Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: 3

Segmentación del estudiantado universitario: el caso de la Udima

---

Autora: Alejandra Bonilla Garzón

Tutora: Laia Subirats Maté

Profesor: Ferran Prados Carrasco

Madrid, 13 de enero de 2023

Créditos/Copyright

Esta obra está sujeta a una licencia de Reconocimiento - NoComercial - SinObraDerivada 3.0 España de CreativeCommons.

FICHA DEL TRABAJO FINAL

Título del trabajo:	Segmentación del estudiantado universitario: el caso de la Udima
Nombre de la autora:	Alejandra Bonilla Garzón
Nombre de la colaboradora docente:	Laia Subirats Maté
Nombre del PRA:	Ferran Prados Carrasco
Fecha de entrega (mm/aaaa):	01/2023
Titulación o programa:	Máster Universitario en Ciencia de Datos (*Data Science*)
Área del Trabajo Final:	TFM - Área 3
Idioma del trabajo:	Español
Palabras clave	segmentación, *clustering*, educación superior, estudiante, características, clasificación, predicción

Dedicatoria

A ti, mamá, que no has podido acompañarme en esta aventura pero sé que estarías muy orgullosa de ella.

Agradecimientos

Mis más sinceros agradecimientos a:

Mi familia, por dedicarme todos los esfuerzos, recursos y tiempo para educarme y formarme a lo largo de la vida. Gracias por vuestra paciencia, apoyo incondicional, amor, comprensión y por confiar en mi más que yo misma. Sin vosotros, nada de esto hubiera sido posible.

A mis amigas, por comprender mis ausencias y aún así apoyarme en cualquiera de mis decisiones, ¡sois fundamentales!

A Pollo, por tu compañía infinita, lealtad, cariño y apoyo incondicional.

A Alberto, por tu amor, compromiso y paciencia. Siento que no siempre haya sido fácil.

A todos los profesores que he tenido, por formarme y motivarme, en especial, a todos aquellos que habéis despertado en mi la curiosidad por profundizar y continuar en el proceso infinito del aprendizaje. A Laia, mi tutora de TFM, por su disponibilidad y cercanía. Gracias por acompañarme y guiarme en esta aventura. Espero que esta colaboración solo sea el principio.

Gracias a los compañeros del máster, por el apoyo y la colaboración. Sin vosotros no hubiera sido posible poner el broche final a esta etapa.

A mis compis de la UTC, por su disposición, colaboración y apoyo en el día a día. En especial, a David, por liderar el departamento con confianza, cercanía, disponibilidad, motivación e inspiración.

A David, Isaac, Ana y Juanje por motivarme y ayudarme a elegir un tema interesante para este TFM. Gracias por vuestra colaboración en todas las fases del proyecto.

Por último, a la Udima, y cómo no a David, Isaac, Arancha y Roque por apostar y confiar en mí en este proyecto. Este TFM es solo una pequeña parte de todo lo que está por descubrir.

¡Gracias a todos los que habéis colaborado para que este TFM sea una realidad!

Resumen

Resumen

Este trabajo final determina las distintas tipologías del estudiantado de la Udima en base a su rendimiento, opiniones, características sociodemográficas y de comportamiento en *Moodle* para poder personalizar su seguimiento.

La población investigada esta formada por el estudiantado universitario (Grado/Máster) de la Udima matriculado en algún curso desde su inicio (2008-09) hasta el último curso (2021-22).

Se identifica el perfil de 30.875 estudiantes que finalizan/abandonan sus estudios mediante la técnica óptima de aprendizaje no supervisado: algoritmo de agrupación *k-means*, distancia euclidiana y 4 *clusters* obteniendo: 25.09 % del conjunto como *dropout*; 59.59 % como *common graduate*; 8.76 % como *motivated*: egresado especialista, con sentido de pertenencia a la universidad y alta satisfacción y 6.55 % como *dissatisfied*: egresado con satisfacción inferior a la media.

Esta clasificación sirve para predecir, mediante técnicas de aprendizaje supervisado, la clasificación de 7.989 estudiantes en progreso con un *dataset* reducido (se eliminan variables relativas al egreso/abandono para no condicionar). Mediante *10-Fold-Cross-Validation* (y según máxima *accuracy*) se aplica el algoritmo de clasificación *Random Forest*. Tras entrenar, se obtiene una *accuracy* del 86.45 %. Se aplica el modelo y la distribución es: *dropout*, 59.36 %; *common graduate*, 30.85 %; *motivated*, 1.11 % y *dissatisfied*, 8.69 %.

Evaluados los resultados frente al problema planteado y frente a los resultados de la literatura, se da por válido el modelo.

La diferencia que aporta este modelo frente a los utilizados en Udima es la integración de información procedente de la satisfacción y la inserción laboral. Los resultados obtenidos son específicos para la Udima, pero la metodología empleada puede adaptarse a cualquier institución de educación superior a distancia.

Palabras clave: segmentación, *clustering*, educación superior, estudiante, características, clasificación, predicción.

Abstract

This final project ascertains the different typologies of Udima students based on their performance, opinions, socio-demographic and behavioural characteristics in Moodle in order to customize their follow-up.

The research population is made up of Udima university students (Bachelor's/Master's) enrolled in a course from the beginning (2008-09) to the final year (2021-22).

The profile of 30.875 students who finish/leave their studies is identified using the optimal unsupervised learning technique: k-means group algorithm, Euclidean distance and 4 clusters obtaining: 25.09 % of the group as dropout; 59.59 % as common graduate; 8.76 % as motivated: specialist graduate, with a sense of belonging to the university and high satisfaction and 6.55 % as dissatisfied: graduate with below-average satisfaction.

This classification is used to predict, by means of monitored learning techniques, the classification of 7.989 students in progress with a reduced dataset (variables related to graduation/dropout are eliminated so as not to condition). Using 10-Fold-Cross-Validation (and according to maximum accuracy), the Random Forest classification algorithm is applied. After training, an accuracy of 86.45 % is obtained. The model is applied and the distribution is: dropout, 59.36 %; common graduate, 30.85 %; motivated, 1.11 % and dissatisfied, 8.69 %.

After evaluating the results in relation to the problem raised and the results of the literature, the model is considered valid.

The difference between this model and those used in Udima is the integration of information from satisfaction and job placement. The results obtained are specific to Udima, but the methodology used can be adapted to any advanced e-learning institution.

Keywords: segmentation, clustering, higher education, students, characteristics, classification, prediction.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto, justificación y motivación del trabajo.

Con la declaración de Bolonia (2) se creó el Espacio Europeo de Educación Superior (EEES), sirviendo como referencia para modificar el marco normativo educativo de distintos países con el fin de homogeneizar y equiparar las titulaciones universitarias. Para ello, se plantean seis objetivos fundamentales (2), entre los que se encuentra el 'Establecimiento de un sistema internacional de créditos (ECTS)'. Estos ECTS (*European Credit Transfer and Accumulation System*) regulan el tiempo de dedicación del estudiantado e introducen requisitos de evaluación continua para superar las asignaturas.

Las universidades españolas adaptaron sus planes de estudios a la normativa de Bolonia y definieron la dedicación del estudiantado y los requisitos de superación de las asignaturas. El sistema de ECTS requiere que el estudiante asista a clase y sea evaluado continuamente durante el desarrollo del semestre. El estudiantado con ocupaciones personales y/o laborales presenta dificultades para el seguimiento habitual de un curso universitario en modalidad presencial. Estos motivos hacen proliferar la creación de titulaciones universitarias a distancia (3) y que sirvan como reclamo tanto para nuevo estudiantado como para estudiantado con titulaciones universitarias inacabadas en sus universidades de origen. El uso de las tecnologías de la información y la comunicación (TIC) disponibles en los LMS (sistema de gestión de aprendizaje) han permitido modificar el paradigma del aula de educación superior. Los avances de la tecnología educativa y la necesidad de involucrar al estudiantado en las metas de aprendizaje contribuyen a que los procesos de enseñanza-aprendizaje se agilicen. (Norambuena et al. 2022) (4)

La Universidad a Distancia de Madrid (Udima) (5) aplica una metodología de enseñanza a distancia en la que, en la totalidad de sus titulaciones, la superación de la mayoría de las asignaturas se consigue mediante evaluación continua y examen final presencial, con una proporción de calificación de 40 % y 60 %, respectivamente. Para poder presentarse al examen final

presencial, el estudiante debe superar la evaluación continua.

El desarrollo del trabajo se justifica ante la necesidad de conocer desde la Dirección y el Rectorado de la Udima la proporción y características del estudiantado de titulaciones oficiales de Grado y Máster, con el fin de poder determinar las necesidades de atención y acompañamiento necesarias y así optimizar el egreso y evitar el abandono del estudiantado. Este estudio también permitirá dar a conocer a la sociedad la proporción y características del estudiantado universitario de la Udima y servir como referente para otros estudios similares.

La principal motivación personal para desarrollar este trabajo se centra en el desarrollo y especialización profesional en el ámbito de la ciencia de datos (*data science*). Actualmente me encuentro desarrollando mi actividad laboral en la Unidad Técnica de Calidad (UTC) de la Udima, obteniendo, gestionando y facilitando datos asociados al Sistema Interno de Garantía de Calidad (SIGC) de la Udima con el fin de facilitar la toma de decisiones basadas en evidencias. Este estudio permitirá determinar la toma de decisiones así como la revisión de los procedimientos asociados al SIGC para facilitar la optimización del egreso y evitar el abandono del estudiantado.

Como se indica en uno de los trabajos de la UNESCO (6) "Al mejorar la capacidad de planificar y gestionar la escolarización, destinar a los docentes a las zonas donde más se les necesita, **promover** el uso de material didáctico y **planes de estudio pertinentes y actualizados** y proporcionar pasarelas entre los diversos niveles y contextos educativos **se garantiza que los sistemas educativos podrán responder a las auténticas necesidades de la sociedad**". Por tanto, la motivación personal se complementa con conocer y tratar de mejorar que los planes de estudios ofertados por la Udima cubran las necesidades de la sociedad y su pertinencia y gestión optimicen la finalización de los estudios y reduzcan el abandono de los mismos.

1.2. Objetivos del trabajo.

Los objetivos general y específicos del proyecto se establecen atendiendo a los principios de que sean claros, concisos, medibles, alcanzables, relevantes y planificables.

Objetivo general:

Conocer las características del estudiantado que finaliza u abandona una titulación universitaria en la Udima en base a su rendimiento académico, opiniones manifestadas, características sociodemográficas y de comportamiento en el *LMS* utilizado para el proceso de enseñanza-aprendizaje. De esta forma, también se pretende **identificar qué proporción** del estudiantado egresado o que abandona la universidad **corresponde a cada tipo** para **poder predecir el perfil del estudiantado en progreso**, con especial atención en los que **están en riesgo de abandono**.

Objetivos específicos:

- Identificar los perfiles del estudiantado que finaliza o abandona sus estudios.
- Determinar las características principales de los perfiles del estudiantado.
- Predecir la clasificación del estudiantado que no ha finalizado los estudios con especial atención en el estudiantado en riesgo de abandono.

