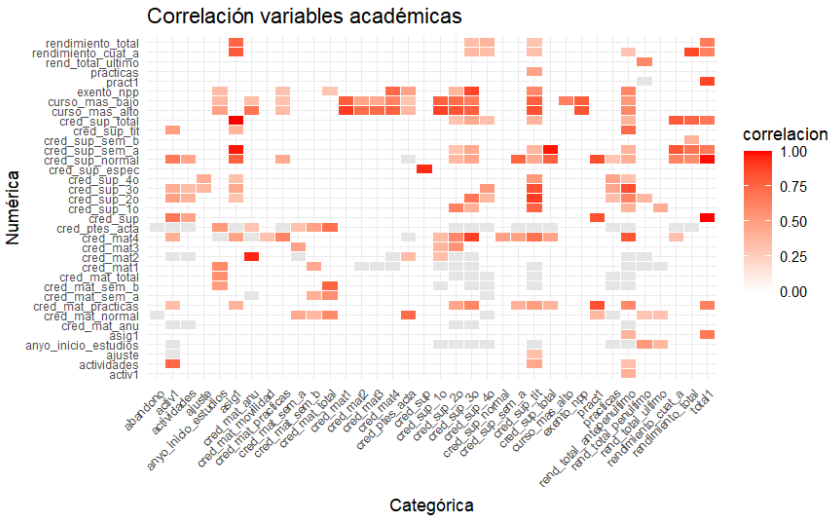


Figura 3.14: Matriz de correlación de las variables académicas

De manera más urgente que con el dataset “sociodemografía”, se encuentra la necesidad de utilizar un Análisis de Componentes Principales para reducir la dimensionaldiad,

vista en la figura3.15. La primera componente refleja por sí sola un 25 % de la varianza del dataset, lo cual es un muy buen dato comparado con la figura 3.5. Sin embargo, el resto de componentes también son bastante bajas, aunque se dispone de un 50 % de la varianza con cuatro componentes.

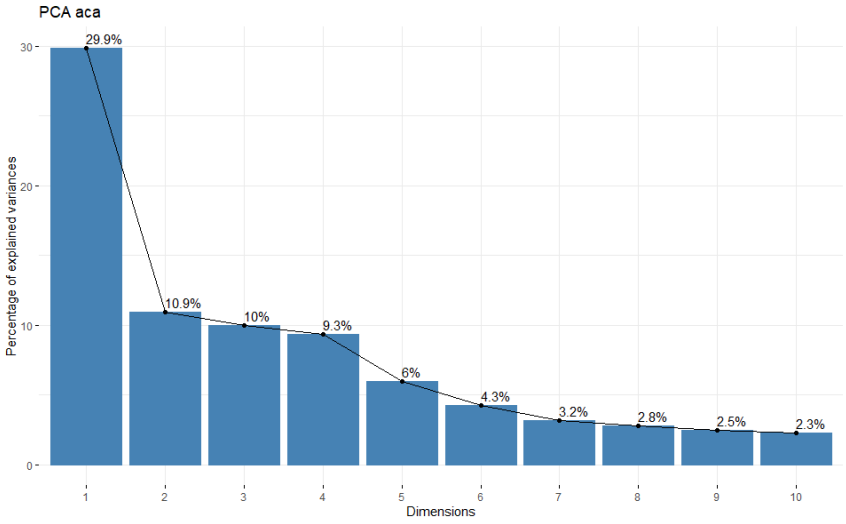
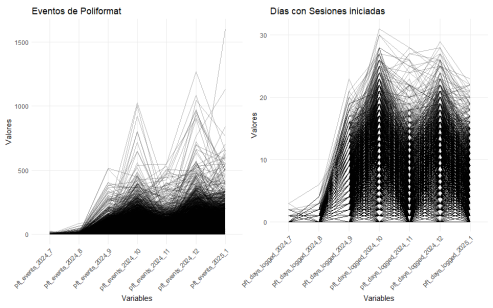


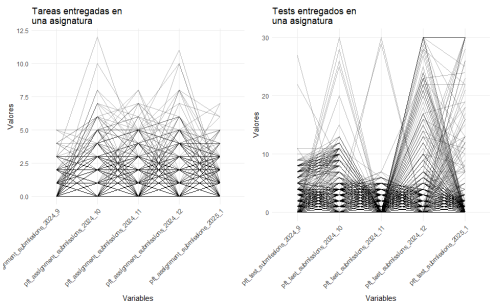
Figura 3.15: PCA de las variables académicas

3.6.3. Análisis del conjunto de datos de Poliformat

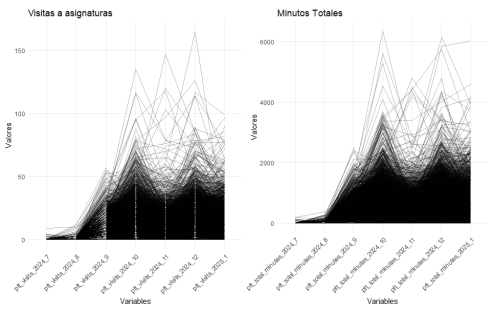
Este dataset tiene 1754 filas y 70 variables. La información sobre las columnas se puede ver en la tabla A.2.



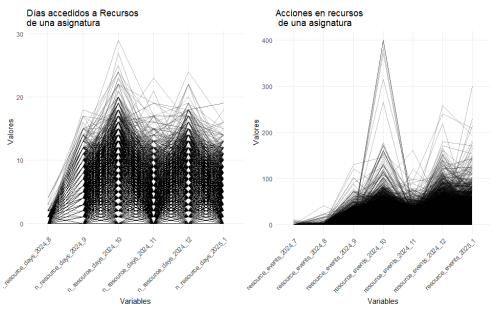
(a) Eventos en Poliformat y días con sesiones iniciadas por estudiante.



(b) Número de tareas y tests entregados por asignatura.



(c) Visitas a Poliformat y tiempo total en minutos por estudiante.

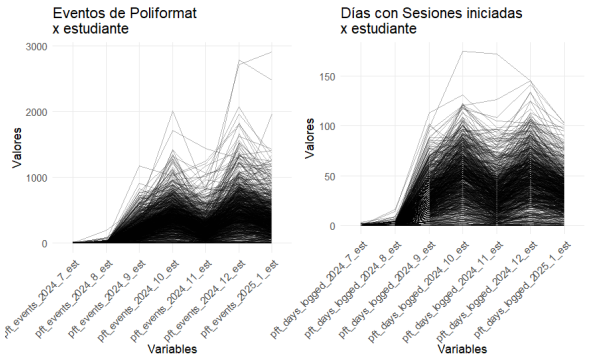


(d) Accesos a recursos y número de eventos en ellos.

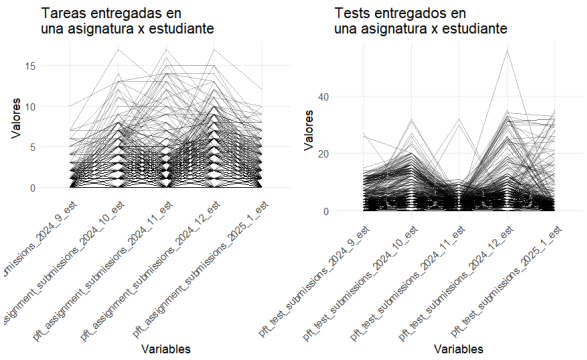
Figura 3.16: Evolución de uso de Poliformat por estudiante: variables originales.

Una vez vista la figura 3.16, se puede ver que las variables que venían en los datos originales ya no presentan ninguna anomalía, aunque sí algunos puntos destacables. Los eventos de Poliformat que, como recordatorio, son todas las acciones realizadas por un estudiante en una asignatura tienen valores de 1000 y hasta de 1500 acciones. No es imposible que un alumno haya hecho tantas acciones, pero es llamativo. Con respecto a las tareas, se puede ver como hay también valores algo altos, como 10 entregas en un mes, pero puede deberse también a reenvíos de la propia tarea. Lo mismo puede ocurrir con los tests, los cuáles tenían valores muy anormales, pero que ahora han sido acotados a 20. Las visitas a una asignatura, los minutos pasados en la plataforma y las acciones en los recursos de una asignatura también son muy dependientes de los estudiantes, por lo que no es de extrañar que tengan valores altos.

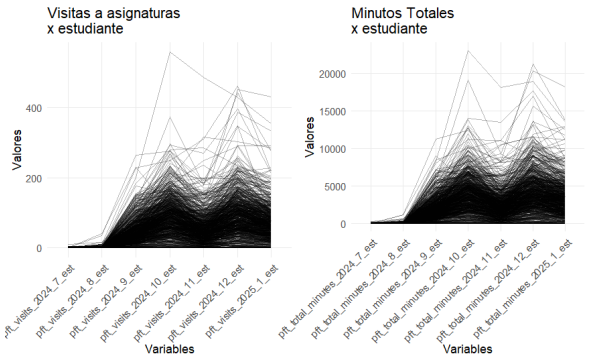
Como apunte final para estas gráficas, se puede apreciar como los meses de mayor actividad, en general, son octubre y diciembre, donde iban a tener lugar los primeros exámenes y dónde tuvieron realmente lugar a la vez que se estudiaba para los de enero, respectivamente.



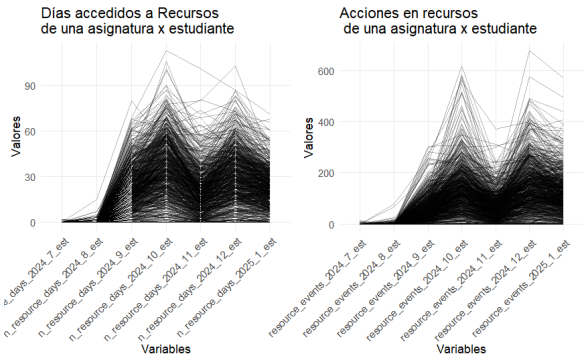
(a) Eventos en Poliformat y días con sesiones iniciadas por estudiante.



(b) Número de tareas y tests entregados por asignatura y estudiante.



(c) Visitas a Poliformat y tiempo total por estudiante.



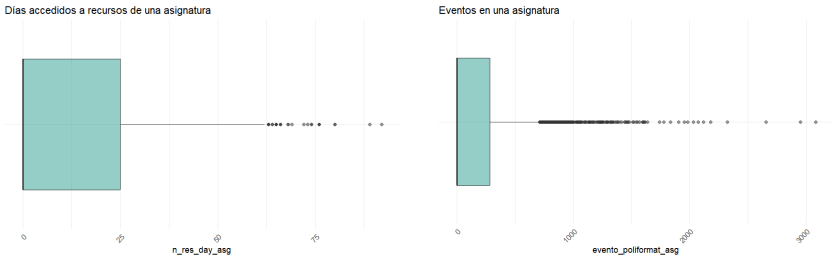
(d) Accesos y acciones sobre recursos por estudiante.

Figura 3.17: Evolución del uso de Poliformat por estudiante en distintas métricas.

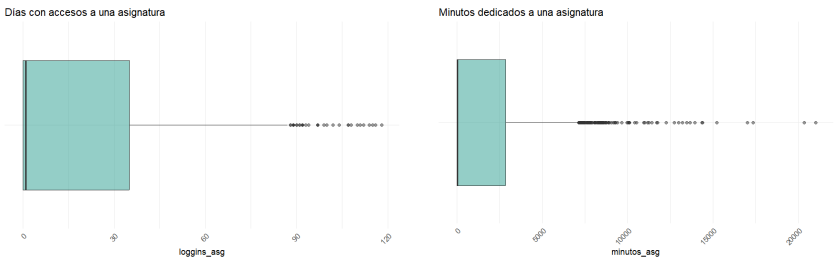
Las siguientes variables de interés son las generadas como suma de todo el interés del estudiante en todas sus asignaturas, que se pueden ver en las gráficas de la figuras 3.17. Se puede observar, al igual que en las gráficas de la figura 3.16 que los meses de mayor actividad fueron octubre y diciembre, no siendo también destacable enero en varios estudiantes.

Acabando con los datos que hay de cada estudiante en total por ámbito informático y asignatura, se disponen de las figuras 3.18 y 3.19. Rápidamente puede ver en sus dia-

gramas de caja y bigotes que, al menos la mitad de los datos (como indica la mediana) se encuentra en 0. Esto indica que, aunque se hayamos eliminado las asignaturas del segundo cuatrimestre (las cuáles tenían 0 actividad) y habiendo sumado todos los meses, muchos estudiantes no han realizado ninguna de las acciones que el sistema monitoriza.

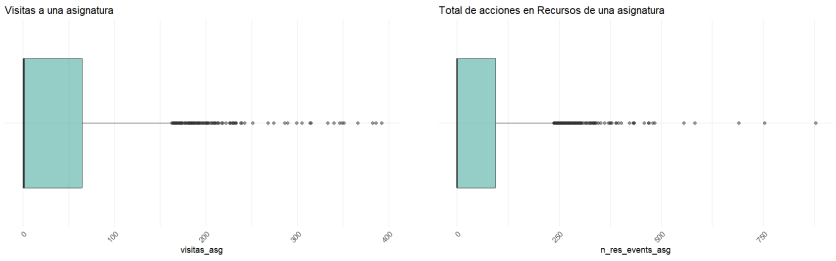


(a) Días accedidos a recursos. **(b)** Eventos registrados en Poliformat.



(c) Días con acceso a la asignatura. **(d)** Minutos dedicados a la asignatura.

Figura 3.18: Indicadores de actividad digital (I) en Poliformat relacionados con una asignatura concreta. Se analizan accesos, eventos y tiempo.



(a) Visitas a la asignatura. **(b)** Acciones en recursos.

(c) Tareas entregadas. **(d)** Tests entregados.

Figura 3.19: Indicadores de actividad digital (II) en Poliformat relacionados con una asignatura concreta. Se recogen visitas, acciones en recursos y entregas.

En cuanto a los datos diferentes a 0, nos encontramos con muchos diagramas de caja y bigote cuya caja (a partir de la mediana) no abarca demasiado rango, mientras que los

valores posteriores y los valores más atípicos (gran actividad total) se reparte por todo el rango.

Los casos más atípicos serían las tareas y los tests entregados. Una gran mayoría de los datos están en torno a 0, siendo valores atípicos los que se encuentran por encima. Esto indica que la mayoría de las asignaturas no utilizan muchos test o muchas tareas, o bien que la no los entregan.

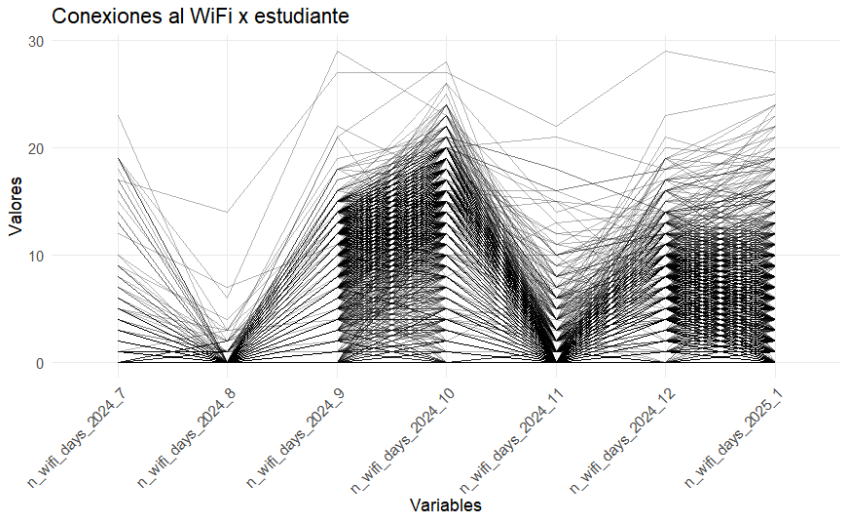


Figura 3.20: Conexiones al WiFi de la UPV de los estudiantes

Además, en la figura 3.20 es posible estudiar las conexiones a la red de la UPV por parte de los alumnos de la ETSINF. Se puede ver como había muchos datos en julio, posiblemente de alumnos que tuviera que asistir a una convocatoria extraordinaria, seguido de un bajón en agosto por el cierre de la universidad. En septiembre y octubre vuelven a crecer las conexiones, indicando que vuelve la actividad normal a la institución, pero se encuentra con un descenso muy notable por las consecuencias de un fenómeno natural que tuvo lugar en Valencia y obligó a suspender las clases presenciales de ese mes y parte de diciembre: la DANA. Finalmente, enero también tiene niveles similares, ya que no es periodo lectivo y los alumnos asisten a los exámenes y a estudiar en la universidad.