1.3. Impacto en sostenibilidad, ético-social y de diversidad.

Para garantizar la adquisición de la competencia de compromiso ético y global (CCEG) definida en la asignatura Trabajo Fin de Máster (TF o TFM) del Máster Universitario en Ciencia de Datos (*Data Science*) como *Actuar de manera honesta, ética, sostenible, socialmente responsable y respetuosa con los derechos humanos y la diversidad, tanto en la práctica académica como en la profesional, y diseñar soluciones para mejorar estas prácticas* se abordarán (7) tres grandes dimensiones:

- Sostenibilidad
- Comportamiento ético y responsabilidad social (RS)
- Diversidad (género, entre otros) y derechos humanos

Estas dimensiones están alineadas con los ODS (8), con los que la UOC está públicamente comprometida (9) y el desarrollo del presente trabajo se encuadra en los siguientes ODS:

- **Sostenibilidad**:
  - ODS 12 - *Responsible consumption and production*: El resultado de este trabajo, con la implementación del producto/servicio pretende impactar positivamente en este ODS para mejorar el consumo responsable de los recursos público-privados tanto de la Udima (universidad privada) como de las instituciones públicas y los recursos que facilitan (por ejemplo, la Fundación Madri+d, agencia evaluadora de los procesos asociados al ciclo de vida de los títulos oficiales o del Sistema Integrado de Información Universitario, en adelante SIIU).

- **Comportamiento ético y responsabilidad social (RS)**:
  - ODS 8 - *Decent work and economic growth*: El resultado de este trabajo, con la implementación del producto/servicio pretende impactar positivamente en este ODS para aumentar el crecimiento económico de la Udima en lo relativo a la retención del estudiantado y reducción del abandono.
  - ODS 16 - *Peace, justice and strong institutions*: El resultado de este trabajo, con la implementación del producto/servicio pretende impactar positivamente en este ODS para aumentar la confianza de los grupos de interés de la Udima, en especial, del estudiantado y consolidar su reputación y prestigio en el panorama universitario español.
- **Diversidad (género entre otros) y derechos humanos**:
  - ODS 5 - *Gender equality*: El resultado de este diseño será indicado para todo el estudiantado, con independencia de su género.
  - ODS 10 - *Reduced inequalities*: El resultado de este diseño será indicado para todo el estudiantado con independencia de su raza, etnia, origen, orientación sexual, ideología, religión, diversidad funcional o posición social.

1.4. Enfoque y método seguido.

La metodología a emplear en el trabajo será la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) dado que permite complementar un servicio existente en la Udima que consiste en el seguimiento personalizado del estudiantado y en la prevención del abandono. Esta metodología permitirá crear este nuevo servicio para conseguir los objetivos propuestos.

A continuación se detallan las fases del ciclo de vida del proyecto según la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*):

- **Fase I. *Business Understanding* o Definición de necesidades**: El objetivo de esta fase del proyecto es definir los objetivos generales y específicos.
- **Fase II. *Data Understanding* o Estudio y comprensión de los datos**: El objetivo de esta fase es estudiar los datos disponibles para evitar problemas en fases posteriores. Para ello se debe recopilar, acceder, describir, explorar y verificar la calidad de datos.
- **Fase III. *Data Preparation* o Análisis de los datos y selección de características**: En esta fase se preparan los datos para asegurar la calidad de los mismos según las

conclusiones y/o necesidades detectadas en la fase anterior (Fase II). Es decir, en esta fase, a los datos iniciales (*input*) se aplican algunas operaciones de limpieza, normalización, discretización y reducción de dimensionalidad para obtener unos datos de calidad (*output*) que permitirán la siguiente fase de modelado (Fase IV).

- **Fase IV. *Modeling* o Modelado**: En esta fase, el objetivo es modelar las técnicas que son necesarias para resolver el objetivo principal del proyecto.
- **Fase V. *Evaluation* o Evaluación (obtención de resultados)**: En esta fase se identifican posibles limitaciones del *dataset* seleccionado y se analizan los riesgos para el caso de uso.
- **Fase VI. *Deployment* o Despliegue (puesta en producción)**: El objetivo de dicha fase se corresponde con aplicar el modelo descrito en la Fase IV y evaluado en la Fase V para obtener los beneficios de su aplicación.

1.5. Planificación del trabajo.

A continuación se muestra la planificación del trabajo en un diagrama de Gantt asociando las fechas límites de las PECs y las tareas asociadas a la metodología propuesta para alcanzar las distintas actividades evaluables asociadas a la asignatura:

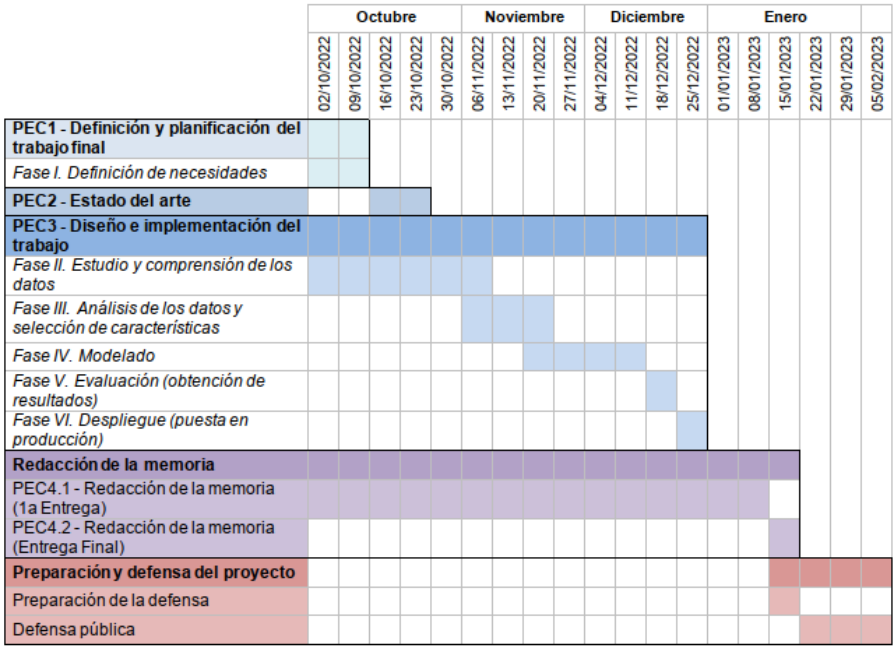

Figura 1.1: Planificación del trabajo.

Capítulo 2

Estado del arte

2.1. Rendimiento académico de la titulación y abandono.

2.1.1. Terminología.

La Subdirección General de Actividad Universitaria Investigadora de la Secretaría General de Universidades dependiente del Ministerio de Universidades establece en el *Capítulo II: Área Académica* del Catálogo oficial de indicadores universitarios del SIIU (10) los indicadores oficiales de rendimiento académico que aplican (11) a todo el Sistema Universitario Español, en adelante SUE.

2.1.1.1. Rendimiento académico de la titulación.

El presente trabajo tiene definido como parte de su objetivo principal el clasificar al estudiantado en función de su comportamiento académico, entre otras variables. El Catálogo oficial de indicadores universitarios del SIIU (10) establece los siguientes indicadores de rendimiento académico de los egresados:

$$Tasa\ de\ graduación = \frac{Egresados\ X + n}{Estudiantes} * 100$$

$$Duración\ media\ de\ los\ estudios = \frac{Cursos}{Egresados\ X}$$

$$Tasa\ de\ eficiencia = \frac{Superados}{Matriculados} * 100$$

$$Tasa\ de\ éxito\ (egresados) = \frac{Superados}{Presentados} * 100$$

$$Tasa\ de\ evaluación\ (egresados) = \frac{Presentados}{Matriculados} * 100$$

Siendo:

- Cursos: Suma de cursos hasta que titulan los egresados en el curso X
- Egresados X: Estudiantes egresados en el curso X
- Egresados X+n: Estudiantes de nuevo ingreso en el curso X egresados en el curso X+n o antes (n: número de cursos completos de la titulación)
- Estudiantes: Estudiantes de nuevo ingreso en el curso X
- Matriculados: ECTS matriculados por los egresados en el curso X
- Presentados: ECTS presentados por los egresados en el curso X
- Superados: ECTS superados por los egresados en el curso X

2.1.1.2. Abandono.

El presente trabajo tiene definido como parte de su objetivo principal predecir el abandono del estudiantado en función de su comportamiento académico, entre otras variables. El Catálogo oficial de indicadores universitarios del SIIU (10) establece los siguientes indicadores al respecto:

$$Tasa\ de\ abandono\ del\ SUE = Tasa\ de\ abandono\ del\ estudio - Tasa\ de\ cambio\ de\ estudio$$

$$Tasa\ de\ cambio\ de\ estudio = \frac{Cambio}{Estudiantes} * 100$$

$$Tasa\ de\ abandono\ del\ estudio = \frac{Abandono}{Estudiantes} * 100$$

Siendo:

- Abandono: Estudiantes con nuevo ingreso en el curso X y sin matrícula en los 2 cursos siguientes
- Cambio: Estudiantes con nuevo ingreso en el curso X y sin matrícula en los 2 cursos siguientes, matriculado en otro estudio en esos cursos

- Estudiantes: Estudiantes de nuevo ingreso en el curso X

Tal y como defiende Álvarez Ferrándiz (2021) en su Análisis del abandono universitario en España: un estudio bibliométrico (12) el abandono universitaria está relacionado con dimensiones sociales y económicas y puede provocar en el estudiante desmotivación, insatisfacción personal y/o baja autoestima, entre otros.

Los tipos de abandono, también conocido como *dropout*, *attrition*, *withdrawal* o *non-completer*, descritos en la literatura, también cuentan con el *stop out* o *break* que se asocia al descanso temporal de un estudiante que eventualmente podría retornar los estudios.

En el estudio bibliométrico de Álvarez Ferrándiz (2021) (12) se identifican cuatro modelos diferentes en los que el estudiantado termina abandonando sus estudios universitarios y son:

- **Modelo de adaptación**: la falta de adaptación al mundo universitario.
- **Modelo economista**: el estudiantado identifica que los costes del estudio no justifican el beneficio que podría obtenerse.
- **Modelo psicopedagógico**: en este caso, el abandono universitario viene motivado por características psicológicas del sujeto como la capacidad para enfrentarse a obstáculos y alcanzar metas o el desconocimiento de estrategias para aprender, entre otras.
- **Modelo estructural**: en este modelo, el abandono universitario es la consecuencia de todas las contradicciones de los distintos subsistemas (político, económico y social).

Las cifras más recientes que se han publicado por el INE (13) sobre el abandono universitario corresponden al estudiantado que iniciaron sus estudios en el curso académico 2018-19 son del 21,99 % en estudios de Grado y del 9,28 % en Máster; mientras que las tasas de cambio del estudio son, respectivamente, 8,79 % y 1,27 %. Es decir, las cifras globales de abandono del SUE serían del 13,2 % en Grado y del 8,01 % en Máster.

2.1.2. Aplicación de la terminología en los estudios a distancia.

Podría decirse que el inicio del SUE tuvo lugar con la creación de la primera universidad española, que fue la Universidad de Salamanca en 1218. Desde ese momento, el sistema fue creciendo progresivamente incorporando universidades y titulaciones hasta la actualidad.

No fue hasta 1972 cuando los avances tecnológicos permitieron crear la primera universidad a distancia en España, la Universidad Nacional de Educación a Distancia (UNED), lo que supuso permitir el acceso a la Universidad en condiciones de equidad e igualdad.

El predominio histórico y el volumen del estudiantado asociados a las enseñanzas presenciales han hecho que la definición de los indicadores oficiales del SIIU de aplicación al SUE den

respuesta, en mayor medida, a estas enseñanzas, no teniendo en cuenta las particularidades asociadas a las enseñanzas a distancia descritas en el apartado 1.1 del presente trabajo.

La metodología descrita en el Catálogo oficial de indicadores universitarios del SIIU (10) establece que la población para calcular los indicadores de rendimiento y abandono definidos en el apartado anterior (2.1.1) es, en mayor medida, **la población óptima con dedicación a lo largo del estudio a tiempo completo**.

Según el catálogo, (10) la definición de **la población óptima con dedicación a lo largo del estudio con dedicación a tiempo completo** está referida al estudiantado que son población óptima a lo largo del estudio y que han matriculado de media más de 45 créditos por curso. Mientras que las condiciones para que el estudiantado pertenezcan a la población óptima son:

En estudios de Grado:

Está referida al conjunto del estudiantado objetivo que empieza un grado y tiene que cursar prácticamente la totalidad de los créditos de ese estudio para ser graduado. Es decir, el estudiantado que:

- Deben ser Población de créditos.
- El número de créditos reconocidos (tanto totales como en el curso) en todos los cursos en los que se matricula en la titulación de grado debe ser < 30, excepto en caso del estudiantado proveniente de FP que se permitirá que tengan hasta 36 créditos reconocidos (< 36).
- El estudiante no puede constar como egresado en la misma titulación en ningún curso anterior.
- En el curso en el que acceden al grado los créditos matriculados, superados y presentados en el curso deben coincidir con los créditos matriculados, superados y presentados acumulados.
- El curso en el que el estudiante egresa, la suma de los créditos superados desde el inicio y los créditos reconocidos desde el inicio debe ser superior a 230 ECTS.
- Al egresar el estudiante debe haber cursado al menos 3 cursos la titulación. (Año curso en el que se titula – año acceso a la titulación) > 3.
- Hayan accedido al estudio de grado por cualquiera de las formas de acceso excepto las siguientes:
  - 09= Convalidación parcial de estudios extranjeros (al menos 30 créditos reconocidos).

- 10= Mediante traslado de Expediente de otro estudio de grado (al menos 30 créditos reconocidos).

En estudios de Máster:

Está referida al conjunto del estudiantado objetivo que empiezan un máster y tienen que completar la totalidad de créditos. A esta población la identificaremos como el conjunto del estudiantado que:

- Deben ser Población de créditos.
- El número de créditos reconocidos (tanto totales como en el curso) en todos los cursos en los que se matricula en la titulación de máster debe ser < 10.
- El estudiante no puede constar como egresado en la misma titulación en ningún curso anterior.
- En el curso en el que acceden al grado (Año curso del fichero de rendimiento = año de inicio de la titulación) los créditos matriculados, superados y presentados en el curso deben coincidir con los créditos matriculados, superados y presentados acumulados.

Dependiendo del perfil predominante del estudiantado de las universidades a distancia es posible que gran parte de ellos queden excluidos de la metodología de cálculo y, por tanto, sus indicadores sean poco representativos frente a los de las universidades presenciales. La mayoría de publicaciones relativas al abandono del SUE se centran en el tratamiento de datos de estudios presenciales, como es el caso del análisis realizado por Fernández-Mellizo, M. (2022) para la Secretaría General Técnica del Ministerio de Universidades (14) sin existir un homólogo para analizar el abandono de las universidades a distancia de España.

Como defiende Troche de Trevisan (2019) en su tesis doctoral titulada *Estudio del rendimiento académico del estudiante en línea como variable predictiva del abandono en educación superior: el caso de la Universitat Oberta de Catalunya* (15). 'En particular, la definición oficial de la deserción en España: dos semestres consecutivos sin matricularse, no refleja la naturaleza de la educación superior en línea (Josep Grau-Valldosera & Minguillón, 2014). Así pues, se favorece la definición empírica propuesta para la Universitat Oberta de Catalunya (UOC) que utiliza un número mínimo de semestres consecutivos de no matriculación para catalogar a un estudiante como desertor de un programa específico, es decir, cuando el estudiante toma N *breaks* o más, se asume que ha abandonado el programa, siendo N diferente para cada programa de estudios (Josep Grau-Valldosera & Minguillón, 2014). Tal particularidad podría ser atribuible a que la definición de abandono es bastante susceptible al contexto, incluso al micro contexto que suponen los diferentes programas de estudio en la misma universidad.'

2.1.3. Aplicación de la terminología en los estudios de la Udima.

Como se ha introducido en el apartado 2.1.1, la definición oficial establecida por el SIIU para las tasas de rendimiento y de abandono no se ajustan al comportamiento de la Udima, en el que la mayoría del estudiantado, por sus características principales, quedarían excluidos del cálculo porque no pertenecerían a la población óptima con dedicación a lo largo del estudio a tiempo completo.

En base a la definición descrita en el apartado anterior 2.1.2 sobre la población óptima con dedicación a lo largo del estudio a tiempo completo, se describen **qué características habituales del estudiantado de la Udima harían que quedasen excluidos de dicha población**:

- El estudiantado realiza matrícula parcial para adecuar los estudios a sus intereses personales (profesión, cuidado de familiares, etc.);
- El estudiantado realiza estudios a distancia para finalizar estudios universitarios inacabados (por lo que reconocen ECTS de sus titulaciones de origen);
- El estudiantado se matricula en nuevas menciones de su titulación anterior (Magisterios, máster en PRL, etc.);
- El estudiantado de Grado puede egresar sin haber cursado al menos 3 cursos la titulación (si se matricula para cursar especialidades concretas de la misma titulación).

Por los argumentos anteriormente expuestos, **la definición de los indicadores de rendimiento y abandono que se aplicarán a la Udima serán idénticos a los descritos en el catálogo oficial de indicadores del SIIU (10) exceptuando la aplicación de la población, que se referirá al total del estudiantado, no solo a los definidos por la población óptima con dedicación a lo largo del estudio a tiempo completo**.

Además, **para el presente trabajo relacionado con los datos de la Udima, de los indicadores descritos asociados al rendimiento académico se aplicarán la *Duración media de los estudios* y *Tasa de eficiencia (Rendimiento de los egresados universitarios)***. Los motivos de dicha decisión se detallan a continuación:

- Se aplicarán los siguientes indicadores:
    - Duración media de los estudios: como el juego de datos estará a nivel de estudiante y plan de estudios se podrá conocer la duración de los estudios para cada uno de ellos, lo que podrá ser una característica diferenciadora a la hora de clasificar al estudiantado, sobre todo, por los distintos perfiles que ingresan a las titulaciones

oficiales (estudiantado que trata de finalizar estudios inacabados, que quiere cursar nuevas menciones, etc.);

- Tasa de eficiencia (Rendimiento de los egresados universitarios): este indicador ofrecerá información sobre cómo es el rendimiento de un estudiante en la titulación, independientemente de las condiciones en las que haya iniciado la misma (estudiantado que trata de finalizar estudios inacabados, que quiere cursar nuevas menciones, etc.);
- Tasa de éxito de los egresados universitarios: *ídem* al argumento del indicador anterior;
- Tasa de evaluación de los egresados universitarios: *ídem* al argumento del indicador anterior.

- Se descartarán los siguientes indicadores:
  - Tasa de graduación: al haber optado por tratar a todos los perfiles posibles del estudiantado (descartando la población óptima que describe el SIIU en su metodología) este indicador deja de resultar interesante porque la duración prevista de los estudios sería distinta para cada estudiante en función de las características que presente a la hora de iniciar sus estudios, lo que dificultaría el cálculo y ya está recogido en los dos indicadores anteriores que se han decidido mantener.

En el presente trabajo, los indicadores descritos asociados al abandono universitario que se aplicarán para el estudio de la Udima serán la *Tasa de abandono del estudio* y *Tasa de cambio de estudio*. Los motivos de dicha decisión se detallan a continuación:

- Se aplicarán los siguientes indicadores:
  - Tasa de abandono del estudio: pese a que este indicador se encuentra mitigado por la Tasa de cambio de estudio, lo relevante a analizar es si el estudiantado abandona o no la universidad, no el SUE.
  - Tasa de cambio de estudio: aunque no se analizará el abandono global del SUE, se tendrá en cuenta este indicador para saber si sus estudios en la Udima son continuados o no en otra universidad, introduciendo como variable a analizar el traslado de expediente.

- Se descartarán los siguientes indicadores:

- Tasa de abandono del SUE: al centrar el estudio en la Udima, parte integrante del SUE, no se considera adecuado tratar el indicador que aporta datos a un nivel superior.

2.2. Métodos no supervisados (*clustering*).

2.2.1. Terminología.

Tal y como se puede comprobar en el explorador de conocimiento, culturas e ideas JSTOR (16) el término de *clustering* de datos surgió en 1954 en un artículo de datos antropológicos aunque la terminología es más antigua y aplica a distintas ramas del saber, donde en cada una de ellas, tiene su término correspondiente. En el caso que nos ocupa, asociado al *data science*, lo relacionaremos con la segmentación de datos y el aprendizaje no supervisado.

Los algoritmos no supervisados o de agrupación (*clustering*) descubren patrones y tendencias de los datos.

En 2012, Soni y Ganatra (1) detallan la agrupación de objetos según el tipo de datos y el objetivo a alcanzar para obtener información sobre la distribución de datos, generar hipótesis, observar características y detectar anomalías. Describen el proceso de agrupamiento y una descripción general de los diferentes métodos de agrupamiento según las propiedades de los agrupamientos generados entre los que destacan los siguientes dos grandes grupos (jerárquicos y de partición):

- **jerárquicos**: estos métodos descomponen el conjunto de datos de n objetos en una jerarquía de grupos. Se representa mediante un diagrama de estructura de árbol (dendrograma); cuyo nodo raíz representa todo el conjunto de datos y cada nodo hoja es un único objeto del conjunto de datos. Este método, a su vez, puede clasificarse en:
  - aglomerativo: de abajo hacia arriba.
    - Basados en métricas de vinculación: Los algoritmos representativos son *Single link: SLINK*; *Avg. link: Voorhe's Method* y *Complete link: CLINK*.
    - Otros: Los algoritmos representativos son *ROCK*, *CURE*, *CHEMELEON*, *AGNES* y *BIIRCH*.
  - división: de arriba hacia abajo. Los algoritmos representativos son *MONA* y *DIANA*.
- **de partición**: divide los datos en subconjuntos o particiones en función de los criterios de evaluación que dependen de:

- re-ubicación iterativa:
  - Basados en probabilidades (Modelo de agrupación): El modelo probabilístico clasifica los datos en varias poblaciones cuyas distribuciones guardan características comunes. Los algoritmos representativos son *EM*, *SNOB*, *AUTOCLASS* y *MCLUST*.
  - Métodos de Función objetivo:
    - Basados en K-means: El modelo construye puntos equidistantes o centros de cada *cluster*, los centroides se obtienen a partir de las observaciones. Los algoritmos representativos son *K-means*, *K-modes*, *K-prototypes*, *Fuzzy K-means*, *kernel k-means*, *bisección de k-means*, *ISODATA* y *Forgy*.
    - Basados en K-medoids: El modelo construye puntos equidistantes o centros de cada *cluster* respecto de los centroides, que no están construidos a partir de las observaciones, si no que están representados por uno de los objetos dentro de cada *cluster*. Los algoritmos representativos son *PAM*, *CLARA* y *CLARANS*.
- zonificación: utilizan espacios multidimensionales. Las agrupaciones se consideran regiones más densas que su entorno. Los algoritmos representativos son *WAVECLUST*, *STING*, *BANG*, *MAFIA* y *CLIQUE*.
- sub-espacios: destinados a juegos de datos de alta dimensionalidad. Los algoritmos representativos son *ENCLUS*, *OPTIGRID*, *OPTIGRID* y *ORCLUS*.
- densidad: los conjuntos de puntos agrupan puntos que están en regiones de alta densidad: Para cada punto de datos dentro de un grupo; la vecindad tiene que contener un número mínimo de puntos. Estos pueden clasificarse, a su vez, en:
  - Basados en la función de conectividad: Los algoritmos representativos son *SNN* y *DENCLUE*.
  - Basados en conectividad de puntos: Los algoritmos representativos son *DBSCAN*, *LDBSCAN*, *OPTICS*, *DBCLASD*, *ST-SCAN*, *VDBSCAN*, *GDBSCAN* y *DVBSCAN*.