Finalmente y al haber tantas variables, solamente es posible ver el Análisis de componentes principales del conjunto de datos de Poliformat, sin su matriz de correlación.

A juzgar por la figura 3.21, se puede afirmar que hay una alta correlación entre las variables del dataset, pues una sola dimensión puede explicar más del 50 % de la varianza. Sin embargo, esto también implica mucha redundancia, ya que la mayoría de variables miden el mismo comportamiento o similares, de manera que esta técnica resulta especialmente útil.

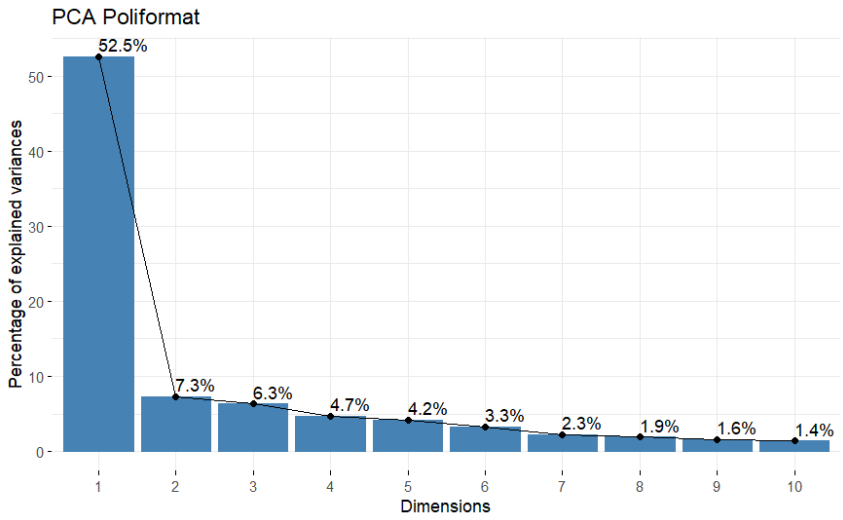


Figura 3.21: PCA de las variables de poliformat

3.6.4. Análisis del conjunto de datos de abandono

Este dataset tiene 52 filas y 5 variables. Este dataset es muy especial, ya que contiene qué estudiantes han abandonado la carrera, cuántas asignaturas han dejado y cuándo. El resumen se puede ver en la tabla 3.10.

Este es un dataset crucial para el estudio, ya que contiene los datos de los estudiantes que han dejado la carrera, cuántas asignaturas y cuándo lo hicieron. Por sí solo no da mucha información, pero el poder concatenarlo cuando sea necesario le da mucho valor.

Variable	**Media**	**SD**	
asi_left	10.04	2.91	
abandono	1	0	
Variable	**Mín Fecha**	**Mediana**	**Máx Fecha**
baja_fecha	2024-07-19	2024-09-04	2025-01-28
Variable	**Únicos**	**Top valores**	
mes	5	jul: 23, sep: 20, oct: 5, enero: 2	

Tabla 3.10: Resumen de variables de abandono

3.6.5. Correlaciones y PCA entre diferentes datasets

En esta última sección del capítulo, se analizarán las relaciones entre las diferentes variables. Se hará un análisis de correlaciones a los datasets de datos académicos y sociodemográficos, así como Análisis de Componentes Principales para los datasets “sociodemografia” y “academicas” y para los tres datasets en conjunto.

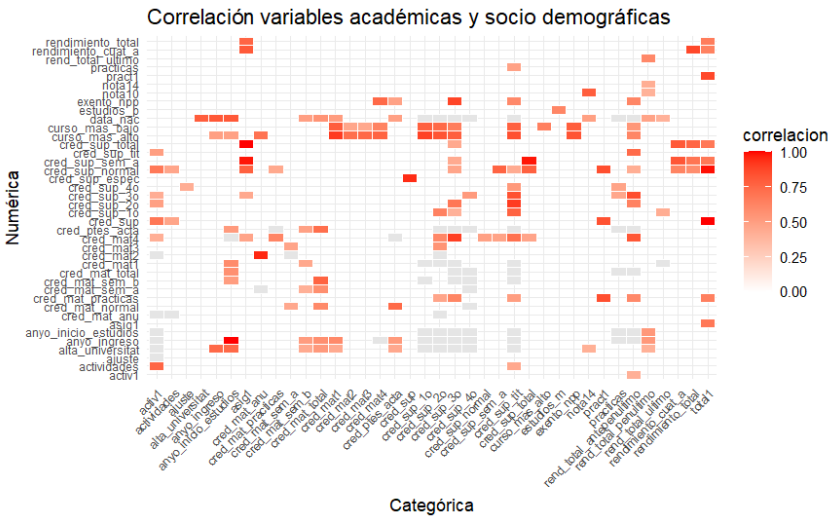


Figura 3.22: Análisis de componentes principales entre variables sociodemográficas y académicas

En primer lugar, se muestra un análisis de correlación entre las variables académicas y sociodemográficas en la figura 3.22.Se identifican correlaciones y significativas entre variables como nota10, nota14, anyo_ingreso o alta_universitat y indicadores académicos como rendimiento_total, cred_sup o curso_mas_bajo, sugiriendo que factores como la edad de ingreso, el nivel educativo familiar o las calificaciones previas influyen en el rendimiento académico.

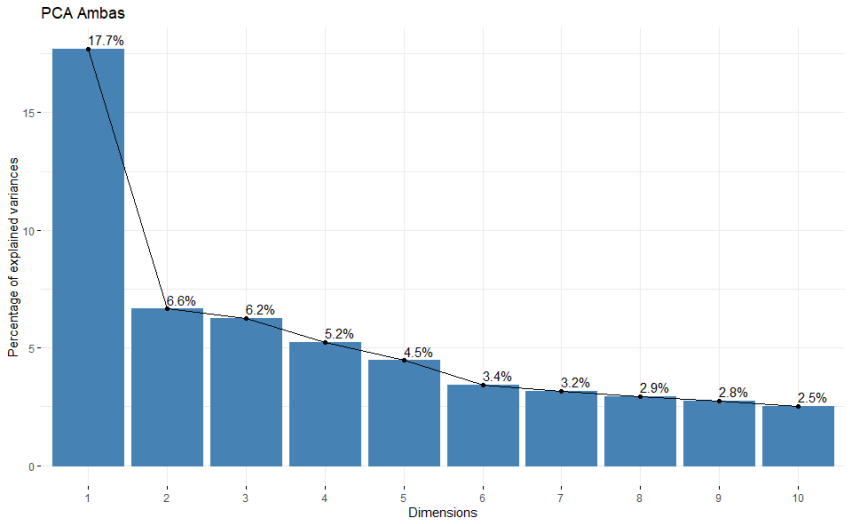


Figura 3.23: PCA de la combinación de datos sociodemográficos y académicos.

En la figura 3.23, se puede observar que, tal como se esperaba viendo la matriz de correlación, existe cierta relación entre las diferentes variables. La primera componente es capaz de capturar un 16 % de la varianza, mientras que las demás bajan al 7 % y al 5 %. La variabilidad, al igual que en las anteriores figuras de PCA, no se encuentran concentradas en una dimensión, si no que se reparten en diferentes factores, reflejando la complejidad de los estudiantes y sus perfiles académicas y sociodemográficos.

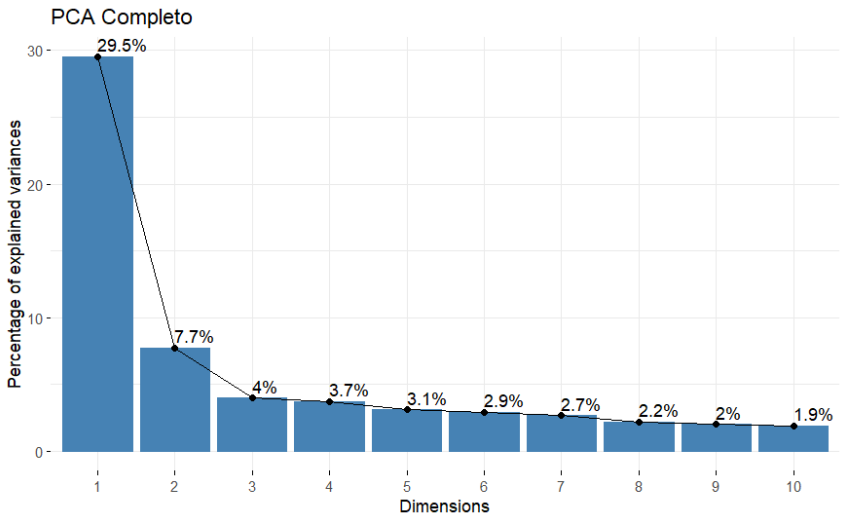


Figura 3.24: PCA del dataset completo.

La figura 3.24 muestra un PCA en el cual están integrados los tres datasets principales: "sociodemogracia", "academicas" y poliformat", teniendo información sobre las características acádemicas, sociodemograficas y de actividad digital de todos los estudiantes. La primera componente explica un 26 %, un valor algo más alto que el resto de análisis, debido probablemente a la fuerte correlación entre variables del entorno digital. Las demás componentes explican menos varianza, pero pueden resultar también importantes para futuros análisis o predicciones.

CAPÍTULO 4
Caracterización del alumnado en relación al abandono

Una vez que ya se ha visualizado la muestra entera, el estudio va a centrarse en los estudiantes que han abandonado, sus características en diferentes ámbitos y las variables que más influyan, así como diferentes test estadísticos usando PERMANOVA para encontrar diferencias entre el alumnado que abandona y el que no.

4.1 Contextualización del abandono a partir de la muestra

En primer lugar, se repasará la cantidad de estudiantes que han abandonado en la carrera de informática este año, vista previamente en la figura 3.10. Se disponen de 52 estudiantes que han dejado la carrera y de los que se disponen datos de este primer cuatrimestre.

Julio	Septiembre	Octubre	Diciembre	Enero
23	20	5	2	2

Tabla 4.1: Abandonos por mes

En la tabla 4.1 es posible visualizar en qué meses la gente tiende a abandonar sus estudios univertarios, siendo los más destacables julio y septiembre.

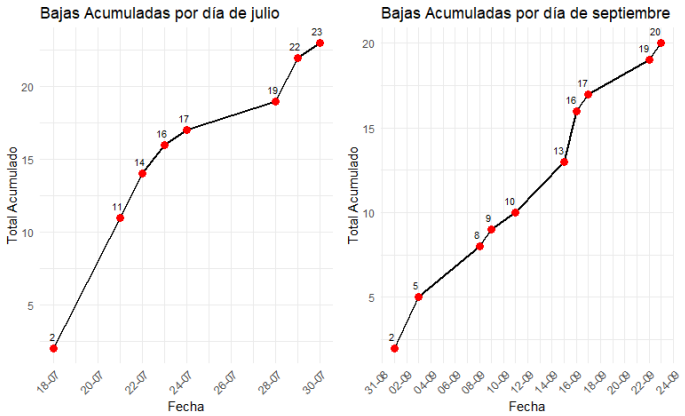


Figura 4.1: Evolución de bajas en julio y septiembre

En la figura 4.1 se dispone de la evolución temporal de las bajas por matrícula. En primer lugar, se puede ver cómo los alumnos han dejado la carrera después de realizar la automatrícula. Los estudiantes de nuevo ingreso realizan la automatrícula del 17 al 18 de junio, mientras que los estudiantes antiguos la hacen del 18 al 22 de junio. Aquí se puede observar como el día 21 hay 19 nuevas bajas, que aumenta progresivamente hasta las 17 el día 24.. Tras eso, hay un parón, volviendo a aumentar el día 28, hasta las 23 bajas.

En septiembre, se puede observar como es más progresivo. Las clases empiezan el día 9 de septiembre, teniendo lugar las jornadas de acogida de la UPV los días 5 y 6 del mismo mes, donde los estudiantes de nuevo ingreso reciben información sobre la carrera por parte del programa PIAE+. Hasta el día antes del inicio de las clases, hay 8 desmatriculaciones. Posteriormente, las bajas aumentan de manera más progresiva, indicando que los estudiantes están tratando de adaptarse a la carrera, sin mucho éxito. Una hipótesis plausible sería la espera para entrar a diferentes carreras que tuvieran prioridad para esos estudiantes.

Adicionalmente, se sabe que, de media los estudiantes se matriculan de, aproximadamente, 10 asignaturas. Por lo tanto, si abandonan la carrera, es habitual que los datos de asignaturas abandonadas estén alrededor de ese valor, tal como se puede observar tanto en la tabla A.3, donde se puede ver por mes, como la anteriormente vista con el resumen de todo el dataset, la tabla 3.10.

4.2 Influencia de los factores sociodemográficos

A continuación, el trabajo volverá a centrase en los factores sociodemográficos y en la infuencia que esta puede tener sobre el abandono universitario en la carrera de informática.

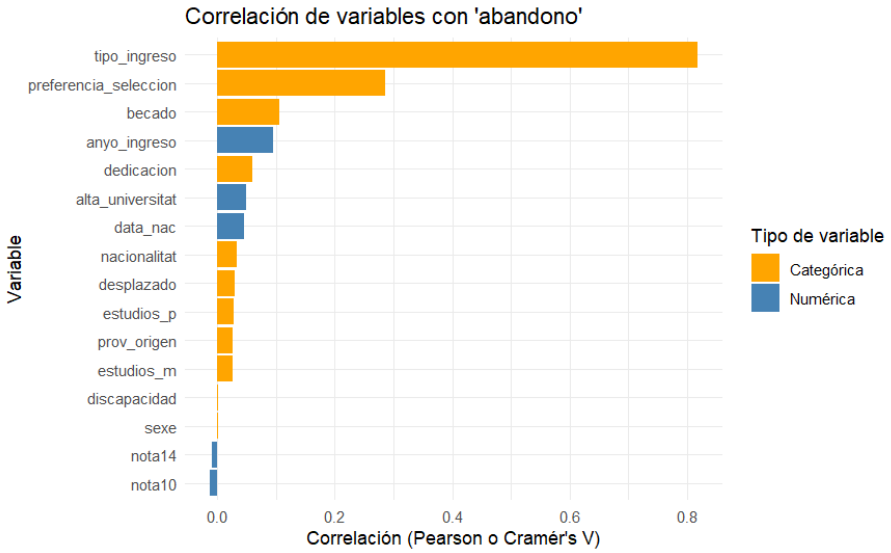


Figura 4.2: Correlación de los factores sociodemográficos con el abandono universitario

En primer lugar, se pueden examinar las diferentes correlaciones con la variable abandono en la figura 4.2. Lo que más llama la atención es la correlación del tipo de ingreso. Sin embargo, no es de extrañar su alta correlación, pues una de las opciones es la baja de la matrícula (el valor "BMA"). Sin embargo, se puede ver en la tabla 4.2 que la mayoría de alumnos que abandona se distribuyen entre alumnos que escogieron Ingeniería Infor-

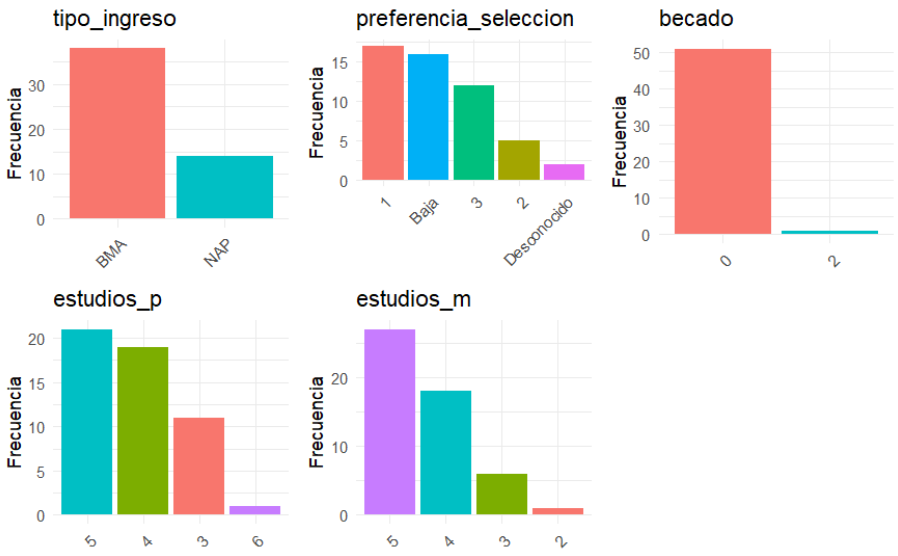


Figura 4.3: Caracterización de las variables categóricas de los estudiantes que abandonan

mática como primera opción (tal vez no satisfechos con la carrera), en tercera selección, y distribuidos de en preferencia baja, con dos personas cuyo dato es desconocido.