Esta clasificación queda representada en la siguiente figura:

Un **método de agrupamiento** es una estrategia que permite resolver un problema de agrupamiento, mientras que un **algoritmo de agrupamiento** es la aplicación de un método.

El diseño y la selección del algoritmo a aplicar depende de las necesidades planteadas en el estudio y de los parámetros y criterios que condicionarán el mismo, ya que como detalló Kleinberg (2002) (17) en su teorema de imposibilidad 'ningún algoritmo de agrupamiento único

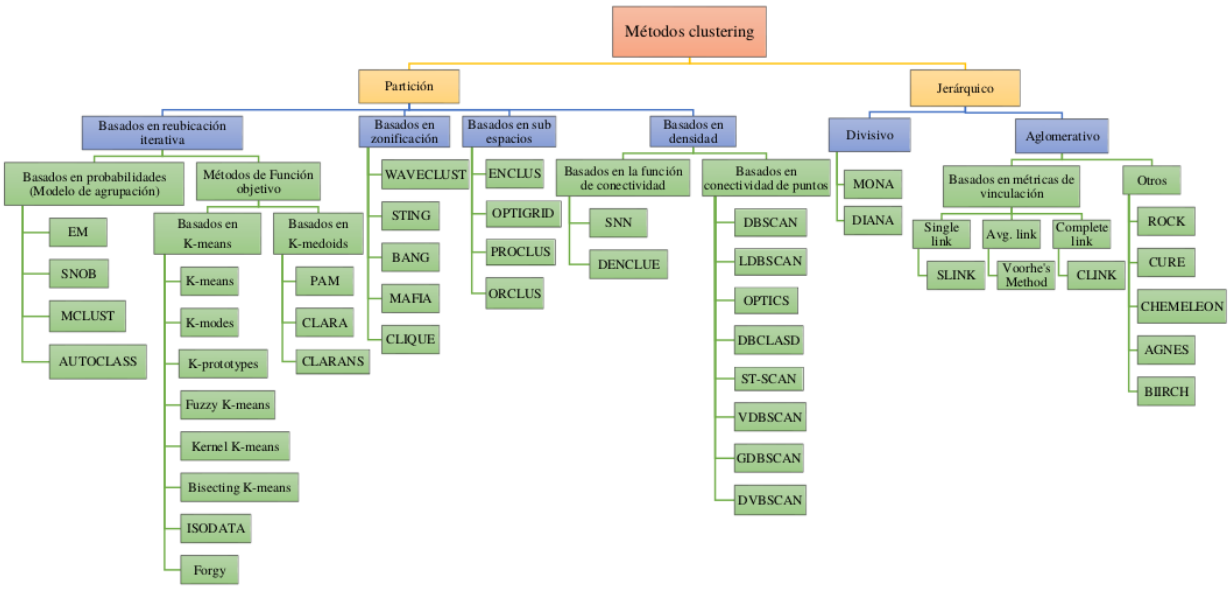

Figura 2.1: Categorización de diferentes métodos de agrupación. Fuente: Soni y Ganatra (2012) (1)

satisface simultáneamente los tres axiomas básicos del agrupamiento de datos: invariancia de escala, consistencia y riqueza'.

Por lo tanto, se debe evaluar el *cluster* según los distintos algoritmos de agrupamiento y parámetros aplicados para posteriormente analizar resultados y obtener conclusiones.

2.2.2. Clasificación del estudiantado universitario: desafíos y resultados.

Según la publicación de Muffo (1987) (18) acerca del estudio realizado al potencial estudiantado de una importante universidad reveló que los encuestados generalmente estaban motivados por consideraciones no económicas. Sin embargo, sí fue un factor determinante en el momento de realizar la matriculación universitaria, poniendo en desventaja competitiva a la institución objeto de estudio.

Fueron Angulo et al. (2010) (19) quienes defendieron que la segmentación del estudiantado de educación superior es única y atiende a grupos con prioridades significativamente diferentes (las diferencias de segmento eran complementarias, en lugar de contradictorias).

Story (2021) (20) sugiere en su estudio un enfoque de segmentación basado en la integración de factores racionales y emocionales que el futuro estudiantado valorará al seleccionar una universidad. **Esta referencia es determinante a la hora de diseñar el presente trabajo,**

teniendo en cuenta variables objetivas (racionales) y subjetivas (emocionales) para clasificar al estudiantado.

Según lo descrito por Wynd y Bozman (2010) (21) el estilo de aprendizaje del estudiantado se ha modelado como una variable moderadora en la eficacia y eficiencia del aprendizaje y puede relacionarse con rasgos demográficos identificables. En su artículo se demuestra que la edad y las calificaciones están significativamente relacionados con el estilo de aprendizaje del estudiantado. La segmentación según estas características aporta como resultado programas de educación alternativos que satisfacen las necesidades de sus electores.

Según el artículo de Chavez y Salinas (2021) (22) el análisis de datos en la educación es una tarea desafiante a la que se enfrentan las investigaciones con el fin de conocer, comprender y gestionar la diversidad del estudiantado que ingresa en las instituciones de educación superior para proponer estrategias educativas que mejoren el modelo de enseñanza-aprendizaje. El resultado de este artículo es, tras aplicar el algoritmo *K-prototypes*, caracterizar al estudiantado en 5 perfiles según sus variables sociodemográficas, económicas y de rendimiento académico. De esta forma, se contribuye a la mejora de las políticas de apoyo, impulsando cambios a favor de la calidad educativa, promoviendo la renovación de la docencia y creando programas de seguimiento personalizados según el perfil de estudiante. Esta referencia y la aplicación del algoritmo *K-prototypes* serán valoradas a la hora de diseñar la metodología del presente trabajo, ya que atiende a las características del juego de datos a utilizar, el cual cuenta con atributos numéricos y categóricos.

Como indican Norambuena et al. (2022) (4) en su estudio sobre los modelos predictivos basados en uso de analíticas de aprendizaje en educación superior, en la mayoría de estudios revisados, los métodos utilizados para la agrupación *clustering* se encuentran: Análisis de componentes principales (PCA), *clustering* de K-medias (K-M C), y agrupación jerárquica (HC).

2.3. Métodos supervisados (predicción).

2.3.1. Terminología.

Los algoritmos supervisados o predictivos predicen un atributo o etiqueta del juego de datos en base al resto de atributos descriptivos. A partir de datos etiquetados y relacionados con otros atributos se predicen datos no etiquetados. Esta metodología de trabajo se conoce como aprendizaje supervisado y consta de dos fases:

- **entrenamiento**: subconjunto de datos con etiqueta conocida.
- **test**: subconjunto de datos restante sobre el que se prueba el modelo.

El proceso completo consta de las siguientes etapas:

- **Determinación de objetivos**.
- **Preparación de los datos**: identificando las fuentes de información y preprocesando los datos para determinar la calidad de los mismos y las técnicas a utilizar.
- **Transformación de los datos**: los datos se convierten en un modelo analítico.
- **Minería de datos**: los datos se tratan automáticamente aplicando la combinación de algoritmos adecuados.
- **Análisis de resultados**: se interpretan los resultados obtenidos en el proceso anterior, preferiblemente, mediante técnicas de visualización.
- **Obtención de conocimiento**: se aplica el conocimiento obtenido.

El orden lógico es el indicado anteriormente aunque el proceso es iterativo y se establece retroalimentación entre los pasos.

2.3.2. Predicción del abandono según el rendimiento académico del estudiantado universitario.

Fueron Rovira, S. et al. (2017) (23) quienes afirmaron que el papel del tutor es determinante para evitar el abandono universitario del estudiantado y mejorar su rendimiento académico. Este estudio, mediante 5 técnicas de clasificación que utilizan diferentes enfoques para resolver un problema de clasificación: Logistic Regression (LR), Gaussian Naive Bayes (GB), Support Vector Machine (SVM), Random Forest (RF) y Adaptive Boosting (AdaBoost) utilizan el vector de características de las muestras del conjunto de entrenamiento, ayuda a los tutores a ofrecer una orientación personal proactiva. El sistema realiza predicciones de riesgo de abandono en función de las calificaciones del estudiantado.

De la interesante revisión sistemática realizada por Norambuena et al. (2022) (4) sobre los modelos predictivos basados en uso de analíticas de aprendizaje en educación superior se extraen las siguientes conclusiones reseñables obtenidas por los autores:

- **Variables a estudiar**: los modelos predictivos requieren incluir variables psicoeducativas para comprender mejor el fenómeno, no solo variables provenientes de la información estática y dinámica sobre el estudiantado y sus contextos de aprendizaje. De los estudios analizados, solo dos incluyeron variables sociodemográficas y otros dos, variables sociocognitivas (determinantes para explicar por qué d¡el estudiantado muestra sus patrones de comportamiento);

- **Entorno virtual de aprendizaje**: El uso de entornos virtuales de aprendizaje genera una huella digital que aporta una gran cantidad de información acerca de las actividades del estudiantado. La aparición de la huella digital tiene asociada una serie de nuevas áreas de estudio, tales como las analíticas o la minería de datos educacionales acerca del comportamiento del estudiantado en estos ambientes o sobre las interacciones que allí ocurren. En un contexto de alto interés por parte de las universidades para mejorar resultados de aprendizaje y éxito académico, la aparición de la analítica contribuye a la satisfacción de estas necesidades. El factor crucial que influye en la satisfacción del estudiantado es que las características disponibles en un LMS satisfaga sus necesidades y facilite su uso;

- **Poder predictivo del rendimiento académico**: El poder predictivo del rendimiento académico en modelos que incorporan las Analíticas de Aprendizaje (AA) es grande y por tanto prometedor para avanzar en las problemáticas actuales de la Educación Superior como el rendimiento académico y la permanencia;

- **Tipo de análisis**: Los métodos de clasificación utilizados con mayor frecuencia son: Clasificación de Bayes (BC), K-Vecino más cercano (KNN), máquina de vectores de soporte (SVM), árboles de clasificación (CT) y la regresión lineal. Otros estudios identifican que han usado en menor medida la red neuronal, la Distancias Levenshtein entre las cadenas relativas a trayectorias y el Algoritmo KNN;

- **Modalidad de estudios**: Todos los estudios analizados se corresponden con programas *blended learning* y *online*, no hay investigaciones de modelos predictivos de estudios presenciales.

Capítulo 3

Materiales y métodos

El presente proyecto se basa en la aplicación de técnicas de aprendizaje no supervisado (*clustering*) para clasificar al estudiantado universitario (de Grado y Máster) de la Udima en función de sus características principales (rendimiento académico, opiniones manifestadas y variables sociodemográficas) y de técnicas de aprendizaje supervisado para predecir el perfil del estudiantado que se encuentra en progreso.

3.1. Aspectos más relevantes del diseño y desarrollo del trabajo.

3.1.1. Aspectos relevantes del diseño del trabajo.

Como ha podido verse en el capítulo anterior (2), la mayoría de estudios similares atienden a variables objetivas de los sujetos a clasificar o predecir. **En el presente trabajo, el aspecto diferenciador es el de integrar variables subjetivas obtenidas a partir de las opiniones manifestadas por el estudiantado a lo largo de su vida académica mediante encuestas de satisfacción y estudios de inserción laboral**. En las encuestas de satisfacción, el estudiantado puede opinar sobre 3 entidades diferenciadoras: las asignaturas cursadas; el desarrollo del curso en la titulación en la que está matriculado/a (estudio iniciado en 2018); y el desarrollo de la totalidad de la titulación (encuesta al estudiantado titulado). Por otra parte, en el estudio de inserción laboral se recogen las variables correspondientes a las expectativas del estudiantado previo a su matriculación, así como la consecuencia que ha supuesto superar el programa.