Preferencia de selección	1	2	3	Baja	Desconocido
No abandona	1250	230	99	69	54
Abandona	17	5	12	16	2
Desconocido	0	0	0	0	0

Tabla 4.2: Distribución del preferencias de selección de carrera según abandono

Aunque es mucho más pequeña, se puede apreciar con cierta correlación con algunas variables, aunque ya es mucho menor, como el año de ingreso, si tiene beca o el año de ingreso. Sin embargo, hay otras variables completamente irrelevantes, como el sexo, la discapacidad o las notas obtenidas en selectividad para el abandono académico.

Viendo la figura 4.3, es posible apreciar características comentadas anteriormente sobre la correlación. Si nos fijamos en la preferencia de selección, se puede ver como muchos estudiantes no tienen como prioridad entrar a la carrera de informática, siendo su tercera opción o una opción muy baja, aunque también se puede apreciar muchos alumnos que sí tenían la carrera como primera preferencia.

Por otro lado, se puede ver que los estudios de los padres permanecen de manera muy similar a todos los estudiantes, aunque los estudios de nivel 4 de los padres es algo mayor que el de las madres en este caso. También se puede apreciar que, prácticamente, todos los estudiantes que abandonan no están becados, lo que puede significar que la beca es un incentivo para continuar con los estudios.

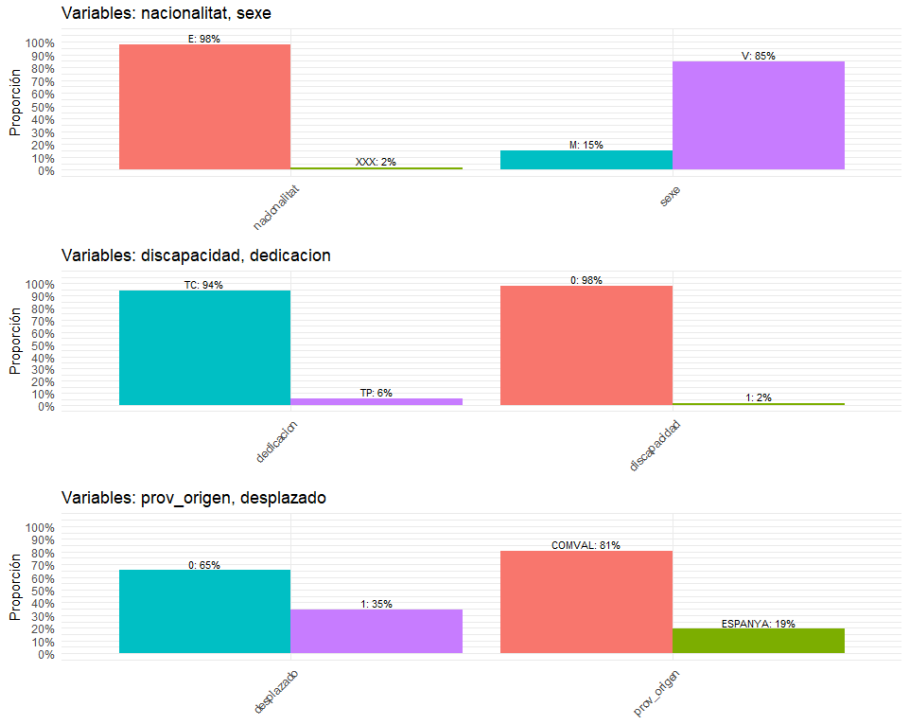


Figura 4.4: Caracterización de las variables binarias sociodemográficas de los estudiantes que abandonan

Continuando con variables no numéricas, en la figura 4.4 se puede apreciar que apenas hay cambios con respecto al alumnado que no ha abandonado la carrera, comparando con la figura 3.3. Prácticamente todos los que abandonan la carrera pertenecen a España al tener un porcentaje del 98 . Se puede ver como hay una o dos personas más que tienen la matrícula a tiempo parcial, así como un ligero aumento en personas desplazadas y provenientes de la Comunitat Valènciana. El sexo y la discapacidad no tienen ningún cambio.

No se esperaban demasiados cambios por parte del muchas variables, pues, como ya se sabía gracias a las correlaciones descritas en la figura 4.2, no tienen demasiada influencia en los alumnos que dejan inconclusa la carrera.

Fuente	Df	Suma de cuadrados	R^2	F	p-valor
Abandono	1	339.7	0.0117	20.037	**0.001**
Residual	1694	28718.1	0.9883		
Total	1695	29057.8	1.0000		

Tabla 4.3: Resultados del análisis PERMANOVA sobre las variables sociodemográficas (distancia euclídea, 999 permutaciones).

Finalmente, se ha realizado un análisis utilizando el método PERMANOVA, en el cual se han introducido las siguientes variables: “nacionalitat”, “data_nac”, “alta_universitat”, “prov_origen”, “anyo_ingreso”, “tipo_ingreso”, “dedicacion”, “desplazado”, “becado”, “preferencia_seleccion” y “abandono”.

Este análisis, cuya información se encuentra resumida en la tabla 4.3, sí existe una diferencia entre los estudiantes que abandonan y los que no, con un p-value menor al 0.001. Sin embargo, el modelo solo es capaz de explicar una mínima parte de la varianza,

siendo poco más del 1 %. Esto nos puede indicar que existen diferencias pequeñas, o debidos a otros factores.

4.3 Diferencias según el perfil académico

En esta sección, se tratará de encontrar diferencias a nivel académico entre los diferentes alumnos que abandonan y los que no.

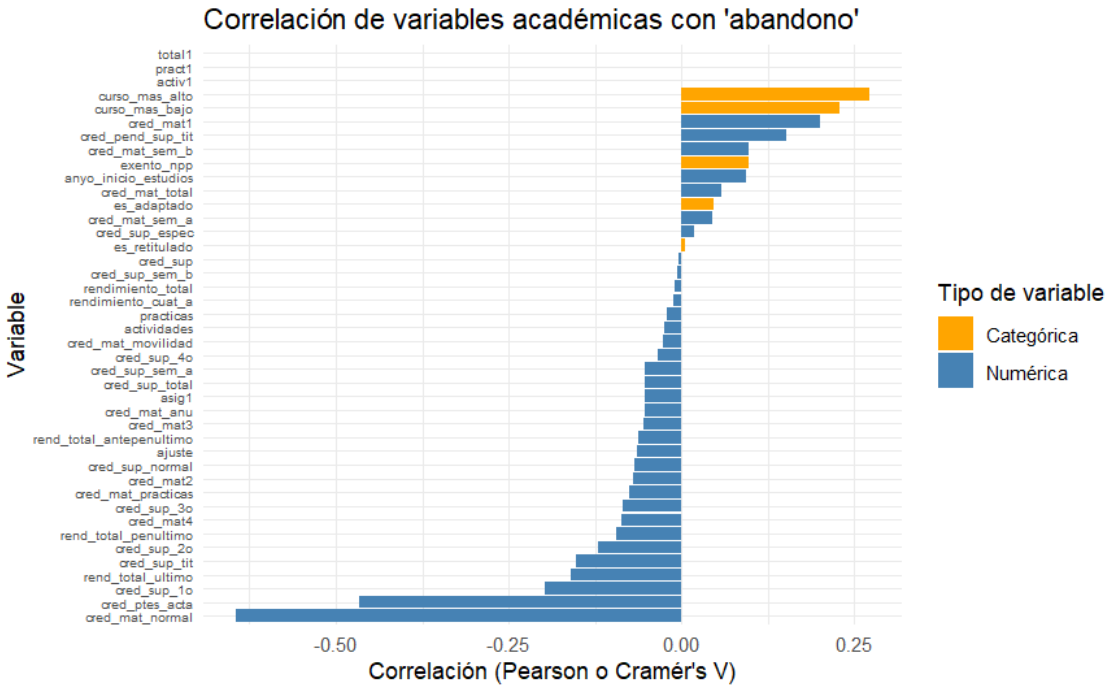


Figura 4.5: Correlación de los factores académicos con el abandono universitario

Al igual que en la anterior sección, se comienza con el análisis de correlación de la variable abandono con el resto de variables académicas del alumnado, que se puede ver en la figura 4.5. Se observa que las variables con mayor correlación positiva son las variables de los cursos en los que están matriculados, lo cual podría dar una aproximación a los cursos en donde el estudiantado decide interrumpir la formación académica. Hay correlación, como era natural, con variables como las matriculaciones y los créditos pendientes de superar. Aunque pequeña, existe cierta correlación con la variable "exento_npp", que indica que al alumno le quedan menos del 25 % de créditos por superar excluyendo el Trabajo de Fin de Grado. Esto podría ser un indicativo de que hay gente que abandona la carrera o bien en cuarto, o bien finalizando tercero de carrera.

Ahora bien, comenzamos a ver correlaciones más negativas entre la mayoría de variables. Las más pequeñas pero significativas corresponden a los créditos superados, lo cual es entendible teniendo en cuenta que, o bien no los superan o bien se desmatriculan tras no haberlos superado, así como los créditos matriculados de cursos posteriores. Además, de manera más variada, se pueden encontrar las variables referidas a créditos superados mediante convalidaciones y prácticas, indicando que el estudiante no está el suficiente tiempo en la universidad (o no tiene interés) para superar créditos por esas vías.

Cabe destacar que también son llamativas las variables de rendimiento, pues son el claro indicativo de que la correlación entre las asignaturas no superadas y el abandono

de la carrera. Destaca el rendimiento total del último curso, pues realmente es donde en más estudiantes puede darse el caso, al disponer de tres años enteros.

Tabla 4.4: Distribución cruzada entre dos variables académicas según abandono

Abandono = 0

	1º	2º	3º	4º
1º como curso más bajo	388	138	16	4
2º como curso más bajo	0	270	191	89
3º como curso más bajo	0	0	189	160
4º como curso más bajo	0	0	0	257

Abandono = 1

	1º	2º	3º	4º
1º como curso más bajo	48	0	2	0
2º como curso más bajo	0	1	1	0
3º como curso más bajo	0	0	0	0
4º como curso más bajo	0	0	0	0

A continuación, estudiaremos los cursos en los que está matriculada la gente que abandona la carrera a partir de las variables "curso_mas_alto" y "curso_mas_bajo". A la vista de los resultados visibles en la figura 4.4, se puede claramente ver que los estudiantes que dejan la carrera son estudiantes de primer curso. Esto puede significar que lo estudiantes están desencantados con la carrera en su primer año, tanto al acabar el año como abandonándola en septiembre.

Se han examinado las demás variables de manera similar al análisis exploratorio. Sin embargo, muchos datos no son demasiado relevantes. Sí que hay puntos bastante claves, como que, originalmente, muchos estudiantes se han matriculado previamente en 60 créditos por curso: 30 en el semestre A y 30 en el semestre B. Esto puede observarse an la figura A.6b. También es interesante destacar que muchos no tienen datos académicos referidos a créditos superados e incluso matriculados, como se puede ver en las figuras **??** y A.7a.

Nuevamente y como cierre de esta sección, se ha utilizado el método PERMANOVA para ver si los alumnos son significativamente diferentes.

Fuente	Df	Suma de cuadrados	R^2	F	p-valor
Abandono	1	875373	0.0325	58.755	**0.001**
Residual	1752	26102606	0.9676		
Total	1753	26977979	1.0000		

Tabla 4.5: Resultados del análisis PERMANOVA sobre las variables académcias.

Como se puede ver en la tabla 4.5, nuevamente resultan significativamente diferentes los estudiantes que abandonan de los que no, dado el p-valor menor a 0.001. Ahora bien, de la misma manera que anteriormente, el modelo solo puede explicar un 3 % de la varianza, una parte muy pequeña del total. Sin embargo y, al igual que las variables sociodemográficas, las diferencias son pequeñas y/o debidos a factores que no se están teniendo en cuenta.

4.4 Comportamiento digital y abandono académico

En la siguiente sección, se tratará de analizar el comportamiento digital de los estudiantes que abandonan, buscando diferencias significativas entre unos y otros.

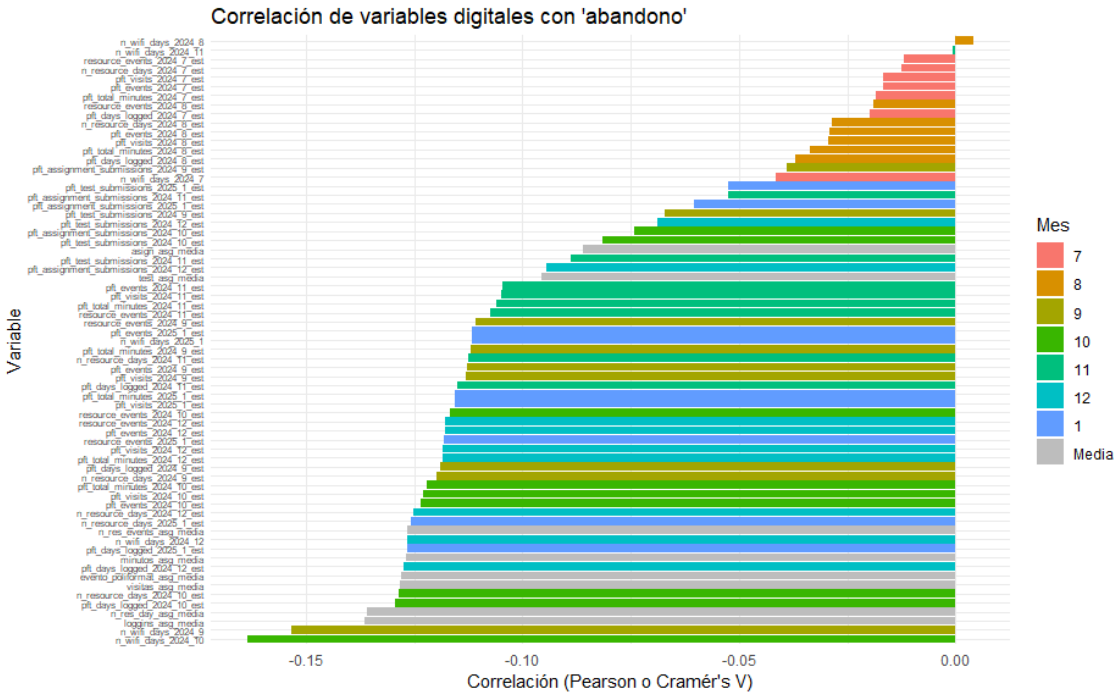


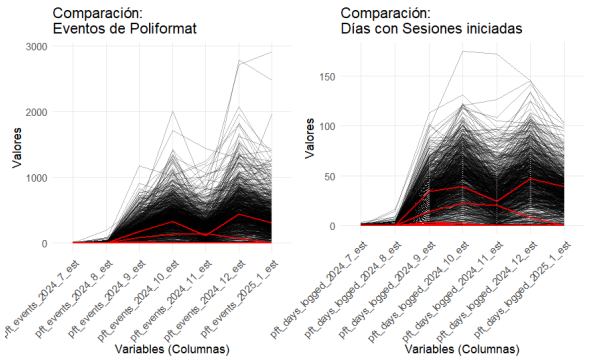
Figura 4.6: Correlación de los factores digitales con el abandono universitario

En la Figura 4.6 se presentan las correlaciones entre las distintas variables de comportamiento digital en Poliformat y la variable "abandono". En este caso, se ha coloreado al mes que pertenecen, para un mejor reconocimiento de los periodos en los que so nrelevantes algunos factores.

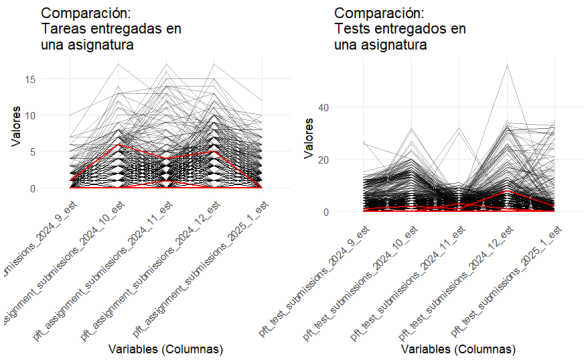
De un vistazo, se puede observar claramente que hay una correlación negativa por parte de casi todas las variables. Esto refuerza la teoría de que, cuanto menos se use el entorno digital, más probable es que se abandone la carrera. Las mayores correlaciones se encuentran en el número de conexiones al wifi en el mes de octubre (el segundo mes) y la media de acceso a recursos por estudiante.

En la figura, también se han coloreado las variables en función del mes al que pertenecen, con la excepción de las variables que son la media de los estudiantes. Continuando con lo último mencionado, se puede ver que muchas de las variables de las medias son relevantes para el abandono académico, siendo las más importantes "loggins_asg" y "n_res_day_asg", las cuáles indican el número de días que se entra al sitio de la asignatura y el número de días que accedió al sitio recursos de la asignatura.

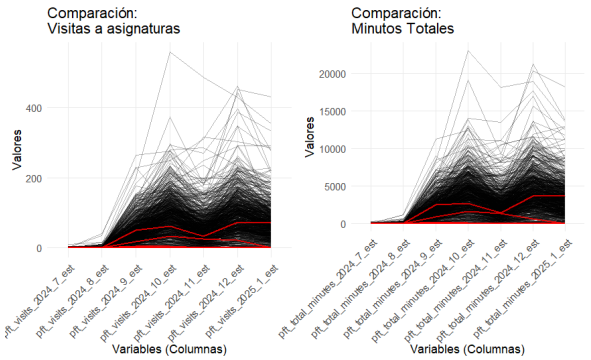
También se puede observar cómo los meses previos al inicio del curso (julio y septiembre) están poco correlacionados, mientras que los meses más centrales del primer cuatrimestre parecen ser los más importantes, mientras que las de diciembre y enero son algo menores. Esto puede indicar que el abandono se materializa a media curso, lo cuál es algo que ya comprobamos en la tabla 4.1.



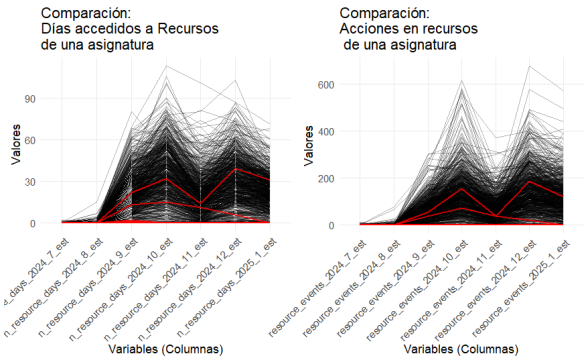
(a) Eventos en Poliformat y días con sesiones iniciadas por estudiante.



(b) Número de tareas y tests entregados por asignatura.



(c) Visitas a Poliformat y tiempo total en minutos por estudiante.



(d) Accesos a recursos y número de eventos en ellos.

Figura 4.7: Evolución del comportamiento digital en Poliformat por estudiante como suma de las asignaturas. Las líneas en rojo representan a quienes abandonan.

A continuación, se pueden observar las distintas líneas de actividad digital y la comparativa entre estudiantes normales y estudiantes que abandonan en la figura 4.7, donde las líneas rojas representan a los estudiantes que han dejado los estudios universitarios. La gente que dejar de ser estudiante tiene, de base, pocas interacciones, pero llaman la atención los valores de acciones y accesos a recursos de las asignaturas, donde hay algunas personas con valores más altos. Estas personas pueden ser aquellas que se dieron de baja posteriormente a septiembre, habiendo tratado de salir adelante con la carrera y no haber tenido éxito.

Se puede apreciar, además, que en la mayoría de variables hay muchas líneas decrecientes, lo que claramente se relaciona con la progresiva desconexión digital del alumnado. En variables como *minutos totales en la plataforma* o *días con sesiones iniciadas*, la diferencia con respecto al resto del alumnado se amplía mes a mes, lo que indica que el abandono suele estar precedido por un proceso de desconexión gradual. Este patrón decreciente sugiere que, en muchos casos, el abandono no es abrupto sino progresivo, reflejando una pérdida de implicación digital antes de la desvinculación formal. No obstante, también se detectan perfiles de estudiantes que, pese a abandonar, registran un volumen notable de actividad en los primeros meses, lo que podría indicar intentos fallidos por continuar los estudios. Destacamos, nuevamente, la excepcionalidad del mes de noviembre con motivo del fenómeno natural de la DANA.

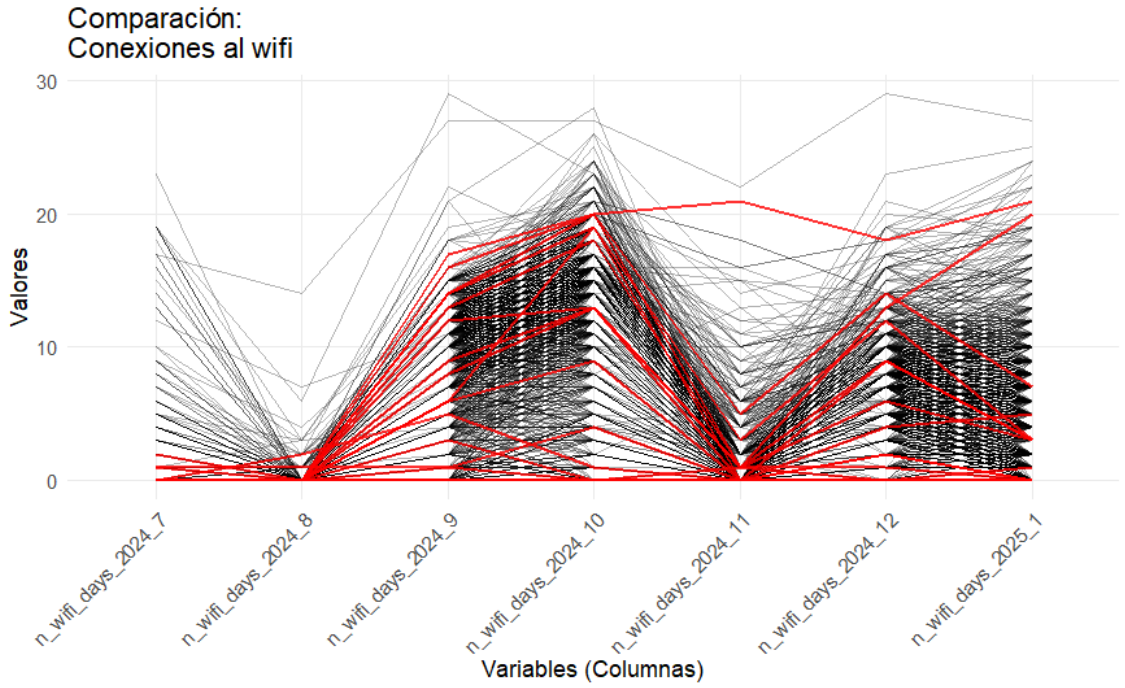


Figura 4.8: Comparación del uso de la conexión a la UPV por estudiante, marcando al alumnado que abandona en rojo.

Siguiendo con el uso de la red de la universidad, se puede observar en la figura 4.8 un patrón similar. Se puede ver como, en septiembre, hay muchas conexiones a la red WiFi, siendo muy similares a los de alumnos normales, aunque en menor medida. Que siga habiendo un número alto de conexiones podría indicar que muchos de esos alumnos siguen yendo a la UPV aun habiendo abandonado los estudios en la ETSINF. Esto podría ser síntoma de que muchos estudiantes no abandonan los estudios universitarios, sino que estos se cambian a otros diferentes dentro de la universidad.

Adicionalmente, se puede comprobar que los valores en las medias de las variables de comportamiento digital son prácticamente 0, indicando una nula relevancia a excepción de ciertos valores atípicos, como se puede comprobar en la figura A.7b.

Finalmente y, al igual que en los anteriores apartados, se procederá a un análisis PERMANOVA para comprobar si los alumnos son significativamente diferentes, algo que sería esperable dadas sus diferencias en la actividad.

Fuente	Df	Suma de cuadrados	R^2	F	p-valor
Abandono	1	3.6862e+09	0.0148	26.357	**0.001**
Residual	1752	2.4503e+11	0.9852		
Total	1753	2.4872e+11	1.0000		

Tabla 4.6: Resultados del análisis PERMANOVA sobre las variables de comportamiento digital.

Se puede observar que, de manera idéntica a los anteriores modelos, el modelo detecta diferencias significativas con un p-value menos a 0.001, pero no llega al 1.5 % de varianza explicada.

4.5 Variables relevantes y estructura del alumnado

Tras analizar y comparar los diferentes conjuntos de datos de los que se disponen y su relación con el abandono académico en la carrera de informática en el anterior capítulo, han sido seleccionadas las variables que se consideran más relevantes y que pueden anticipar el abandono en el grado. Han sido elegidas en función de la capacidad para explicar el abandono, su presencia y relevancia en los análisis exploratorios y la contribución a modelos de reducción de la dimensionalidad. Las variables seleccionadas han sido las siguientes:

- **Mes del abandono**: este mes es clave para entender los momentos críticos en los que se producen las bajas, ya sean en julio o septiembre.
- **Preferencia de selección**: este recoge el orden de preferencia, que resulta ser clave al existir alumnos que siguen registrando conexiones a la red de la UPV y que han cambiado de carrera. Esto pone de manifiesto la motivación que el alumno pudiera tener al comienzo de los estudios y la propensión a abandonar la carrera.
- **Becado**: la condición de becado se ha relacionado con un menor riesgo de abandono. Además, refleja el contexto socioeconómico del estudiante, factor clave en la permanencia universitaria.
- **Curso más alto y más bajo matriculado**: Estas variables indican el curso en el que los estudiantes se han matriculado. La mayoría son de primero, pero existen alumnos matriculados en otros cursos.
- **Año de inicio de estudios**: Nos permite saber si hay alguna relación entre el año de iniciar los estudios con el reciente abandono.
- **Rendimientos académicos (si los hay)**: el rendimiento total o por cuatrimestre se ha demostrado como un indicador directo de riesgo de abandono. Un bajo rendimiento suele anteceder la baja.
- **PCA de las variables digitales más interesantes**: dada la alta dimensionalidad del dataset "poliformat"(68 variables), se optará por usar su Análisis de Componentes Principales como reducción de su dimensionalidad, ya que puede llegar a sintetizar de manera completa y suficiente la información de comportamiento digital e implicación académica.

Estas variables han sido preferidas frente a otras por distintas razones:

- Existen muchas variables que no tienen demasiada variabilidad ni relevancia lo que reducía su valor de cara a futuros análisis.
- Otras, como variables altamente correlacionadas (*anyo_ingreso*, *alta_universitat*, *data_nac*), ya se ven representadas con una sola variable.
- Finalmente, variables correlación ni aparente relevancia con el abandono académico.

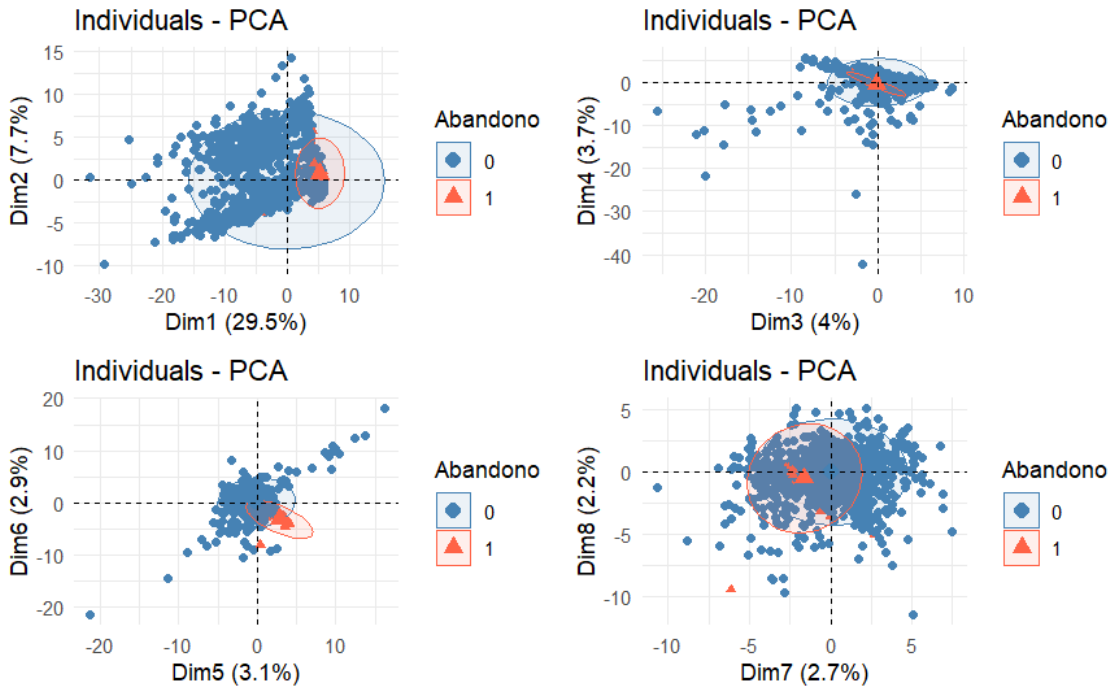


Figura 4.9: Proyección de los estudiantes en el espacio PCA según su pertenencia al grupo de abandono.

Finalmente, se puede ver en la figura 4.9 se puede observar la proyección de los estudiantes en el espacio del Análisis de Componentes Principales realizado en el capítulo anterior, en la figura 3.24. A pesar de haber leves patrones de separación, existe un alto grado de solapamiento entre los alumnos que abandonan y los que no. Esto puede anticipar que los modelos lineales tendrán dificultades para distinguir ambos grupos, ya que no existen relaciones lineales que puedan separar ambos grupos de manera clara.

CAPÍTULO 5
Identificación de perfiles estudiantiles con técnicas de clustering

Tras haber comparado las diferentes características que pueden permitir identificar a los estudiantes con mayor probabilidad de abandono, el siguiente paso es tratar de encontrar sus perfiles mediante técnicas de agrupamiento no supervisado: clustering. El objetivo del clustering será tratar de encontrar patrones comunes en estudiantes o grupos con comportamientos similares que puedan estar relacionados con el abandono académico.

5.1 Preparación de los datos y técnicas de agrupamiento

Con el fin de identificar los perfiles de grupo, se van a realizar diferentes operaciones con el fin de obtener un dataset apropiado para el clustering. En primer lugar y, a la vista de los resultados de la anterior sección, se ha realizado un Análisis de Componentes Principales a los datos del dataset "poliformat" de los estudiantes que han abandonado la carrera. Observando el resultado en la figura A.8, se han decidido añadir las tres primeras componentes, ya que explican un 96 % de la variabilidad.

Una vez hecho esto, se han concatenado los datasets "academicas" y "sociodemografia", pero solamente los estudiantes que han dejado la carrera. Además, se ha añadido el mes de abandono. Una vez hecho esto, se ha realizado la técnica de One-Hot-Encoding para poder representar las variables categóricas de manera efectiva y, con ello, se han escalado los datos y concatenado el PCA de "poliformat"

Al echar un vistazo a los datos finales, se puede ver que todavía es necesario limpiarlo. Se han realizado los siguientes cambios:

- Se ha eliminado la variable "actividades" porque contenía nulos para todos los estudiantes, lo que no aportaba valor y hacía necesaria su eliminación
- La variable "nota14" presentaba un valor faltante, de manera que, para preservar esta variable, se imputó su valor ausente utilizando la mediana de la variable.
- Las variables de rendimiento académico histórico ("rend_total_ultimo", "rend_total_penultimo" y "rend_total_antepenultimo") se han eliminado al haber una gran proporción de estudiantes que no disponían de datos de estas variables. Su altísima tasa de nulos hacía inviable la imputación por la mediana o métodos similares.