3.1.2. Fases a desarrollar en el trabajo.

Tal y como se ha descrito en el apartado 1.4, la **metodología a emplear en el trabajo será la metodología CRISP-DM** (*Cross Industry Standard Process for Data Mining*) que permitirá crear un nuevo servicio para clasificar al estudiantado en función de sus características.

Las fases del ciclo de vida del proyecto según la metodología CRISP-DM han quedado introducidas en en el apartado 1.4 y se resumen a continuación:

- **Fase I. Definición de necesidades**: Se definen los objetivos generales y específicos y las técnicas necesarias a aplicar para conseguirlos.
- **Fase II. Estudio y comprensión de los datos**: Se estudian los datos disponibles.
- **Fase III. Análisis de los datos y selección de características**: Se preparan los datos (limpieza, normalización y discretización) según las necesidades detectadas en la Fase II que permitirán la siguiente fase de modelado (Fase IV).
- **Fase IV. Modelado**: Se modelan las técnicas (de aprendizaje no supervisado y supervisado) que son necesarias para resolver el objetivo principal del proyecto.
- **Fase V. Evaluación de resultados**: Se identifican las limitaciones y riesgos de aplicar el proyecto.
- **Fase VI. Despliegue (puesta en producción)**: Dado que la implementación del proyecto dependerá de la planificación e intereses de la organización, su puesta en producción no se evaluará en el alcance del presente proyecto.

3.1.3. Tecnologías utilizadas para el desarrollo del trabajo.

Las tecnologías escogidas para el desarrollo del proyecto serán *SqlDeveloper*, *dbeaver-ce* y *RStudio*. Las dos primeras son utilizadas para la obtención de información de las bases de datos de origen y la última, *RStudio*, para la gestión y almacenado de datos no disponibles en bases datos (satisfacción e inserción laboral) y la aplicación de técnicas de aprendizaje no supervisado (*clustering*) y aprendizaje supervisado.

En primer lugar, los datos de rendimiento académico y variables sociodemográficas son extraídas a través de consultas realizadas al programa de gestión académica de la universidad (Universitas XXI o UXXI) mediante *SqlDeveloper*.

La obtención de información a la plataforma de formación *Moodle* se consulta mediante el *software dbeaver-ce*.

A través de *RStudio* se gestiona y almacena la información no disponible en bases de datos oficiales y que se corresponde con las datos obtenidos en las encuestas de satisfacción y los estudios de inserción laboral, que se recogen mediante la plataforma de encuestas *LimeSurvey*. También en *RStudio* se gestiona la construcción del juego de datos global a utilizar en el proyecto y que relaciona los datos provenientes de las seis bases de datos mencionadas (gestión académica, *Moodle*, tres de satisfacción e inserción laboral).

Para las tareas asociadas a la aplicación de técnicas de aprendizaje no supervisado (*clustering*) y aprendizaje supervisado se utilizará *RStudio*.

3.2. Metodología empleada.

A continuación se describe la metodología (algoritmos, datos y librerías específicas seguidas) empleada para realizar el proyecto, describiendo las alternativas posibles, las decisiones tomadas y los criterios utilizados para cada una de las tareas principales.

3.2.1. Metodología empleada (algoritmos, datos y librerías específicas seguidas) para realizar el proyecto.

3.2.1.1. Tareas de ETL (*Extract, Transform and Load*).

Para las **tareas de ETL** se han utilizado las siguientes librerías y funciones:

- Para la **carga y guardado de datos**:
  - Las funciones *read_csv* y *read_delim* de la librería *readr* para cargar ficheros .csv;
  - La función *read_excel* de la librería *xlsx* para cargar ficheros .xlsx;
  - La función *load* de la librería *base* para cargar ficheros .Rda (tipo de base de datos generada en R, correspondiente a los históricos de resultados de satisfacción e inserción laboral);
  - La función *write_csv* de la librería *readr* para guardar ficheros .csv;
  - La función *save* de la librería *base* para guardar ficheros .Rda.
- Para la **transformación de datos**:
  - Las siguientes funciones de la librería *stringr*:
    - *str_replace_all*, para sustituir espacios y puntuaciones en los comentarios realizados por el estudiantado en las distintas encuestas/estudios;

  - *substr*, para extraer caracteres de una variable dada una condición;
  - *str_to_lower* para transformar a minúsculas el nombre de las variables;
- Las funciones *as.numeric*, *as.character* y *as.factor* de la librería *base* para, respectivamente, convertir una variable en un tipo de datos: numérico, carácter y factor/categórico;
- La función *cut* de la librería *base* para discretizar variables, categorizando los valores de una variable continua en diferentes niveles de un factor (definiendo el número de grupos o *breaks* y los niveles del factor, definiendo *labels*);
- La función *ifelse* de la librería *base* para dicotomizar variables, categorizando los valores de una variable en valores 1 y 0;
- La función *dummy_cols* de la librería *fastDummies* para transformar las variables cualitativas mediante la técnica *One Hot Encoding* (obteniendo una columna con valores dicotómicos, 0-1, por cada variable y categoría);
- La función *scale* de la librería *base* para normalizar las variables numéricas pertenecientes a distintas escalas para que todas aporten la misma importancia en el cálculo de los centroides, centrándolas (media 0 y desviación típica 1);
- La función *melt* de la librería *reshape2* para en un mismo *dataframe*, transformar varias variables en una única;
- La función *rbind* de la librería *SparkR* para fusionar en uno varios *dataframes* con la misma estructura;
- Las siguientes funciones de la librería *dplyr*:
  - *rename*, para renombrar variables;
  - *select*, para seleccionar variables;
  - *filter*, para filtrar valores dada una o varias condiciones;
  - *group_by*, para agrupar datos;
  - *summarise*, para calcular variables;
  - *arrange*, para ordenar resultados de un juego de datos;
  - y *anti_join* para eliminar de una juego de datos los contenidos en otro juego;
- La función *reduce(left_join)* de la librería *reshape2* para añadir atributos de una tabla a otra a partir de variable/s común/es;
- Las funciones *removeWords*, *get_tokens* y *get_nrc_sentiment* de las librerías *syuzhet* y *tm* para aplicar técnicas de procesamiento del lenguaje natural NLP mediante diccionario de léxico *NRC* y extraer la connotación (positiva o negativa, en rango -1:1) de los comentarios realizados por el estudiantado en las distintas encuestas/estudios.

- Para la **comprobación de la calidad de los datos**:
  - De la librería *base*, se aplican las funciones;
    - *unique* para comprobar los valores únicos de una variable;
    - *dim* para comprobar las dimensiones (filas y columnas) de un *dataframe*;
    - *sort* para ordenar valores de una variable;
    - *round* para redondear los valores numéricos de una variable según el número de dígitos indicado;
    - *sapply* combinada con *range* para comprobar los valores extremos de todas las variables numéricas de un *dataframe*;
    - *colSums* combinada con *is.na* para comprobar el total de valores ausentes de las variables de un *dataframe*.

3.2.1.2. Análisis de datos.

Para el **análisis de datos** se han utilizado las siguientes librerías y funciones:

- La función *summary* de la librería *base* para mostrar las estadísticas básicas de las variables de un *dataframe*;
- La función *dfSummary* de la librería *summarytools* para mostrar las estadísticas básicas de las variables de un *dataframe*, su distribución y la cantidad de valores faltantes;
- La función *datatable* de la librería *DT* para mostrar y poder navegar por un *dataframe*;
- La función *boxplot* de la librería *graphics* para mostrar gráficamente la distribución de una variable respecto de otra;
- Para el **análisis de correlaciones** se aplican las siguientes funciones y librerías;
  - *cor* de la librería *stats* para calcular el coeficiente de correlación entre las variables numéricas y determinar si es estadísticamente significativo (*p-value*);
  - *diag* de la librería *base* para sustituir la correlación de la diagonal por 0;
  - *chart.Correlation* de la librería *PerformanceAnalytics* para representar gráficamente las correlaciones (debido a las dimensiones del *dataframe* no se visualiza correctamente);
  - *tab_corr* de la librería *sjPlot* para mostrar y poder navegar por la matriz de correlaciones.

3.2.1.3. Reducción de la dimensionalidad.

Para la **reducción de la dimensionalidad** se han utilizado las siguientes librerías y funciones:

- Para el **análisis PCA (Análisis de Componentes Principales)**:
  - Mediante la función *prcomp* de la librería *stats* a partir de datos escalados, centrados y sin valores ausentes, se conoce la cantidad de varianza explicada que aporta cada variable y la varianza explicada acumulada del juego de datos;
  - Mediante las funciones *get_pca_var* y *get_pca* de la librería *factoextra* ordena las variables por orden de contribución.
- Para el **análisis SVD (*Single Value Decomposition*)**:
  - Al combinar las funciones *cumsum* y *ggplot* de las librerías *base* y *ggplot*, respectivamente, se muestra mediante gráfica cómo varía la varianza explicada acumulada en las variables del juego de datos;
  - Mediante la función *fviz_screeplot* del paquete *factoextra* se genera un *scree_plot* que muestra los *eigenvalores* (de mayor a menor). De esta forma se obtienen 10 dimensiones, es decir, las variables a analizar se agrupan en 10 dimensiones de modo que para cada agrupación la variación de las varianzas está optimizada;
  - La función *fviz_pca_var* muestra el gráfico con la contribución de variables PCA;
  - La función *fviz_contrib* representa las variables que están por encima de la media para las 2 dimensiones principales.

3.2.1.4. Objetivo específico 1: Identificar los perfiles del estudiantado que finaliza o abandona sus estudios.

Para **identificar los perfiles del estudiantado que finaliza o abandona sus estudios** (objetivo específico 1) se aplican **técnicas de aprendizaje no supervisadas basados en el concepto de distancia**, métodos de agregación o análisis *cluster*.

- Mediante la función *clValid* del paquete *clValid* se evalúan distintas métricas y métodos para conocer el algoritmo y número óptimo de *clusters* según las características del juego de datos a estudiar.
- Mediante la función *fviz_nbclust* de la librería *factoextra*, se obtiene el número óptimo de agrupaciones o *clusters* mediante dos métodos: *elbow* o del codo (method = "wss") y método de silueta promedio (method = "silhouette").

- Mediante la función *set.seed* de la librería *base* se define una raíz para forzar a que los valores de la muestra aleatoria sean siempre los mismos, y por tanto, los resultados sean reproducibles.

- La función *kmeans* del paquete *stats* permite entrenar el modelo y obtener la correspondencia con las agrupaciones indicadas y se fuerza al algoritmo a elegir 100 conjuntos de centros de inicio aleatorios (precisando el argumento *'nstart = 100'*).

- Se visualizan los *clusters* mediante la función *clusplot* de la librería *cluster*.

- Se obtiene una estimación de la calidad del agrupamiento mediante la función *silhouette* del paquete *cluster*. Para ello, previamente, es necesario calcular la matriz de disimilitud con la función *daisy* del paquete *cluster*.

- Se visualiza la clasificación óptima mediante la función *fviz_cluster* del paquete *factoextra*.

3.2.1.5. Objetivo específico 2: Determinar las características principales que definen a un estudiante que finaliza o abandona sus estudios.

Para **determinar las características principales que definen a un estudiante que finaliza o abandona sus estudios** (objetivo específico 2) se muestran y analizan las **características principales de cada grupo según las estadísticas básicas de las variables**.

- La función *boxplot* de la librería *graphics* muestra gráficamente la distribución de las variables más representativas respecto del tipo de agrupamiento o *cluster*, para identificar y comparar visualmente sus características;

- Además, para cada uno de los cuatro perfiles del estudiantado, mediante la función *dfSummary* de la librería *summarytools* se muestran las estadísticas básicas de las variables y su distribución;

- A partir de las características principales se asigna una etiqueta que define, sintetiza e identifica a cada perfil.