Finalizando con el tratamiento de datos, se concatena con el PCA de "poliformat", de manera que ya se puede para las técnicas que utilizaremos.

A continuación se explican brevemente las técnicas de agrupamiento que se utilizarán, así como el motivo de su uso. Las técnicas usadas son:

- **Método de Ward**: Es una técnica de clustering jerárquico y aglomerativo que trata de minimizar la varianza total dentro de cada grupo. Se encuentra útil debido a su capacidad para generar grupos que sean compactos y esféricos.
- **Método de la media** (*average linkage*): Es también un método jerárquico y aglomerativo, pero calcula la distancia entre todos los pares de elementos de cada grupo. Este método genera clústeres de manera más equilibrada debido al cálculo y uso de medias.
- **K-means**: Es un método de partición que divide los datos en K clústeres, tratando de minimizar la varianza entre individuos del mismo grupo. Es un método eficiente siempre y cuando los datos estén escalados y no haya valores atípicos.
- **PAM (Partitioning Around Medoids)**: Es un método parecido a K-means, pero es más robuste frente a valores extremos. En vez de usar medias, usa individuos reales como centro de los clústeres, los cuáles tienen el nombre de *medoides*. Es más costoso computacionalmente pero es capaz de soportar mejor el rudio.

Finalmente y, tras proceder previamente con todas las técnicas que hay a continuación, se encontraron dos valores atípicos muy separados de los clústeres que hacían muy difícil la interpretación. Por ello, se procedió a su eliminación. Es posible ver la determinación de clústeres en la figura 5.3 y los clústeres generados en la figura A.11 previos a la eliminación de los valores atípicos.

5.2 Evaluación de la agrupabilidad y determinación del número de clústeres

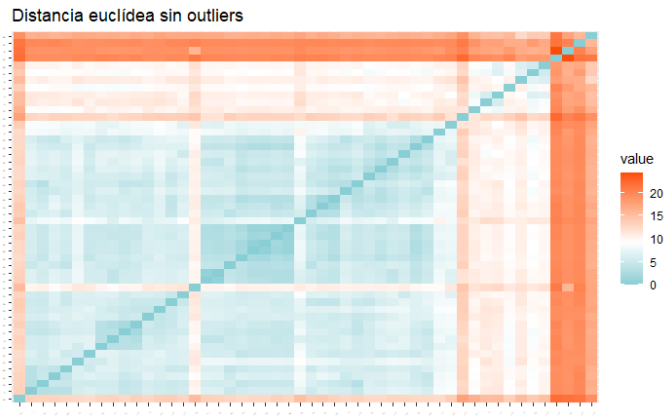


Figura 5.1: Mapa de calor de distancia euclídea entre estudiantes que dejaron la carrera tras eliminar outliers

Disponiendo de datos limpios y habiendo explicado y justificado las técnicas de agrupamiento a utilizar, se procede a ver la capacidad de agrupamiento que tienen los datos.

En primer lugar, se ha calculado la distancia existente entre los individuos usando la distancia euclídea, ya que es una métrica muy utilizada que funciona muy bien una vez los datos han sido estandarizados, como es el caso tras haber escalado. No se han utilizado otros métodos como la distancia de Manhattan o la Mahalanobis al haber solventado previamente problemas de multicolinealidad eliminando variables muy correlacionadas ("anyo_nacimiento" y ""anyo_inicio" con "anyo_ingreso").

En la figura 5.1, se puede observar que existen una gran cantidad de individuos similares, observables en el cuadrado de tonalidades azul claro. Dicha región indica similitud al haber poca distancia entre los diferentes individuos. Sin embargo, existen estudiantes alejados de otros, como se puede ver en la región de color rojo, especialmente 4, que seguramente se podrán identificar más adelante en el clustering. Aunque haya tonos rojizos, las distancias no parecen ser insalvables en la zona superior derecha, observando, incluso, tonos azules. También es posible observar el mapa de calor previo a la eliminación de los outliers, en la figura A.9.

Estadístico	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
Hopkins	0.7682	0.7767	0.7985	0.8042	0.8223	0.8603

Tabla 5.1: Resumen del coeficiente de silhouette tras la eliminación de valores atípicos.

Una vez examinadas las distancias mediante el mapa de calor, es necesario tratar de ver la similitud entre los diferentes individuos de manera analítica. Para ello, utilizaremos el estadístico de Hopkins, un estadístico que evalua la tendencia de agrupamiento de los datos, con un valor que varía entre 0 y uno.. Se puede ver en la tabla 5.1 que los estudiantes son bastante similares, obteniendo de media un 0.8 de media y no menos de un 0.76, muy cercano a 1. Esto significa que es coherente utilizar clustering, ya que existe una cierta tendencia al agrupamiento. También es posible revisar el coeficiente de clustering previo a la eliminación del los dos outliers en la tabla A.4.

Habiendo demostrado el sentido del análisis, el siguiente paso es encontrar el número apropiado de clústeres en cada método. Para ello, se utilizará el coeficiente de Sillhouette, una métrica que permite evaluar la calidad del agrupamiento dado un número concreto de clústeres. Se suele utilizar, además, con la suma de cuadrados intra-grupo para determinar el númeor óptimo de agrupaciones.

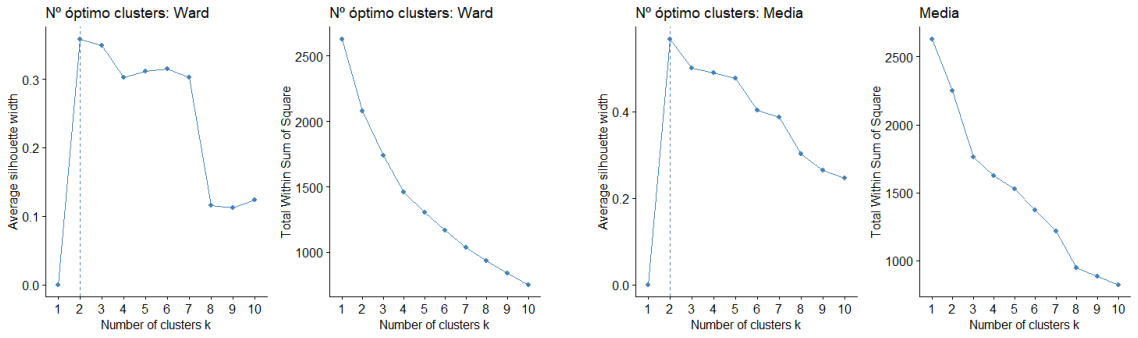
En este caso y tal como se puede ver en la figura 5.2, se han utilizado los dos métodos mencionados anteriormente para cuatro técnicas diferentes: Método de Ward, método de la media, K-means y PAM. La figura muestra una caída progresiva de la suma de cuadrados para todas las técnicas, con varias diferencias en el coeficiente de Sillouette. Todos los algoritmos coinciden en el mejor coeficiente lo ofrecen dos clústeres. Sin embargo, la suma de cuadrados es demasiado grande en la mayoría de casos, por lo que dependerá de cada técnica.

Para todas las técnicas a excepción del método de la media coinciden en que 4 clústeres o más supone una baja importante del coeficiente de Sillhouette, por lo que, en esas técnicas, no es conveniente elegir 4 clústeres. Sin embargo, en el método de la media parece razonable, ya que no supone una disminución demasiado grande del coeficiente a la vez que se disminuye la suma de cuadrados. Además, es el método que mayor coeficiente de Sillhouette alcanza, muy por encima de los demás, llegando al acercarse de manera significativa a 0.6 con dos clústeres.

Para K-means, es indicutible la selección de dos clústeres, pues un valor diferente supondría una caída significativa del coeficiente de Sillhouete, hasta el 0.1, donde el 0

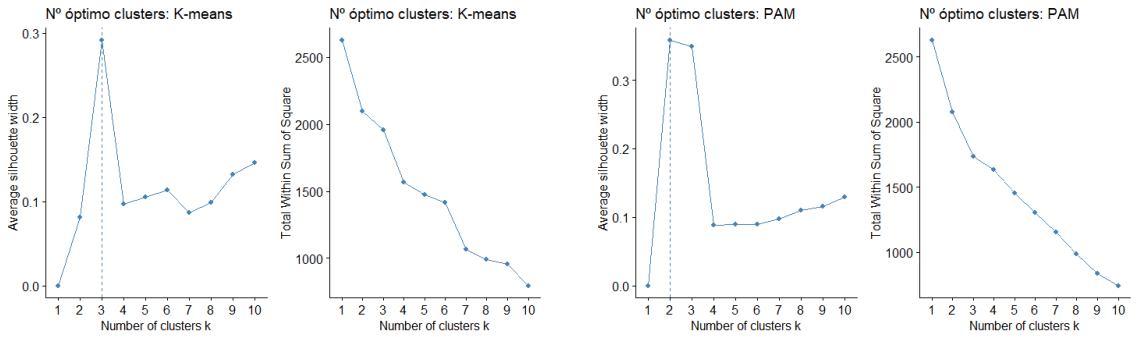
significa una clasificación aleatoria. De manera similar ocurre con el algoritmo PAM que, a partir de 3 clústeres, hay una caída importante en el coeficiente hasta el 0.1, por lo que puede ser razonable utilizar 3 clústeres, con un valor de suma de cuadrados similar al método de la media.

Finalmente, en el método de Ward parece que lo razonable también es utilizar 4 clústers. Dos clústeres tienen una suma de cuadrados alta, mientras que 3 parece más razonable y similar a otros métodos. Con 4 clústeres la suma de cuadrados baja significativamente, por lo que, aun perdiendo cierto coeficiente de Sillhouette, sea conveniente utilizar 4 clústeres.



(a) Número óptimo de clústeres con método Ward.

(b) Número óptimo de clústeres con método de la media.



(c) Número óptimo de clústeres con K-means.

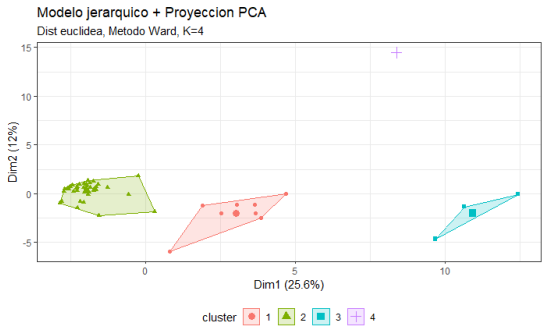
(d) Número óptimo de clústeres con método PAM.

Figura 5.2: Determinación del número óptimo de clústeres tras eliminar valores atípicos. Se muestran los resultados según los métodos del coeficiente de silhouette.

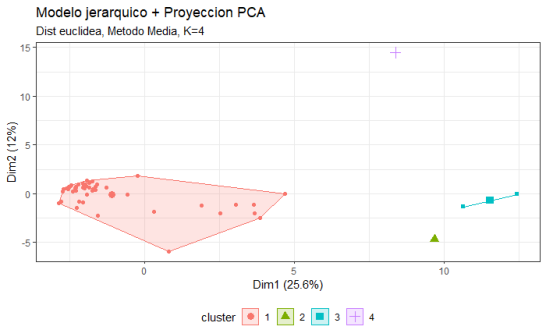
Por lo tanto, el número de clústeres decidido será el siguiente:

- **Método de Ward**: Se utilizarán **4** clústeres.
- **Método de la media** (*average linkage*): Se utilizarán **4** clústeres.
- **K-means**: Se utilizarán **2** clústeres.
- **PAM (Partitioning Around Medoids)**: Se utilizarán **3** clústeres.

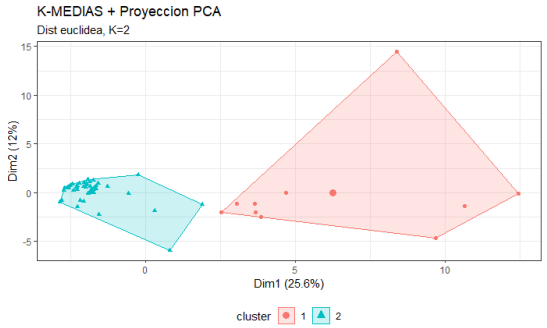
5.3 Comparativa de modelos de agrupamiento



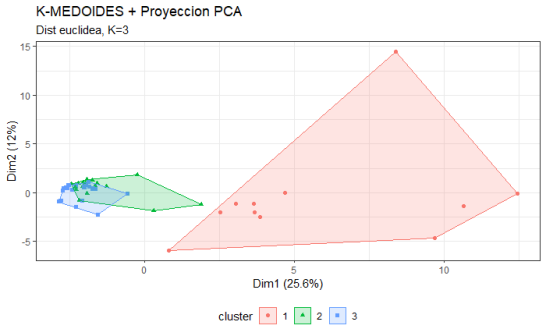
(a) Modelo jerárquico, método Ward, K=4.



(b) Modelo jerárquico, método de la media, K=3.



(c) K-means con distancia euclídea, K=3.



(d) Modelo PAM con distancia euclídea, K=2.

Figura 5.3: Comparativa de modelos de clustering tras la eliminación de valores atípicos. Se observa cómo varía la partición del conjunto según el algoritmo utilizado.

En esta sección, el primer paso es mostrar la distribución de los estudiantes en los clústeres de cada técnica, así como su proyección en las dos primeras componentes. A simple vista, se puede observar que cada método ha agrupado de manera diferente a los estudiantes.

Por la parte de los métodos jerárquicos (subfiguras 5.3a y 5.3b), se observa de manera bastante marcada como dejan apartado el outlier de la zona superior. Ambos tienen 4 como número de clústeres y, sin embargo, los tratan de manera bastante diferente. Donde en el método de Ward se crean 3 clústeres en la parte inferior bastante separados y compactos, el método de la media crea un solo clúster no tan compacto a la izquierda, dejando aisladas a las tres observaciones de la izquierda, con las que crea clústeres de 2 y de 1 observación.

El método de Ward y el de la media () son los únicos que proponen una partición en cuatro clústeres, reflejando una mayor granularidad en la segmentación. Esto puede deberse a que ambos son métodos jerárquicos, más sensibles a pequeñas variaciones en las distancias.

Por otra parte, tanto K-means como PAM (subfiguras 5.3c y 5.3d) han tendido a crear menos grupos (2 y 3, respectivamente). El algoritmo de K-means crea dos clústeres bien separados, pero el primero no es nada compacto debido a la lejanía de muchos datos. En cuanto al algoritmo de PAM, tienen el mismo problema con el primer clúster, además de crear dos clústeres solapados donde, en K-means, solamente hay un clúster.

Tras todos estos análisis, parece que los adecuado es elegir el método de Ward. Los algoritmos aglomerativos no generan clústeres compactos, teniendo individuos muy alejados y diferentes, con grupos solapados en el caso de PAM. Por otra parte, tanto Ward como la media dejan aislado al individuo de la parte superior, que parece tener características más especiales. Sin embargo, el método de la media separa las 3 observaciones que se encuentran en la parte inferior derecha, siendo, aparentemente, bastante cercanas. En particular, Ward consigue clústeres bien separados y visualmente compactos, justificando su uso pese a la leve caída del coeficiente de Silhouette.