3.2.1.6. Objetivo específico 3: Predecir la clasificación del estudiantado que no ha finalizado los estudios.

Para **predecir la clasificación del estudiantado que no ha finalizado los estudios** (objetivo específico 3) se aplican **técnicas de aprendizaje supervisadas** para valorar cuál es el algoritmo de clasificación que aporta mejores resultados. Para **determinar el mejor**

modelo se aplicará validación cruzada, optando por el que mayor calidad de predicción (*accuracy*) aporte.

Los métodos de clasificación tienden al sobre entrenamiento u *overfitting*, para evitarlo, se aplica *10-Fold-Cross-Validation*, que consiste en:

- Mediante la función *set.seed* de la librería *base* se define una raíz para forzar a que los valores de la muestra aleatoria sean siempre los mismos, y por tanto, los resultados sean reproducibles;
- Se parte de los datos originales y se dividen en conjuntos de entrenamiento (*train*), de *test* y de validación, para ello:
  - La función *nearZeroVar* de la librería *caret* identifica variables con valor único (predictores de varianza cero) para posteriormente, eliminarlas del juego de datos que entrenará el modelo;
  - Con ayuda de la función *sample.split* de la librería *caTools* se obtienen muestras proporcionadas en base al tipo de clasificación. Se crea un vector de partición sobre la variable de clasificación que dividirá la muestra al 80 % para el conjunto *train* y al 20 % para el *test* (los conjuntos de *train* y *test* cuentan con la misma proporción de datos de cada perfil).
- Para aplicar la validación cruzada (*10-Fold-Cross-Validation*) se asigna a una variable el resultado de la función *trainControl* del paquete *caret* precisando los argumentos *method = "cv"* y *number = 10* para introducirla como parámetro en la validación de los algoritmos a valorar;
- Para determinar el algoritmo y la métrica óptima que maximice el *accuracy* o exactitud del modelo se hace uso de la función *train* del paquete *caret*. De esta forma, se comparan el método *k-NN* y *Random forest*. Para ello:
  - En **el método *k-NN***:
    - El algoritmo *k-NN* mide distancias entre conjuntos de datos para identificar patrones sin un aprendizaje específico;
    - Se aplican los argumentos *method = "knn"*, *tuneGrid = expand.grid(k = seq(3, 15, 2))*, para indicar valores impares de k y evitar empates, *trControl* para aplicar *10-Fold-Cross-Validation* y *metric = "Accuracy"* como métrica comparativa de los algoritmos a la función *train*.

- En **el método bosque aleatorio de decisión (*Random Forest*)**:
  - *Random forest* consiste en aplicar de manera iterativa el algoritmo de árboles de decisión con diferentes parámetros sobre los mismos datos aplicando el promedio de muchos modelos reduciendo la variabilidad final del conjunto. Esta iteración crea modelos más robustos de los que se obtendrían creando un solo árbol de decisión;
  - Se aplican los argumentos *method = "rf"*, *trControl* para aplicar *10-Fold-Cross-Validation* y *metric = "Accuracy"* como métrica comparativa de los algoritmos a la función *train*. La función *train* aplica *tunegrid* automáticamente para el *method = "rf"*;
- **Se opta por el modelo y métrica que mayor *accuracy* reporta**.

- Conocido el algoritmo con óptimos resultados, se entrena el juego de datos de entrenamiento para predecir su categoría a partir de la función *randomForest* del paquete *randomForest*, precisando los argumentos *y_train:* (variable a predecir), *rto_fin_abandono_train* (juego de datos a partir del que se predice), *ntree = 100* (numero de árboles en el bosque) y *mtry = x* (máximo de variables en modelos, siendo x el valor óptimo obtenido en el proceso de validación);

- A partir de la función *varImpPlot* de la librería *randomForest* se grafica la importancia de las variables del modelo para determinar cuáles tienen mayor correlación con la variable a predecir;

- Posteriormente, se realiza la predicción sobre el conjunto de *test* mediante la función *predict* del paquete *stats* y se obtiene la bondad de su ajuste mediante la función *mean* (librería *base*) de la igualdad entre la predicción y el valor real de la variable (*y_test*);

- Tras asegurar que la bondad del ajuste del conjunto de *test* es similar a la *accuracy* o exactitud del modelo sobre el conjunto *train*, se realiza la predicción sobre el conjunto de datos del estudiantado en progreso (de manera análoga al paso anterior).

3.2.1.7. Evaluación de los resultados obtenidos.

Para **evaluar el resultado obtenido se comparan los resultados de la clasificación frente a los de la predicción** a partir de:

- La construcción y análisis de una tabla comparativa que muestra los resultados obtenidos para cada etiqueta (o tipo de estudiante), el método (clasificación o predicción) y las estadísticas básicas de las variables principales;

- Se evalúa si los resultados obtenidos son coherentes con el problema planteado;
- En caso negativo, se realizan los ajustes correspondientes en el modelo y se vuelven a obtener resultados;
- En caso afirmativo, se valida el modelo.

3.2.2. Alternativas posibles y decisiones tomadas.

De manera paralela al desarrollo de la metodología empleada se han ido aplicando las siguientes decisiones en las distintas fases del proyecto:

En la fase inicial de **obtención de datos**, la técnica de procesamiento del lenguaje natural NLP aplicada para el análisis de sentimientos manifestados por el estudiantado en los comentarios de los distintos estudios y encuestas se ha realizado con el *diccionario de léxico NRC*. Previamente también se valoró aplicar el diccionario AFFIN obteniendo resultados similares. Se comprobó que ambos diccionarios no siempre aportan un valor adecuado para el contenido del comentario al analizar los términos de manera aislada, por lo que se producen sesgos, obteniendo valoraciones que no se corresponden con el contenido o connotación real del comentario. Además, las ironías o metáforas utilizadas en estos comentarios no son correctamente evaluadas. Esto se considera una limitación.

En la fase de **transformación de datos**, inicialmente se aplicó la técnica *One Hot Encoding* a las variables categóricas. Esto suponía que las variables del juego de datos aumentasen considerablemente. El tamaño y la variabilidad de este juego de datos introducía mucho ruido. Para evitarlo, **se redujo la dimensionalidad del juego de datos tomando las siguientes decisiones**:

- Convirtiendo variables categóricas en dicotómicas:
  - *ste_codalf*, toma el valor 0 si es 'Grado' y 1 si es 'Máster';
  - *sexo*: 0, 'Hombre' y 1, 'Mujer';
  - *nacionalidad*: 1, 'española' y 0, 'extranjera' (sin esta transformación se contaba con 87 categorías);
  - *acceso*: 1, 'Titulado Universitario' y 0, 'No' (sin esta transformación se contaba con 166 categorías);
  - *complementos*: 1, 'Sí' y 0, 'No';
  - *traslado_expediente*: 1, 'Sí' y '0', No (sin esta transformación se contaba con 3 categorías);

- *il_motivo_estudios*: 1 si 'Mejora laboral', 0 si 'Otros' (sin esta transformación se contaba con 5 categorías);
- *il_consecuencia_estudios*: 1 si 'Conseguir o mejorar un empleo', 0 si 'No variación u Otro' (sin esta transformación se contaba con 8 categorías);
- *il_trabajo_actual*: 1, 'Sí' y 0, 'No';
- *il_trabajo_actual_relacionado_estudios*: 1, 'Sí' y 0, 'No';

- Discretizando variables categóricas en rangos:
  - *any_anyaca_acceso*, rangos '1965-80', '1981-95', '1996-2010' y '2011-25';
  - *any_anyaca_inicio*: rangos '2008-13', '2014-19' y '2020-25'.

- Por último, **se aplica la técnica *One Hot Encoding* a las 3 variables categóricas** (*any_anyaca_acceso*, *any_anyaca_inicio* y *cen_codnum*);

- Aplicando las técnicas anteriores, **se consigue disminuir la dimensionalidad del juego de datos de partida hasta un total de 68 variables**.

- En cuanto a la calidad de los datos, se comprueba la existencia de valores ausentes y **se decide imputar datos en función del significado de las variables**: algunas de ellas, al no tener valor, se imputa 0; y otras, que su valor es desconocido, principalmente porque el estudiante no ha participado en las encuestas, se imputa la mediana (métrica robusta a los *outliers*). Los motivos y las variables a las que aplica son las siguientes:
  - **Se imputa 0 en las variables que aún no han registrado valor**: *sa_titulacion_R* y *sa_titulacion_participacion*, relacionadas con el total del estudiantado que no ha egresado ni participado en el estudio; *il_participacion* e *il_R* total del estudiantado que no ha egresado ni participado en el estudio; si no se han registrado comentarios en las encuestas (el valor 0 aporta connotación neutra en los sentimientos expresados); las variables dicotómicas del estudio de inserción se imputa el valor 'No' = 0;
  - **Se imputa la mediana a las variables que, pese a no tener valor, imputar 0 como en el caso anterior supondría tener el peor valor posible del aspecto medido**:
    - *nota_media*: estudiantado, en progreso o que ha abandonado, que no ha registrado calificaciones;

- *sa_asig_participacion* total del estudiantado que su participación no está recogida en el histórico (se recuerda que este no cuenta con datos anteriores a 2012-13 y *sa_curso_participacion* total del estudiantado que su participación no está recogida en el histórico (se recuerda que el estudio se inició en 2017-18. **Esto se considera una limitación**;
- Los distintos ítems de las encuestas de satisfacción de asignaturas (*sa_asig_x*) tienen distinto nivel de participación debido a que el formato ha variado en el tiempo y no es obligatorio responder a ningún ítem;
- Los distintos ítems de las encuestas de satisfacción con el curso (*sa_curso_x*) tienen distinto nivel de participación debido a que el formato ha variado en el tiempo y no es obligatorio responder a ningún ítem;
- Los distintos ítems de las encuestas de satisfacción de egresados con la titulaciones (*sa_titulacion_x*) tienen distinto nivel de participación debido a que el formato ha variado en el tiempo y no es obligatorio responder a ningún ítem;
- Las variables de utilización de la plataforma de formación *Moodle* (*dias_laborable, dias_festivo, horario_madrugada, horario_mañana, horario_tarde, horario_noche*) coinciden en valores, total del estudiantado que no ha registrado movimiento en *Moodle* entre los cursos recogidos ("2012-13" a "2020-21"). **Esto se considera una limitación**.

De nuevo, para reducir el ruido de los datos y facilitar la clasificación de los mismos, se decide **reducir la dimensionalidad del juego de datos a partir de los resultados obtenidos en los análisis PCA y SVD**: con 33 de las 68 variables disponibles en el juego de datos se consigue obtener el 85 % de la varianza acumulada. Estas decisiones afectaron a las fases posteriores, donde inicialmente se había valorado clasificar al estudiantado mediante *k-prototypes*, que admite variables categóricas.

Para la consecución del objetivo principal se han tomado las siguientes decisiones en cada uno de los objetivos específicos:

- **Justificación de la técnica idónea a aplicar para resolver la clasificación definida en el objetivo específico 1**: Pese a que la estimación de la calidad del agrupamiento y la comparación entre *k-means-PAM* entre 2 y 5 *clusters* determina que los resultados óptimos de la técnica de aprendizaje no supervisado basada en el concepto de distancia se obtienen con el algoritmo *k-means*, distancia euclidiana y 2 *clusters*, **se descarta agrupar en 2 *clusters* y se decide agrupar en el siguiente nivel máximo: 4 *clusters*, con el algoritmo *k-means* y métrica: distancia euclidiana** como sugieren el *método elbow* (el codo) y el método de la silueta promedio y que, dado el problema planteado

permite una clasificación más rica en detalles, ya que agrupar en 2 *clusters* significaría obtener un grupo de estudiantado egresado y otro de abandono.