5.4 Análisis del modelo seleccionado y caracterización de perfiles

Una vez seleccionado el método a utilizar y el número de clusters, el último paso es analizar los 4 diferentes perfiles plasmados en los cuatro clústeres. Estos pueden verse nuevamente en la figura 5.4. Para poder comprender los perfiles que caen en cada clúster, es necesario observar los centros de estos, los cuáles son vectores de medias de las variables de los individuos dentro del grupo. Conociendo los valores presentes en la tabla **??**

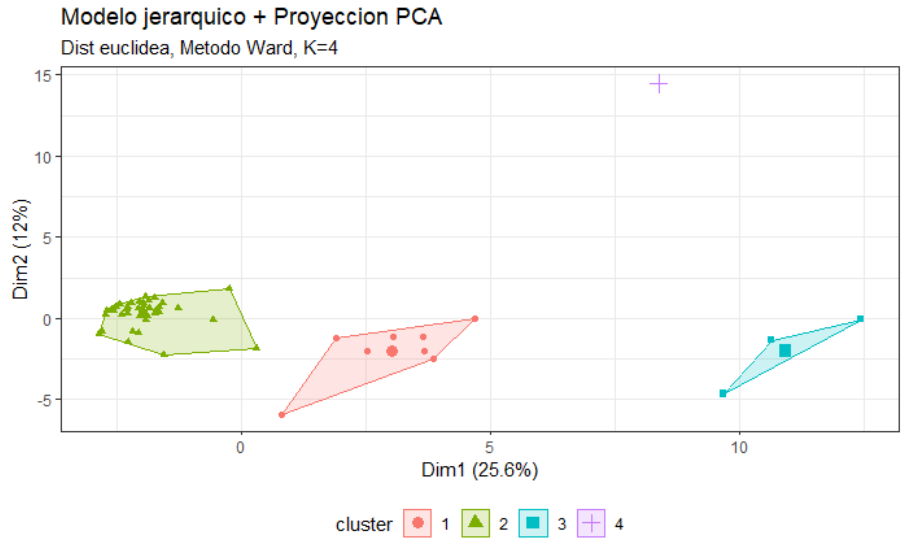


Figura 5.4

- **Clúster 1: Perfil equilibrado y participativo**: En este grupo se representa a los estudiantes con un perfil caracterizado por estudiantes que tienen un volumen y créditos superados dentro de los esperado. Los valores del entorno digital se sitúan por encima de la media, dando a entender que están involucrados en la carrera en un principio. Además, los alumnos de este grupo tienen altos niveles en preferencia. Finalmente, el abandono parece producirse en julio. Este clúster parece representar un perfil académico estable, que decide abandonar tras el año de carrera.

- **Clúster 2: Perfil de baja carga y escasa implicación**: En este grupo se encuentran estudiantes con baja carga lectiva y rendimientos dispares, de haberlos (se comprobó en el anterior capítulo que no había demasiados valores), con bajos niveles de preferencias. Las participación digital indica un bajo nivel de actividad, y los abandonos parecen producirse en septiembre. Este parece el caso de estudiantes que

entraron al grado al no poder acceder a la carrera de preferencia principal, que luego deciden ese mismo mes abandonar la carrera para dejar de estudiar o cambiarse a otra donde los hayan aceptado.

- **Clúster 3: Perfil de abandono temprano y desorientación**: Este grupo es algo particular, pues solo tiene tres estudiantes. A la vista de sus valores, se puede decir que es un grupo con bajo rendimiento generalizado (de haber valores), tanto en créditos como en las variables digitales. Las fechas de abandono se concentran en los primeros meses del curso, pero especialmente en octubre. También se destaca que se desconoce la preferencia de selección, lo que podría indicar indicar una falta de orientación por parte de los estudiantes, que puede desembocar en el abandono de la carrera. Son estudiantes caracterizados por abandonar de manera temprana, poco satisfechos con el grado de informática.
- **Clúster 4: Perfil atípico**: Este es el clúster más especial, ya que representa a un único individuo. Hay una alta carga de créditos superados en asignaturas, pero contrasta con otras variables académicas. El comportamiento digital, aunque alto en la primera componente, refleja una baja en la segunda, y un valor muy negativo en la tercera, indicando un comportamiento algo errático, pero en principio bueno. Presenta extremos en algunas variables, como la preferencia desconocida.

CAPÍTULO 6
Anticipación del abandono mediante modelos predictivos

En este capítulo, el estudio se centrará en la predicción del abandono escolar. En primer lugar, se tratará de paliar el evidente desbalanceo de datos de estudiantes que han abandonado, pues solo un 3 % de los estudiantes han dejado lo estudios frente al 97 %. Para ello, se utilizarán diferentes métodos que permitan generar datos sintéticos para utilizarlos posteriormente como *input* de diferentes modelos de predicción. De entre los modelos utilizados, se seleccionará el que mejor pueda anticipar el abandono escolar, teniendo especial cuidado en el uso de métricas sensibles al desbalanceo de clase, como *F1-score* y el *recall* de la clase minoritaria.

6.1 Tratamiento del desbalanceo de clases

http://eio.usc.es/pub/mte/descargas/proyectosfinmaster/proyecto_1469.pdf

El desbalanceo de clases es una situación en la que el número de muestras de las clases de una base de datos no se encuentran balanceadas. En el caso de muchos estudios, existe un desbalanceo que puede desembocar en la pérdida de eficiencia de modelos de clasificación, pues estos podrían predecir sistemáticamente la clase mayoritaria. Este comportamiento por parte de los algoritmos supone un gran coste en la vida real al ser las clases minoritarias el objetivo de muchos estudios.

En este estudio se ha podido comprobar en varias ocasiones que existe un alto desbalanceo entre estudiantes, disponiendo de solo 52 estudiantes que han abandonado la carrera frente a 1752 que no lo han hecho. Como se ha comentado antes, es necesario tomar medidas, pues un clasificador que acierte más del 95 % de las ocasiones (prediciendo siempre que el estudiante continua) no resulta de utilidad para anticipar que un universitario deje la carrera.

Para tratar dicho problema, se van a utilizar dos técnicas de "resampling", la cual consiste en crear nuevas muestras a través de ya existentes con el objetivo de paliar el desbalanceo, mejorando los modelos futuros que se puedan realizar. Las dos técnicas utilizadas son:

- **Método SMOTE**: Esta técnica genera nuevas instancias de la clase minoritaria mediante una interpolación entre observaciones existentes y sus vecinos más cercanos. Esto evita el sobreajuste asociado a la duplicación directa de ejemplos y proporciona una representación más robusta del espacio de características [1].

- **Método ROSE**: ROSE genera observaciones sintéticas a partir de una estimación de la densidad condicional, combinando sobremuestreo y submuestreo de manera aleatoria. Esto permite crear un conjunto de entrenamiento balanceado que mantiene la variabilidad de los datos originales [2].

  De esta manera,

6.2 Justificación de los modelos utilizados

6.3 Evaluación y comparación de resultados

6.4 Modelo elegido y análisis

CAPÍTULO 7
Creación del conjunto de datos para docencia

El objetivo de este estudio es tratar de mejorar la educación universitaria a través de la identificación de perfiles de estudiantes que abandonen la carrera. Además de ello, en este capítulo también se presentará una base de datos derivada dataset original. Este conjunto de datos tiene como meta su uso en la asignatura *Análisis de Datos en Educación*, ofertado en el Grado de Ciencia de Datos de la Universitat Politècnica de València. El dataset sirve como material de apoyo para futuros curso de la asinatura, donde los alumnos podrán trabajar con él para hacer diferentes descubrimientos y estudios.

A diferencia del dataset tratado a lo largo de todo el estudio, este tiene contiene datos de todos los estudiantes de la UPV.

7.1 Objetivos del conjunto de datos

Preguntar qué se quiere añadir exactamente. También qué créditos sería mejor añadir de entre todos.

7.2 Descripción de la base de datos

La base de datos, como se ha comentado anteriormente, se ha creado utilizando datos de todos los estudiantes de la UPV y todos sus diferentes grados y másteres. La descripción de los campos de la base de datos es la misma a la mostrada anteriormente. Se puede ver la descripción de la base de datos en la figura A.12.

La base de datos se ha diseñado siguiendo una estructura relacional y con el objetivo de ofrecer claridad y simplicidad, que se puede ver a continuación en la figura 7.1. Contiene la siguientes tablas:

Tengo que cambiar los nombres de digital por unos más profesionales

- **Estudiante**: contiene los datos de los estudiantes. Estos datos incluyen la totalidad del dataset "sociodemografia", al contener todos sus datos sociodemográficos, pero no contiene datos académicos.
- **Título**: esta tabla contiene simplemente el código del título y su nombre, al igual que en el archivo *.csv* presentado en el tercer capítulo.

- **Cursa**: esta es una tabla asociación entre el título y el estudiante, que indica que un universitario está estudiando una carrera en un curso académico, con un campus, centro, su rendimiento de ese año y cursos más altos y bajos dados, así como la variable más importante: si ha abandonado el título que cursaba.
- **Asignatura**: esta taba contiene los datos de una asignatura en un título universitario y su código en la plataforma *PoliformaT*.
- **Matrícula**: la tabla *Matrícula* es una clase asociación entre cursa y Asignatura, en la cual un alumno que cursa un título en un año académico se matrícula de una asignatura y todos los datos asociados a esa matrícula, diferentes al ámbito académico y a los datos.
- **Tablas de créditos**: los créditos del alumno se reparten en tres tablas diferentes: *Créditos varios*, que contiene todos los campos de créditos relacionados con convalidaciones, prácticas o actividades; *Créditos superados*, que contiene todas las variables de créditos que el alumno ya ha superado de diferentes tipos y cursos; y *créditos matriculados*, que contiene la información de todos los créditos en los que el alumno se ha matriculado.
- **Comportamiento digital por asignaturas**: esta tabla contiene todas las variables relacionadas con las interacciones del alumno de *PoliformaT* para cada asignatura, además de las variables que son su suma para cada asignatura.
- **Comportamiento digital por estudiante**: esta tabla tiene los campos relcionados con el uso de *PoliformaT* acabados en "_est", que determinan la media de uso por cada estudiante, así como el uso de la red WiFi de la universidad.

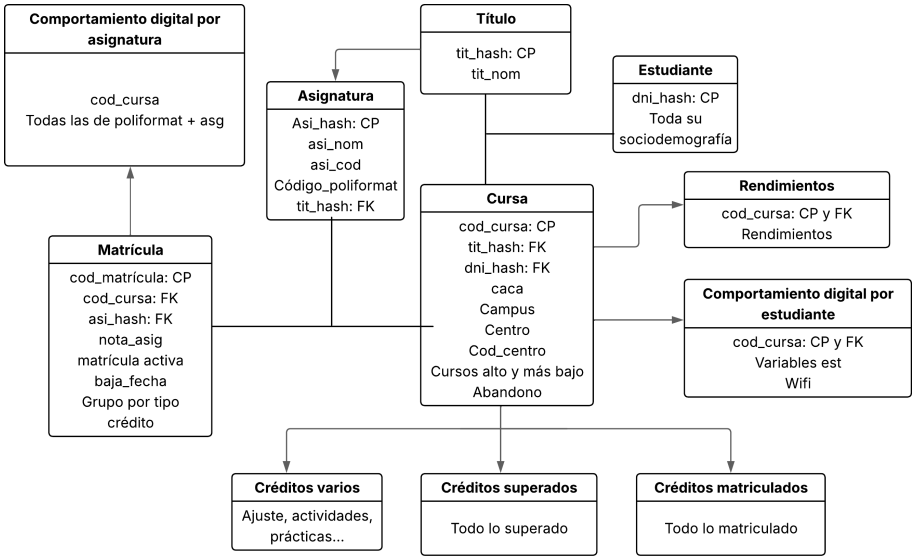


Figura 7.1: UML de la base de datos

El diseño de las tablas responde a varias cuestiones. En primer lugar, las tablas *Estudiante, Título y Asignatura* resultan de las tablas naturales con las que inició el estudio (con la excepción de *Estudiantes*, que ha recibido más información).

Cursa se plantea como eje del esquema, pues que un estudiante cursa un título en un año dado es la base sobre la que se fundamenta toda la estructura de la base de datos. Sin embargo, se han eliminado todos los rendimientos diferentes del total de ese curso.

Esta decisión tiene como fundamento que no es necesario almacenar el rendimiento de cursos anteriores en una misma fila, pues una fila que represente al mismo estudiante en otro año académico (y mismo título) ya lo reflejará. Además, se han eliminado también la diferenciación por cuatrimestres, pues no existe tal diferenciación explícita por asignaturas, por lo que no hay motivos para mantenerlos. Posteriormente, *Matricula* contiene los datos de las asignaturas en las que el estudiante se ha matriculado.

Las tablas restantes se han separado de las tablas originales de las que proceden con el objetivo de ofrecer claridad y sencillez a la hora de trabajar con la base de datos. Las tablas relacionadas con los créditos resulta de la división en tres ámbitos: los créditos superados, matriculados y de otro tipo. Su diferenciación en diferentes clases permita un enfoque más claro y una selección más directa.

Finalmente, nos encontramos con las tablas de comportamiento digital. Por un lado, a tabla *Comportamiento digital por estudiante* tiene como objetivo almacenar el uso de PoliformaT de manera total, correspondiendo solamente a un curso académico, así como las variables relacionadas con el WiFi. Por otro lado, la tabla *Comportamiento digital por asignatura* guarda el comportamiento en los diferentes ámbitos de la plataforma para cada asignatura, así como el uso total para esa asignatura concreta.

Esta base de datos se proporciona en (inserta enlace del github luego). Cada archivo *.csv* representa una tabla diferente de las presentadas previamente, por lo que resulta necesario cargarlas todas para poder conformar la base de datos al completo y trabajar con ella de manera normal.

7.3 Sugerencias de uso

Escenarios académicos

Igualdad entre estudiantes (discapacidad, beca, desplazamiento, etc.).

Patrones de regularidad académica.

Análisis de comportamiento digital entre titulaciones.

Comparación entre asignaturas de un mismo grado que pudieran relacionarse con el abandono.

Aplicación en proyectos de fin de grado o máster por otros estudiantes.

Estudios de más años en caso de disponer de más datos.

Comparación de entre titulaciones de una misma escuela (o no).

CAPÍTULO 8

Conclusiones y recomendaciones

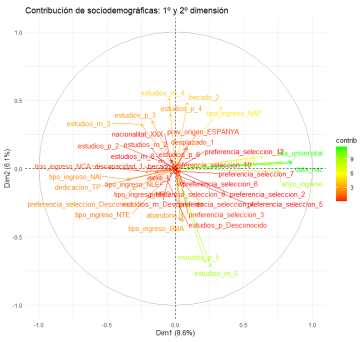
????? ????????????? ????????????? ????????????? ????????????? ?????????????

Bibliografía

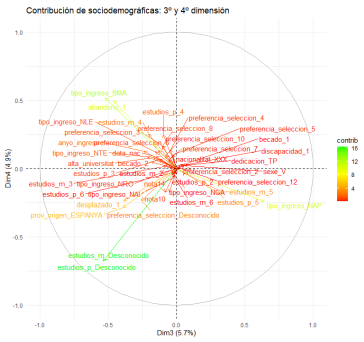
[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321–357.
Disponible en: https://www.jair.org/index.php/jair/article/view/10302

[2] Lunardon, N., Menardi, G., & Torelli, N. (2014). *ROSE: A Package for Binary Imbalanced Learning*. The R Journal, 6(1), 82–92.
Disponible en: https://journal.r-project.org/archive/2014/RJ-2014-008/index.html

APÉNDICE A
Anexos

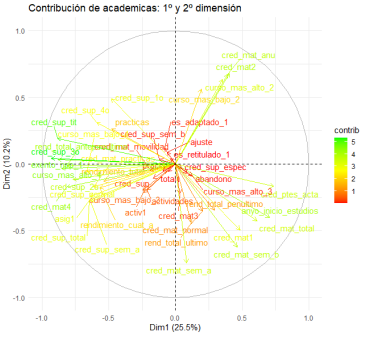


(a) "Sociodemografía" - Componentes 1 y 2

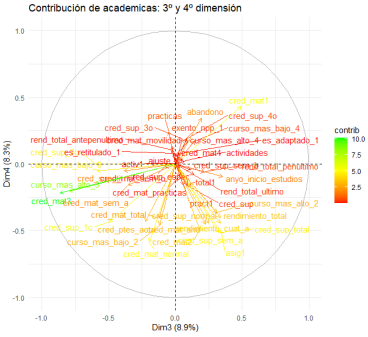


(b) "Sociodemografía" - Componentes 3 y 4

Figura A.1: Contribuciones de las variables sociodemográficas a las cuatro primeras componentes principales

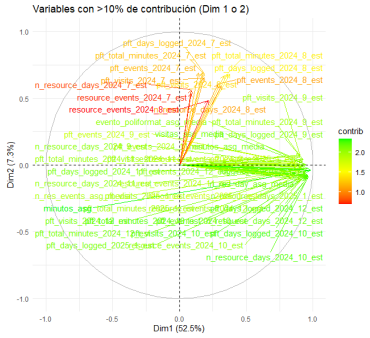


(a) "Académicas" - Componentes 1 y 2

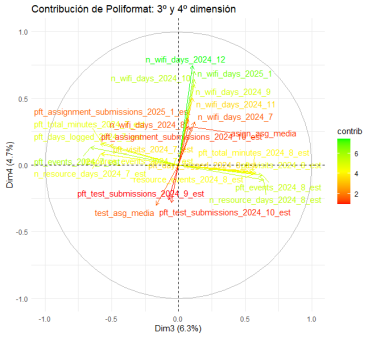


(b) "Académicas" - Componentes 3 y 4

Figura A.2: Contribuciones de las variables académicas a las cuatro primeras componentes principales