- Para **determinar las características principales que definen a un estudiante que finaliza o abandona sus estudios** (objetivo específico 2) **se determina una etiqueta que define de manera abreviada las características principales de cada grupo** según las estadísticas básicas de las variables.

- **Justificación de la técnica idónea a aplicar para resolver la predicción de clasificación definida en el objetivo específico 3**: Se comparan distintas técnicas de aprendizaje supervisadas (*k-NN* y *Random forest*) aplicando *10-Fold-Cross-Validation* y se determina que **el algoritmo de clasificación que aporta mejores resultados es *Random forest* atendiendo a la máxima *accuracy* o exactitud**. En la evaluación de resultados del primer modelo se obtiene una bondad del ajuste muy alta, así como un bajo porcentaje de estudiantado clasificado en riesgo de abandono. En el primer modelo, el gráfico de importancia de las variables muestra que **las variables más importantes del modelo están relacionadas con el egreso/abandono del estudiantado. Mantenerlas en el modelo con el fin de predecir la clasificación del estudiantado en progreso se considera una limitación**. Al estar en progreso, es lógico que sus resultados se vean condicionados y no sean comparables. Por dicho motivo, **se decide repetir el proceso eliminando las variables que dependen del egreso u abandono del estudiantado**, para no condicionar la clasificación del estudiantado en progreso. De esta forma, **se reduce la *accuracy* del modelo y aumenta el porcentaje de predicción del estudiantado que se encuentra en riesgo de abandono. Una vez comparados con los resultados de la literatura, se da por válido el modelo**.

A continuación, y a modo resumen, **se facilita una tabla descriptiva del juego de datos empleado en cada una de las fases del proyecto con el detalle del tipo de imputación de datos aplicado**.

Variable	Tipo	Explotación	Imputación	Tras SVD	Tras clasificación
est_codmec	Categórica	No	–	No	No
pla_codalf	Categórica	No	–	No	No
exp_numord	Categórica	No	–	No	No
id_usuldap	Categórica	No	–	No	No
numdocsinletra	Categórica	No	–	No	No
ste_codalf	Cardinal	Sí	–	Sí	Sí
cen_codnum	Categórica	Sí	–	No	No
sexo	Cardinal	Sí	–	No	No
edad_inicio	Cardinal	Sí	–	Sí	Sí
nacionalidad	Cardinal	Sí	–	No	No
any_anyaca_acceso	Categórica	Sí	–	No	No
any_anyaca_inicio	Categórica	Sí	–	No	No
titulado	Cardinal	Sí	–	Sí	No
tasa_eficiencia	Cardinal	Sí	–	Sí	Sí
t_exito	Cardinal	Sí	–	Sí	Sí
t_evaluacion	Cardinal	Sí	–	Sí	Sí
t_rec	Cardinal	Sí	–	Sí	Sí
t_sup_req	Cardinal	Sí	–	Sí	No
acceso	Cardinal	Sí	–	Sí	Sí
complementos	Cardinal	Sí	–	No	No
traslado_expediente	Cardinal	Sí	–	No	No
n_nodos_fin_titulacion	Cardinal	Sí	–	Sí	No
duracion_media_estudios	Cardinal	Sí	–	Sí	No
tasa_abandono	Cardinal	Sí	–	Sí	No
nota_media	Cardinal	Sí	mediana	Sí	Sí
sa_asig_R	Cardinal	Sí	mediana	Sí	Sí
sa_asig_participacion	Cardinal	Sí	mediana	Sí	Sí
sa_asig_organizacion	Cardinal	Sí	mediana	Sí	Sí
sa_asig_manual	Cardinal	Sí	mediana	Sí	Sí
sa_asig_materiales	Cardinal	Sí	mediana	Sí	Sí
sa_asig_actividades	Cardinal	Sí	mediana	Sí	Sí
sa_asig_docente	Cardinal	Sí	mediana	Sí	Sí
sa_asig_aprendiendo	Cardinal	Sí	mediana	Sí	Sí
sa_asig_comentarios	Cardinal	Sí	0	No	No
sa_curso_R	Cardinal	Sí	mediana	Sí	Sí
sa_curso_participacion	Cardinal	Sí	mediana	Sí	Sí
sa_curso_nivel_academico	Cardinal	Sí	mediana	No	No
sa_curso_implicacion	Cardinal	Sí	mediana	No	No
sa_curso_aprovechamiento_recursos	Cardinal	Sí	mediana	No	No
sa_curso_titulacion	Cardinal	Sí	mediana	Sí	Sí
sa_curso_recomendacion	Cardinal	Sí	mediana	No	No
sa_curso_comentarios	Cardinal	Sí	0	No	No
sa_titulacion_R	Cardinal	Sí	0	Sí	No
sa_titulacion_participacion	Cardinal	Sí	0	Sí	No
sa_titulacion_nivel_academico	Cardinal	Sí	mediana	No	No
sa_titulacion_implicacion	Cardinal	Sí	mediana	Sí	No
sa_titulacion_aprovechamiento_recursos	Cardinal	Sí	mediana	No	No
sa_titulacion_mejora_profesional	Cardinal	Sí	mediana	No	No
sa_titulacion_obtener_trabajo	Cardinal	Sí	mediana	No	No
sa_titulacion_titulacion	Cardinal	Sí	mediana	No	No
sa_titulacion_recomendacion	Cardinal	Sí	mediana	No	No
sa_titulacion_comentarios	Cardinal	Sí	0	No	No
il_R	Cardinal	Sí	0	Sí	No
il_participacion	Cardinal	Sí	0	Sí	No
il_motivo_estudios	Cardinal	Sí	mediana	Sí	No
il_consecuencia_estudios	Cardinal	Sí	mediana	Sí	No
il_trabajo_actual	Cardinal	Sí	mediana	Sí	No
il_trabajo_actual_relacionado_estudios	Cardinal	Sí	mediana	Sí	No
il_comentarios	Cardinal	Sí	0	No	No
dias_laborable	Cardinal	Sí	mediana	No	No
dias_festivo	Cardinal	Sí	mediana	No	No
horario_mañana	Cardinal	Sí	mediana	No	No
horario_tarde	Cardinal	Sí	mediana	No	No
horario_noche	Cardinal	Sí	mediana	No	No

Tabla 3.1: Variables utilizadas en el proyecto.

La definición de las variables no identificativas son:

- ***ste_codalf***: Tipo de estudios oficiales (0 si es 'Grado' y 1 si es 'Máster')
- ***cen_codnum***: Código MEC del centro
- ***sexo***: Sexo (0, 'Hombre' y 1, 'Mujer')
- ***edad_inicio***: Edad de inicio de los estudios
- ***nacionalidad***: 1ª Nacionalidad (1, 'española' y 0, 'extranjera')
- ***any_anyaca_acceso***: Curso de la vía de acceso presentada (rangos '1965-80', '1981-95', '1996-2010' y '2011-25')
- ***any_anyaca_inicio***: Curso de inicio de los estudios oficiales (rangos '2008-13', '2014-19' y '2020-25')
- ***titulado***: Titulación finalizada (1, 'Sí' y 0, 'No')
- ***tasa_eficiencia***: 100xECTS superados/ECTS matriculados
- ***t_exito***: 100xECTS superados/ECTS presentados
- ***t_evaluacion***: 100xECTS presentados/ECTS matriculados
- ***t_rec***: 100xECTS reconocidos/total ECTS a superar
- ***t_sup_req***: 100xECTS (superados+reconocidos)/total ECTS a superar
- ***acceso***: Tipo de acceso presentado (1, 'Titulado Universitario' y 0, 'No')
- ***complementos***: Requiere complementos (1, 'Sí' y 0, 'No')
- ***traslado_expediente***: Traslado de expediente (1, 'Sí' y 0, 'No')
- ***n_nodos_fin_titulacion***: Nº de nodos de finalización (si 2 o más, obtiene distintas especialidades)
- ***duracion_media_estudios***: Número de cursos donde formaliza matrícula
- ***tasa_abandono***: 0: si se titula, no se traslada y tiene matrícula activa en dos últimos cursos; 1: en el resto de casos
- ***nota_media***: Media ponderada de los ECTS superados en Udima

- ***sa_asig_R***: Representatividad de la muestra de encuestas respondidas por el estudiante sobre las asignaturas

- ***sa_asig_participacion***: Proporción de encuestas respondidas por el estudiante sobre las asignaturas del plan respecto de las que han sido enviadas

- ***sa_asig_organizacion***: Promedio de valoraciones aportadas por el estudiante sobre las asignaturas del plan para el ítem *La organización de la asignatura facilita el aprendizaje de la misma*. Escala 1-5

- ***sa_asig_manual***: Promedio de valoraciones aportadas por el estudiante sobre las asignaturas del plan para el ítem *El manual facilita el aprendizaje de esta asignatura*. Escala 1-5

- ***sa_asig_materiales***: Promedio de valoraciones aportadas por el estudiante sobre las asignaturas del plan para el ítem *Otros materiales didácticos aportados facilitan el aprendizaje de esta asignatura*. Escala 1-5

- ***sa_asig_actividades***: Promedio de valoraciones aportadas por el estudiante sobre las asignaturas del plan para el ítem *Las actividades didácticas planteadas facilitan el aprendizaje de esta asignatura*. Escala 1-5

- ***sa_asig_docente***: Promedio de valoraciones aportadas por el estudiante sobre las asignaturas del plan para el ítem *La labor del docente facilita el aprendizaje de esta asignatura*. Escala 1-5

- ***sa_asig_aprendiendo***: Porcentaje de respuestas Sí al ítem *¿Consideras que estás aprendiendo?* respecto de las encuestas respondidas por el estudiante sobre las asignaturas. Escala 0-100 (ítem introducido en 2017-18)

- ***sa_asig_comentarios***: Promedio de la connotación del contenido de los comentarios realizados por el estudiante en las encuestas de las asignaturas. Escala -1:1 (léxico NRC)

- ***sa_curso_R***: Representatividad de la muestra de encuestas respondidas por el estudiante sobre los cursos (encuesta iniciada en 2017-18)

- ***sa_curso_participacion***: Proporción de encuestas respondidas por el estudiante sobre los cursos matriculados en el plan respecto de las que han sido enviadas (encuesta iniciada en 2017-18)

- ***sa_curso_nivel_academico***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción los cursos matriculados en el plan desde 2017-18 para el ítem *Mi nivel académico y de conocimientos a la hora de afrontar los estudios*. Escala 1-5

- ***sa_curso_implicacion***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción los cursos matriculados en el plan desde 2017-18 para el ítem *Mi implicación con los estudios*. Escala 1-5

- ***sa_curso_aprovechamiento_recursos***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción los cursos matriculados en el plan desde 2017-18 para el ítem *Mi aprovechamiento de los recursos didácticos disponibles*. Escala 1-5

- ***sa_curso_titulacion***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción los cursos matriculados en el plan desde 2017-18 para el ítem *Grado de satisfacción general con la titulación*. Escala 1-5

- ***sa_curso_recomendacion***: Proporción de respuestas Sí al ítem *¿Recomendarías estudiar en la Udima a otra persona?* respecto de las encuestas respondidas por el estudiante sobre los cursos matriculados en el plan desde 2017-18. Escala 0-100

- ***sa_curso_comentarios***: Promedio de la connotación del contenido de los comentarios realizados por el estudiante en las encuestas respondidas por el estudiante sobre los cursos matriculados en el plan desde 2017-18. Escala -1:1 (léxico NRC)

- ***sa_titulacion_R***: Representatividad de la muestra de encuestas respondidas por el egresado en los cursos de egreso

- ***sa_titulacion_participacion***: Proporción de encuestas respondidas por el egresado en los cursos de egreso en el plan respecto de las que han sido enviadas