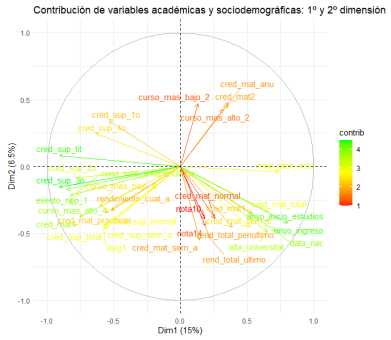


(a) "Poliformat" - Componentes 1 y 2

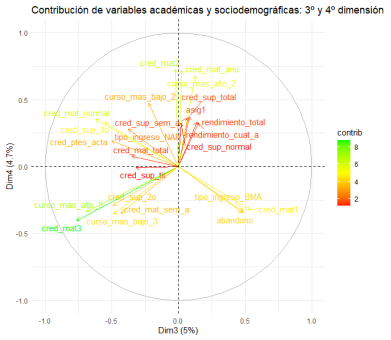


(b) "Poliformat" - Componentes 3 y 4

Figura A.3: Contribuciones de las variables de actividad en Poliformat a las cuatro primeras componentes principales

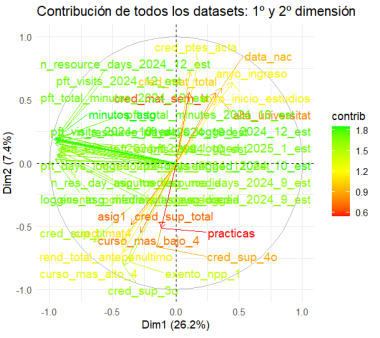


(a) Sociodemográficas y académicas - Componentes 1 y 2

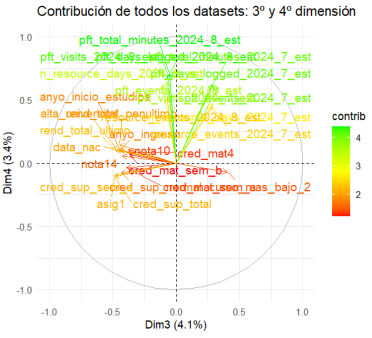


(b) Sociodemográficas y académicas - Componentes 3 y 4

Figura A.4: Contribuciones de variables académicas y sociodemográficas a las cuatro primeras componentes principales



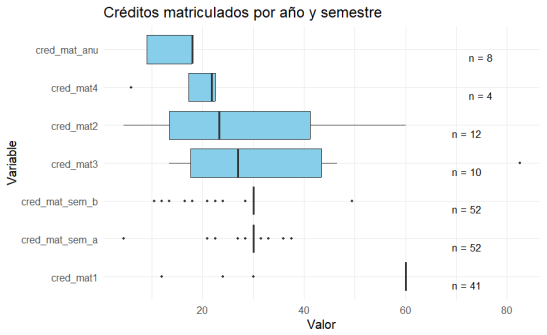
(a) Todas las variables - Componentes 1 y 2



(b) Todas las variables - Componentes 3 y 4

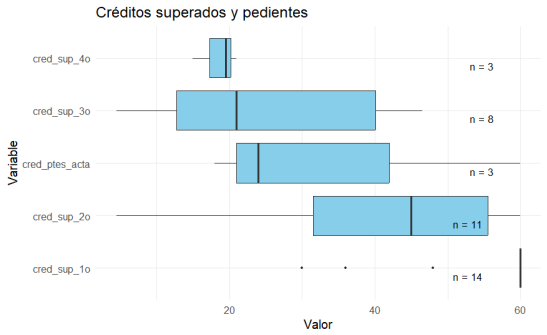
Figura A.5: Contribuciones de todas las variables a las cuatro primeras componentes principales

(a) Distribución de créditos totales de las personas que han abandonado.

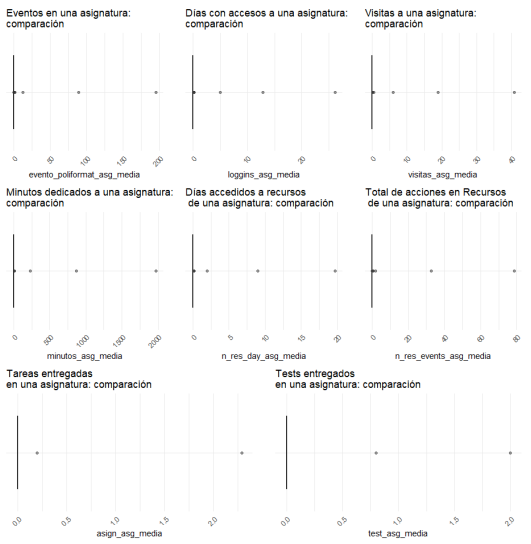


(b) Créditos matriculados de los estudiantes que han abandonado.

Figura A.6: Distribuciones de créditos totales y matriculados en estudiantes que han abandonado.



(a) Distribución de créditos superados de los estudiantes que han dejado la carrera.



(b) Distribución de medias de diferentes de alumnos que han abandonado los estudios.

Figura A.7: Distribuciones de créditos superados y comportamiento digital de estudiantes que han abandonado.

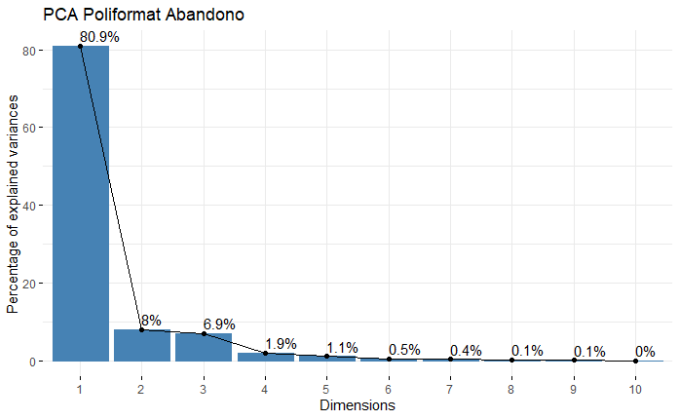


Figura A.8: Ánalisis de Componentes Principales del los datos digitales de los estudiantes que han dejado los estudios de informática

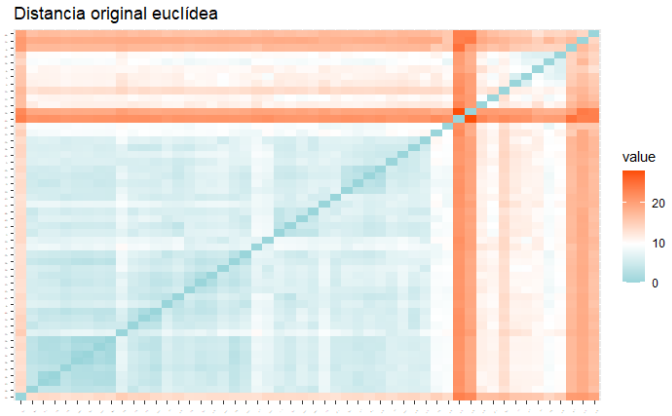
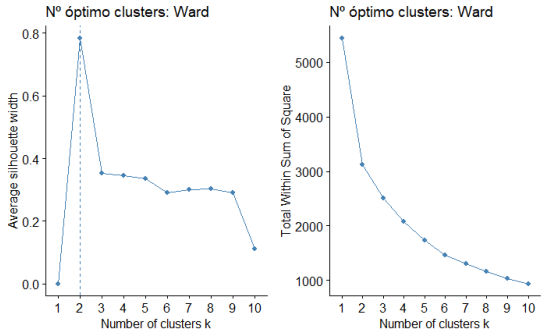
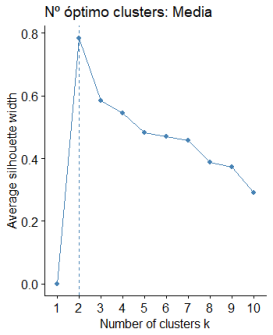


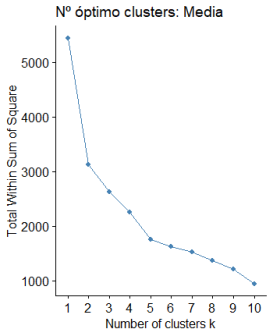
Figura A.9: Ánalisis de Componentes Principales del los datos digitales de los estudiantes que han dejado los estudios de informática

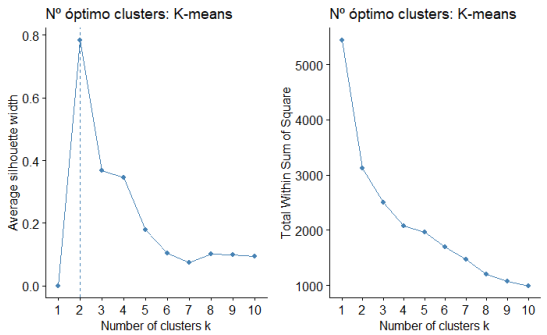


(a) Número óptimo de clústeres con método Ward.

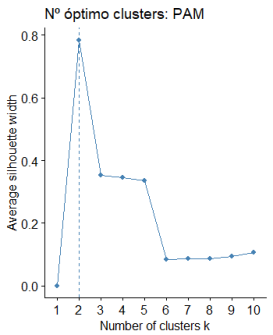


(b) Número óptimo de clústeres con método de la media.





(c) Número óptimo de clústeres con K-means.



(d) Número óptimo de clústeres con método PAM.

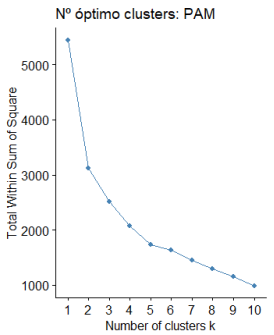
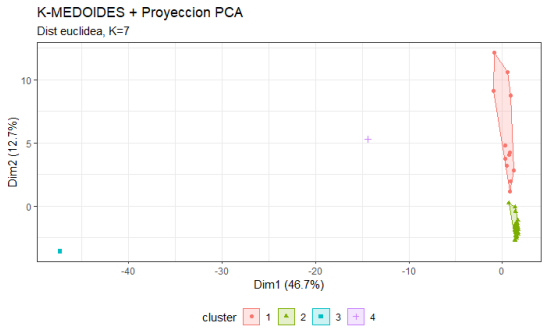
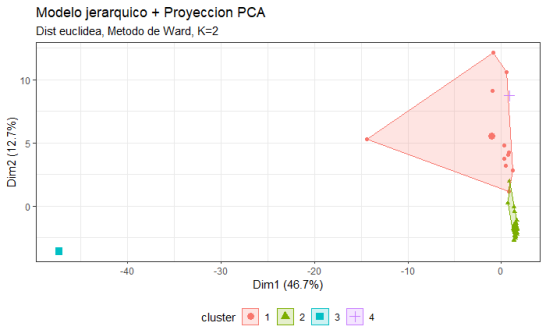


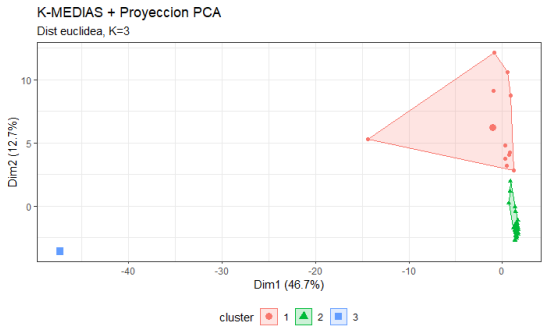
Figura A.10: Evaluación del número óptimo de clústeres mediante el coeficiente de silhouette para diferentes algoritmos de agrupamiento, antes de eliminar valores atípicos.



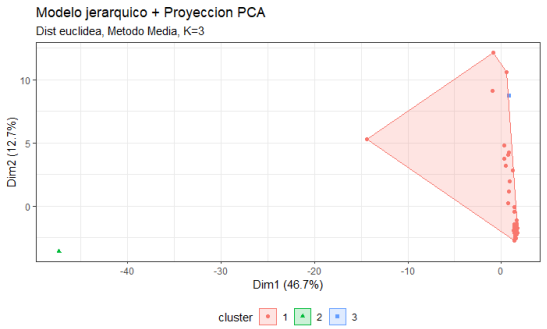
(a) Modelo PAM con distancia euclídea, K=7.



(b) Modelo jerárquico Ward, K=2.



(c) K-means con distancia euclídea, K=3.



(d) Modelo jerárquico método de la media, K=3.

Figura A.11: Comparativa de métodos de agrupamiento proyectados en el plano principal del PCA antes de eliminar los dos valores atípicos.

Variable	NAs	Compl.	Media	SD	Mín	P25	P50	P75	Máx
cred_mat1	0	1.00	16.09	25.24	0.00	0.00	0.00	36.00	60.00
cred_mat2	0	1.00	15.37	23.05	0.00	0.00	0.00	30.00	64.50
cred_mat3	0	1.00	13.28	21.93	0.00	0.00	0.00	19.50	82.50
cred_mat4	0	1.00	10.69	18.92	0.00	0.00	0.00	13.13	73.50
cred_sup_normal	0	1.00	1.31	3.37	0.00	0.00	0.00	0.00	30.00
cred_sup_espec	0	1.00	1.33	9.86	0.00	0.00	0.00	0.00	126.00
cred_sup	0	1.00	2.64	10.27	0.00	0.00	0.00	0.00	126.00
cred_mat_normal	0	1.00	55.84	14.61	0.00	57.00	60.00	60.50	114.00
cred_mat_movilidad	0	1.00	0.86	5.79	0.00	0.00	0.00	0.00	66.00
cred_ptes_acta	0	1.00	49.06	17.64	0.00	39.00	60.00	60.00	78.00
cred_mat_practicas	0	1.00	2.60	6.03	0.00	0.00	0.00	0.00	18.00
cred_mat_sem_a	0	1.00	27.28	7.08	4.50	21.00	30.00	30.00	57.00
cred_mat_sem_b	0	1.00	23.65	7.43	0.00	21.00	25.50	30.00	55.50
cred_mat_anu	0	1.00	4.50	7.47	0.00	0.00	0.00	9.00	18.00
cred_mat_total	0	1.00	55.43	11.38	4.50	52.50	60.00	60.00	96.00
cred_sup_sem_a	0	1.00	0.62	2.07	0.00	0.00	0.00	0.00	30.00
cred_sup_sem_b	0	1.00	0.01	0.41	0.00	0.00	0.00	0.00	12.00
cred_sup_total	0	1.00	0.63	2.11	0.00	0.00	0.00	0.00	30.00
rendimiento_cuat_a	54	0.97	1.98	7.98	0.00	0.00	0.00	0.00	100.00
rendimiento_total	54	0.97	1.31	5.95	0.00	0.00	0.00	0.00	100.00
anyo_inicio_estudios	0	1.00	2022.03	1.95	2010.00	2021.00	2022.00	2024.00	2024.00
cred_sup_1o	0	1.00	43.65	25.41	0.00	18.00	60.00	60.00	60.00
cred_sup_2o	0	1.00	28.16	28.06	0.00	0.00	18.00	60.00	64.50
cred_sup_3o	0	1.00	15.52	24.39	0.00	0.00	0.00	37.50	82.50
cred_sup_4o	0	1.00	2.43	7.01	0.00	0.00	0.00	0.00	60.00
practicas	0	1.00	1.66	5.00	0.00	0.00	0.00	0.00	18.00
actividades	1345	0.23	3.99	2.81	0.20	1.00	4.40	5.77	9.00
ajuste	1702	0.03	9.46	4.30	7.50	7.50	7.50	9.00	30.00
cred_sup_tit	0	1.00	92.79	72.68	0.00	24.00	79.50	166.50	244.50
cred_pend_sup_tit	0	1.00	147.25	72.59	0.00	73.50	160.50	216.00	240.00
asig1	0	1.00	0.63	2.11	0.00	0.00	0.00	0.00	30.00
pract1	1707	0.03	7.31	5.16	0.50	4.00	6.00	9.50	18.00
activ1	1512	0.14	3.49	2.32	0.20	1.00	3.50	5.20	9.00
total1	1733	0.01	12.87	5.82	4.80	9.50	12.91	14.50	27.92
rend_total_ultimo	445	0.75	82.04	27.30	0.00	70.59	100.00	100.00	100.00
rend_total_penultimo	822	0.53	80.83	28.27	0.00	69.13	100.00	100.00	100.00
rend_total_antepenultimo	0	1.00	25.07	40.00	0.00	0.00	0.00	54.54	100.00

Tabla A.1: Resumen de variables numéricas académicas según `skimr`.

Variable	Media	SD	Mín/P25	P50	P75	Máx
pft_events_2024_7_est	0.12	1.22	0.00	0.00	0.00	25.00
pft_events_2024_8_est	1.26	7.51	0.00	0.00	0.00	198.00
pft_events_2024_9_est	105.47	154.13	0.00	0.00	174.00	1178.00
pft_events_2024_10_est	207.40	280.70	0.00	0.00	394.75	2014.00
pft_events_2024_11_est	110.76	176.52	0.00	0.00	183.75	1442.00
pft_events_2024_12_est	245.02	348.44	0.00	0.00	448.75	2782.00
pft_events_2025_1_est	178.56	270.27	0.00	0.00	318.00	2910.00
pft_days_logged_2024_7_est	0.02	0.20	0.00	0.00	0.00	3.00
pft_days_logged_2024_8_est	0.20	0.93	0.00	0.00	0.00	16.00
pft_days_logged_2024_9_est	14.70	20.00	0.00	0.00	27.00	113.00
pft_days_logged_2024_10_est	24.06	30.84	0.00	0.00	48.00	175.00
pft_days_logged_2024_11_est	14.78	21.19	0.00	0.00	27.00	172.00
pft_days_logged_2024_12_est	22.90	29.97	0.00	0.00	45.00	145.00
pft_days_logged_2025_1_est	15.74	20.68	0.00	0.00	31.00	103.00
pft_visits_2024_7_est	0.03	0.34	0.00	0.00	0.00	9.00
pft_visits_2024_8_est	0.28	1.67	0.00	0.00	0.00	41.00
pft_visits_2024_9_est	24.20	35.15	0.00	0.00	42.00	264.00
pft_visits_2024_10_est	45.55	62.08	0.00	0.00	88.00	559.00
pft_visits_2024_11_est	26.06	41.55	0.00	0.00	44.00	487.00
pft_visits_2024_12_est	46.56	66.04	0.00	0.00	86.75	461.00
pft_visits_2025_1_est	33.34	48.42	0.00	0.00	61.00	431.00
pft_total_minutes_2024_7_est	1.21	11.43	0.00	0.00	0.00	197.64

Variable	Media	SD	Mín/P25	P50	P75	Máx
pft_total_minutes_2024_8_est	11.10	57.85	0.00	0.00	0.00	1115.49
pft_total_minutes_2024_9_est	1047.34	1526.21	0.00	0.00	1779.89	11284.18
pft_total_minutes_2024_10_est	1976.43	2712.82	0.00	0.00	3657.21	23054.90
pft_total_minutes_2024_11_est	1114.36	1747.07	0.00	0.00	1891.35	18176.08
pft_total_minutes_2024_12_est	2076.22	2941.07	0.00	0.00	3805.51	21339.98
pft_total_minutes_2025_1_est	1543.17	2231.06	0.00	0.00	2827.20	18270.85
n_wifi_days_2024_7	0.57	2.08	0.00	0.00	0.00	23.00
n_wifi_days_2024_8	0.05	0.47	0.00	0.00	0.00	14.00
n_wifi_days_2024_9	7.18	5.45	0.00	9.00	12.00	29.00
n_wifi_days_2024_10	10.16	7.38	0.00	13.00	16.00	28.00
n_wifi_days_2024_11	0.66	2.08	0.00	0.00	0.00	22.00
n_wifi_days_2024_12	4.96	4.45	0.00	5.00	8.00	29.00
n_wifi_days_2025_1	4.36	4.83	0.00	3.00	7.00	27.00
n_resource_days_2024_7_est	0.01	0.11	0.00	0.00	0.00	2.00
n_resource_days_2024_8_est	0.10	0.60	0.00	0.00	0.00	15.00
n_resource_days_2024_9_est	11.04	15.00	0.00	0.00	21.00	80.00
n_resource_days_2024_10_est	17.79	22.97	0.00	0.00	35.00	113.00
n_resource_days_2024_11_est	9.50	14.00	0.00	0.00	17.00	101.00
n_resource_days_2024_12_est	15.34	20.20	0.00	0.00	30.00	103.00
n_resource_days_2025_1_est	10.65	13.97	0.00	0.00	21.00	71.00
resource_events_2024_7_est	0.03	0.51	0.00	0.00	0.00	12.00
resource_events_2024_8_est	0.31	2.88	0.00	0.00	0.00	78.00
resource_events_2024_9_est	31.97	47.49	0.00	0.00	54.00	302.00
resource_events_2024_10_est	68.47	96.02	0.00	0.00	126.00	617.00
resource_events_2024_11_est	31.38	48.66	0.00	0.00	54.00	370.00
resource_events_2024_12_est	72.10	101.06	0.00	0.00	132.00	676.00
resource_events_2025_1_est	56.46	80.07	0.00	0.00	104.00	570.00
pft_assignment_submissions_2024_9_est	0.18	0.72	0.00	0.00	0.00	10.00
pft_assignment_submissions_2024_10_est	0.91	1.87	0.00	0.00	1.00	17.00
pft_assignment_submissions_2024_11_est	0.67	1.91	0.00	0.00	0.00	17.00
pft_assignment_submissions_2024_12_est	1.44	2.48	0.00	0.00	2.00	17.00
pft_assignment_submissions_2025_1_est	0.45	1.31	0.00	0.00	0.00	12.00
pft_test_submissions_2024_9_est	1.01	2.59	0.00	0.00	0.00	27.00
pft_test_submissions_2024_10_est	1.99	4.19	0.00	0.00	2.00	32.00
pft_test_submissions_2024_11_est	1.16	2.13	0.00	0.00	2.00	32.00
pft_test_submissions_2024_12_est	2.12	4.95	0.00	0.00	2.00	56.00
pft_test_submissions_2025_1_est	1.18	3.79	0.00	0.00	1.00	35.00
abandono	0.03	0.17	0.00	0.00	0.00	1.00
evento_poliformat_asg_media	156.26	205.63	0.00	0.00	291.25	2245.25
loggins_asg_media	16.87	20.53	0.00	0.00	33.19	114.00
visitas_asg_media	32.11	42.04	0.00	0.00	60.00	355.25
minutos_asg_media	1420.53	1875.37	0.00	0.00	2671.52	18398.17
n_res_day_asg_media	11.78	14.37	0.00	0.00	23.66	85.25
n_res_events_asg_media	47.93	63.20	0.00	0.00	92.00	753.00
asign_asg_media	0.66	1.24	0.00	0.00	0.83	9.50
test_asg_media	1.40	2.46	0.00	0.00	1.83	18.00
minutos_asg	7769.18	10562.52	0.00	2.00	14514.39	82185.77

Tabla A.2: Resumen de variables de poliformat según `skimr`, sin ningún dato ausente.

Mes	Media de asignaturas abandonadas	Porcentaje del total
Julio	10.00	44.06 %
Septiembre	10.45	40.04 %
Octubre	9.00	8.62 %
Diciembre	11.00	4.21 %
Enero	8.00	3.07 %

Tabla A.3: Media de las asignaturas abandonadas por mes

Estadístico	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
Hopkins	0.7353	0.7629	0.7902	0.7801	0.7961	0.8128

Tabla A.4: Resumen del coeficiente de silhouette previo a la eliminación de valores atípicos.

Tabla A.5: Resumen de valores medios por clúster utilizando el método de Ward.

Variable	**Clúster 1**	**Clúster 2**	**Clúster 3**	**Clúster 4**
cred_mat1	0.23	-1.79	-1.79	-1.79
cred_mat2	-0.04	1.24	-0.44	0.61
cred_mat3	-0.22	0.92	0.54	1.20
cred_mat4	-0.24	4.04	3.90	-0.27
cred_sup_espec	-0.14	-0.14	-0.14	7.07
cred_sup	-0.14	-0.14	-0.14	7.07
cred_mat_normal	-0.17	2.40	-0.21	-0.21
cred_ptes_acta	-0.17	2.40	-0.21	-0.21
cred_mat_sem_a	0.10	0.55	-3.06	-0.48
cred_mat_sem_b	0.07	-1.61	-1.12	-2.35
cred_mat_anu	-0.05	2.81	-0.40	-0.40
cred_mat_total	0.08	1.21	-3.27	-2.72
anyo_inicio_estudios	0.23	-2.02	-3.51	0.46
cred_sup_1o	-0.23	1.79	1.79	1.79
cred_sup_2o	-0.28	1.45	2.73	1.93
cred_sup_3o	-0.30	3.25	3.80	1.59
cred_sup_4o	-0.24	3.17	4.36	-0.24
practicas	-0.25	4.00	4.00	-0.25
cred_sup_tit	-0.29	2.52	3.06	1.70
data_nac	0.22	-1.68	-2.33	-2.65
alta_universitat	0.25	-1.45	-1.71	-5.90
anyo_ingreso	0.23	-2.02	-3.51	0.46
nota10	0.12	0.67	-0.85	-3.02
nota14	0.09	-0.06	-1.32	0.02
curso_mas_bajo_2	-0.20	4.95	-0.20	-0.20
curso_mas_alto_2	-0.14	7.07	-0.14	-0.14
curso_mas_alto_3	-0.15	-0.25	1.88	-0.25
es_adaptado_1	-0.14	-0.14	-0.14	7.07
nacionalitat_XXX	-0.14	-0.14	3.47	-0.14
sexe_V	0.00	0.42	-0.95	0.42
prov_origen_ESPANYA	0.01	-0.48	-0.48	2.03
tipo_ingreso_BMA	0.16	-1.63	-1.63	0.60
tipo_ingreso_NAP	-0.16	1.63	1.63	-0.60
estudios_p_3	-0.04	-0.51	0.70	-0.51
estudios_p_4	0.01	1.31	0.28	-0.75
estudios_p_5	0.06	-0.82	-0.82	-0.82
estudios_p_6	-0.14	-0.14	-0.14	7.07
estudios_m_2	0.02	-0.14	-0.14	-0.14
estudios_m_3	-0.09	-0.36	1.19	2.74

Continúa en la siguiente página

Tabla A.5 – *Continuación de la página anterior*

Variable	Clúster 1	Clúster 2	Clúster 3	Clúster 4
estudios_m_4	0.00	1.36	0.32	-0.72
estudios_m_5	0.05	-1.03	-1.03	-1.03
dedicacion_TP	-0.06	4.00	-0.25	-0.25
desplazado_1	0.05	-0.72	-0.72	1.36
discapacidad_1	0.02	-0.14	-0.14	-0.14
becado_2	-0.14	-0.14	-0.14	-0.14
preferencia_seleccion_2	0.04	-0.32	-0.32	-0.32
preferencia_seleccion_3	0.02	-0.54	0.63	-0.54
preferencia_seleccion_Baja	0.09	-0.66	-0.66	-0.66
preferencia_seleccion_Desconocido	-0.20	-0.20	2.38	4.95
mes_julio	0.12	-0.88	-0.88	-0.88
mes_septiembre	-0.03	1.25	0.23	1.25
mes_octubre	-0.03	-0.32	1.36	-0.32
mes_diciembre	-0.09	-0.20	-0.20	-0.20
PC1	1.21	1.05	0.66	0.81
PC2	-0.20	0.22	0.47	0.40
PC3	-0.04	-0.60	-0.73	-0.78

Figura A.12: Descripción de la base de datos creada para docencia.

- *Estudiante:*
  - **dni_hash**: Texto, Clave primaria
  - **nacionalitat**: Categórica
  - **data_nac**: Fecha
  - **sexe**: Texto
  - **alta_universitat**: Fecha
  - **prov_origen**: Texto
  - **estudios_p**: Categórica
  - **estudios_m**: Categórica
  - **tipo_ingreso**: Categórica
  - **anyo_ingreso**: Entero
  - **anyo_inicio_estudios**: Entero
  - **discapacidad**: Booleano
  - **desplazado**: Booleano
  - **exento_npp**: Booleano
  - **es_retitulado**: Booleano
  - **es_adaptado**: Booleano
  - **preferencia_seleccion**: Categórica
  - **becado**: Booleano
  - **abandono**: Booleano
  - **nota10**: Numérico (escala 0–10)

- **nota14**: Numérico (escala 0–14)

- *Título:*
  - **tit_hash**: Texto, Clave primaria
  - **titnom**: Texto
- *Asignatura:*
  - **asi_hash**: Texto, Clave primaria
  - **asinom**: Texto
- *Cursa:*
  - **cod_cursa**: Texto, Clave primaria
  - **dni_hash**: Clave foránea (Estudiante)
  - **tit_hash**: Clave foránea (Título)
  - **caca**: Entero
  - **curso_mas_bajo**: Entero
  - **curso_mas_alto**: Entero
  - **campus**: Texto
  - **cod_centro**: Texto
  - **Abandono**: Booleano
- *Matrícula:*
  - **cod_matricula**: Texto, Clave primaria
  - **cod_cursa**: Clave foránea (Cursa)
  - **asi_hash**: Clave foránea (Asignatura)
  - **nota_asig**: Numérico
  - **matricula_activa**: Booleano
  - **impagado_curso_mat**: Booleano
  - **baja_fecha**: Fecha
  - **grupos_por_tipocredito**: Texto
- *Rendimientos:*
  - **cod_cursa**: Texto, Clave primaria y foránea
  - **rendimiento_cuat_a**: Numérico
  - **rendimiento_total**: Numérico
  - **rend_total_ultimo**: Numérico
  - **rend_total_penultimo**: Numérico
  - **rend_total_antepenultimo**: Numérico
- *Créditos matriculados:*
  - **cod_cursa**: Texto, Clave primaria y foránea
  - **cred_mat1**–**cred_mat6**: Numéricos
  - **cred_mat_sem_a**: Numérico
  - **cred_mat_sem_b**: Numérico

- **cred_mat_anu**: Numérico
- **cred_mat_normal**: Numérico
- **cred_mat_movilidad**: Numérico
- **cred_mat_practicas**: Numérico
- **cred_mat_total**: Numérico

- *Créditos superados:*
  - **cod_cursa**: Texto, Clave primaria y foránea
  - **cred_sup_1o–cred_sup_6o**: Numéricos
  - **cred_sup_sem_a**: Numérico
  - **cred_sup_sem_b**: Numérico
  - **cred_sup_anu**: Numérico
  - **cred_sup_normal**: Numérico
  - **cred_sup_espec**: Numérico
  - **cred_sup**: Numérico
  - **cred_sup_tit**: Numérico
  - **cred_pend_sup_tit**: Numérico
  - **cred_sup_total**: Numérico
- *Créditos resumen:*
  - **cod_cursa**: Texto, Clave primaria y foránea
  - **total1**: Numérico
  - **asig1**: Numérico
  - **pract1**: Numérico
  - **activ1**: Numérico
  - **ajuste1**: Categórica
- *Actividad digital por estudiante:*
  - **cod_cursa**: Texto, Clave primaria y foránea
  - **pft_events_yyyy_mm_est**: Numéricos
  - **pft_visits_yyyy_mm_est**: Numéricos
  - **pft_days_logged_yyyy_mm_est**: Numéricos
  - **pft_total_minutes_yyyy_mm_est**: Numéricos
  - **n_resource_days_yyyy_mm_est**: Numéricos
  - **resource_events_yyyy_mm_est**: Numéricos
  - **pft_assignment_submissions_yyyy_mm_est**: Numéricos
  - **pft_test_submissions_yyyy_mm_est**: Numéricos
- *Actividad digital por asignatura:*
  - **cod_matricula**: Texto, Clave foránea
  - **evento_poliformat_asg**: Numérico
  - **loggins_asg**: Numérico
  - **visitas_asg**: Numérico

- **minutos_asg**: Numérico
- **n_res_day_asg**: Numérico
- **n_res_events_asg**: Numérico
- **asign_asg**: Numérico
- **test_asg**: Numérico