- ***sa_titulacion_nivel_academico***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción de egresados en el plan para el ítem *Mi nivel académico y de conocimientos a la hora de afrontar los estudios*. Escala 1-5

- ***sa_titulacion_implicacion***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción de egresados en el plan para el ítem *Mi implicación con los estudios*. Escala 1-5

- ***sa_titulacion_aprovechamiento_recursos***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción de egresados en el plan para el ítem *Mi aprovechamiento de los recursos didácticos disponibles*. Escala 1-5

- ***sa_titulacion_mejora_profesional***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción de egresados en el plan para el ítem *La titulación obtenida ha supuesto una mejora profesional.* Escala 1-5

- ***sa_titulacion_obtener_trabajo***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción de egresados en el plan para el ítem *La titulación obtenida ha supuesto obtener un puesto de trabajo.* Escala 1-5

- ***sa_titulacion_titulacion***: Promedio de valoraciones aportadas por el estudiante en las encuestas de satisfacción de egresados en el plan para el ítem *Grado de satisfacción general con la titulación.* Escala 1-5

- ***sa_titulacion_recomendacion***: Proporción de respuestas Sí al ítem *¿Recomendarías estudiar en la Udima a otra persona?* respecto de las encuestas respondidas por el egresado en las encuestas de satisfacción con la titulación. Escala 0-100

- ***sa_titulacion_comentarios***: Promedio de la connotación del contenido de los comentarios realizados por el egresado en las encuestas de satisfacción con la titulación. Escala -1:1 (léxico NRC)

- ***il_R***: Representatividad de la muestra de estudios de inserción laboral respondidos por el egresado

- ***il_participacion***: Proporción de estudios de inserción laboral respondidos por el egresado en los cursos de egreso y estudios en el plan respecto de las que han sido enviadas (estudios tipo I y II, realizados respectivamente, al año y dos años de egresar)

- ***il_motivo_estudios***: Valoración aportada por el egresado en los estudios tipo I y II en el plan para el ítem *¿Por qué realizaste tus estudios en Udima?* (1 si 'Mejora laboral', 0 si 'Otros')

- ***il_consecuencia_estudios***: Valoración aportada por el egresado en los estudios tipo I y II en el plan para el ítem *¿Qué ha supuesto la realización de los estudios?* (1 si 'Conseguir o mejorar un empleo', 0 si 'No variación u Otro')

- ***il_trabajo_actual***: Valoración aportada por el egresado en los estudios tipo I y II en el plan para el ítem *¿Trabajas actualmente?* (1, 'Sí' y 0, 'No')

- ***il_trabajo_actual_relacionado_estudios***: Valoración aportada por el egresado en los estudios tipo I y II en el plan para el ítem *¿Trabajas actualmente en alguna actividad relacionada con los estudios finalizados?* (1, 'Sí' y 0, 'No')

- ***il_comentarios***: Promedio de la connotación del contenido de los comentarios realizados por el estudiante egresado en el estudio de inserción laboral del título. Escala -1:1 (léxico NRC)
- ***dias_laborable***: Proporción de conexiones del usuario a la plataforma LMS *Moodle* en días de diario (lunes a viernes) respecto del total de sus conexiones
- ***dias_festivo***: Proporción de conexiones del usuario a la plataforma LMS *Moodle* en días festivos (sábado y domingo) respecto del total de sus conexiones
- ***horario_mañana***: Proporción de conexiones del usuario a la plataforma LMS *Moodle* en horario de mañana (entre las 7:00 y las 12:59 respecto del total de sus conexiones
- ***horario_tarde***: Proporción de conexiones del usuario a la plataforma LMS *Moodle* en horario de tarde (entre las 13:00 y las 20:59) respecto del total de sus conexiones
- ***horario_noche***: Proporción de conexiones del usuario a la plataforma LMS *Moodle* en horario de tarde (entre las 21:00 y las 6:59) respecto del total de sus conexiones

3.3. Productos obtenidos.

Del desarrollo del presente proyecto se han obtenido los siguientes productos:

- Modelo que clasifica al estudiantado que ha finalizado su experiencia en la Udima, ya sea egresando u abandonando los estudios universitarios, para conocer las características que les definen;
- En base a las características anteriores, definir un modelo capaz de predecir la clasificación del estudiantado que se encuentra en progreso, con especial interés en identificar a los que se encuentran en riesgo de abandono.

3.4. Valoración económica del trabajo.

La valoración económica del presente producto y la planificación del despliegue de los modelos quedan fuera del alcance del presente proyecto.

En esta fase inicial, no procede valorar económicamente el proyecto dado que dependería de la planificación e intereses de la organización. Este proyecto se presentará a la Udima para valorar la viabilidad de su puesta en producción, analizando los gastos asociados al desarrollo y mantenimiento del trabajo, así como los beneficios económicos a obtener.

El mantenimiento del trabajo debería realizarse de manera anual, incorporando los resultados generados en cada curso académico finalizado teniendo en consideración que las actas indefinidas han de estar cerradas en su totalidad.

Capítulo 4

Resultados

4.1. Detalle de resultados obtenidos.

El detalle de todos los resultados obtenidos está disponible en *github*.

4.2. Población investigada.

La población investigada esta formada por el estudiantado de la Udima que ha formalizado matrícula en uno o varios cursos desde '2008-09' (inicio de la universidad) hasta '2021-22' (último curso finalizado) en el momento de realización del presente proyecto. Esta población de 38864 observaciones se divide en dos subconjuntos de datos con la misma estructura y que:

- contiene los datos del estudiantado que ha egresado u abandonado los estudios (para acometer los objetivos específicos 1 y 2), con un total de 30875 observaciones;

- contiene los datos del estudiantado que se encuentra cursando estudios (para acometer el objetivo específico 3), con un total de 7989 observaciones.

4.3. Resultados de los objetivos específicos.

Se adjuntan los resultados obtenidos para cada uno de los objetivos específicos planteados en el proyecto:

4.3.1. Objetivo específico 1: Identificar los perfiles del estudiantado que finaliza o abandona sus estudios.

Tal y como se ha detallado en el apartado de metodología 3.2.1, para el desarrollo de este objetivo se aplica: **algoritmo de agrupación** *k-means*; **métrica** distancia euclidiana y **4 *clusters*** como sugieren el **método *elbow* (el codo) y el método de la silueta promedio**:

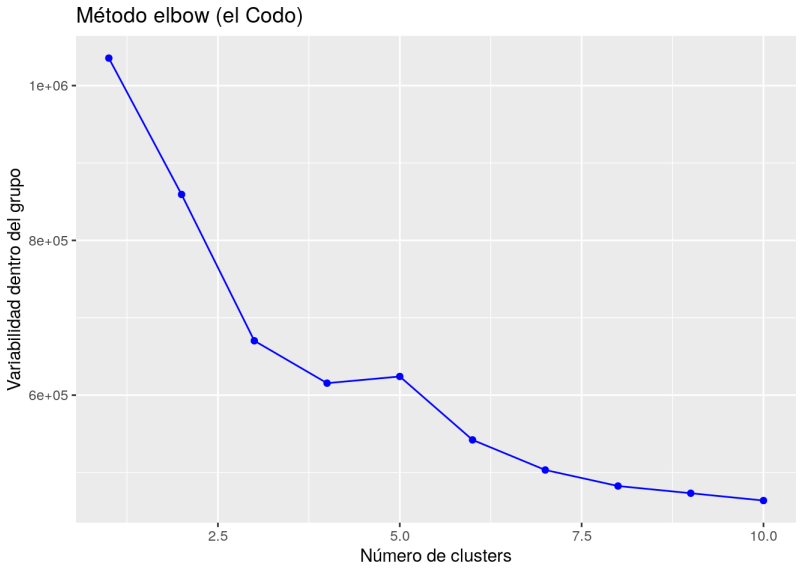

Figura 4.1: Aplicación del método *elbow* (el codo)

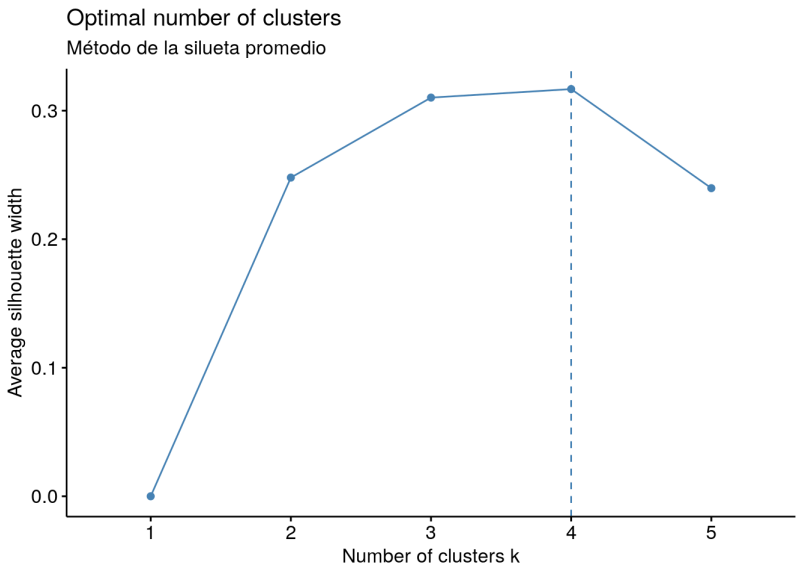

Figura 4.2: Aplicación del método de la silueta promedio

4.3.2. Objetivo específico 2: Determinar las características principales que definen a un estudiante que finaliza o abandona sus estudios.

Los resultados que se obtienen de la clasificación se identifican como:

- **Dropout**: Estudiante que abandona los estudios.
- **Common graduate**: Egresado común.
- **Motivated**: Egresado especialista, con sentido de pertenencia a la universidad (alta implicación en la mejora una vez ha finalizado los estudios), alta satisfacción y con trabajo relacionado con la titulación realizada.
- **Dissatisfied**: Egresado con satisfacción inferior a la media.

Se muestra un tabla comparativa de las distintas categorías:

Nota: En el caso de la proporción, se indica el valor absoluto y entre paréntesis, la proporción respecto al total del estudiantado. En el resto de variables se muestra el valor promedio y entre paréntesis, la desviación típica.

	dropout	common_graduate	motivated	dissatisfied
proporción	7747 (25.09 %)	18399 (59.59 %)	2706 (8.76 %)	2023 (6.55 %)
edad_inicio	32.6 (9.2)	29.9 (7.5)	31.1 (8.4)	31.3 (8)
titulado	0 (0)	1 (0.1)	1 (0.1)	0.9 (0.3)
n_especialidades	0 (0)	1 (0.3)	1.1 (0.3)	1 (0.5)
t_sup_req	24.1 (30.4)	102.6 (7.1)	102.5 (6.6)	94.5 (23)
t_rec	3.9 (10.5)	7.9 (14.3)	9.3 (16.3)	11.2 (19.3)
nota_media	7.6 (0.9)	8 (0.7)	8 (0.7)	8 (0.8)
tasa_eficiencia	41.4 (37.5)	97.5 (6.8)	97.1 (7.2)	94 (15.9)
t_exito	58.5 (43.5)	99.2 (3.3)	99 (3.7)	97.1 (11.8)
t_evaluacion	50.6 (38.2)	98.2 (5.5)	98.1 (5.6)	95.9 (12.6)
sa_asig_participacion	10.5 (25.4)	15.1 (30.1)	37.3 (40.6)	62 (33.4)
sa_asig_docente	4.5 (0.2)	4.5 (0.2)	4.4 (0.5)	3.3 (0.9)
sa_asig_aprendiendo	90.3 (7.1)	90.2 (8.4)	85.3 (18.6)	72.3 (26.5)
sa_curso_participacion	3.7 (17.6)	10.6 (28.3)	34.7 (43.4)	42.1 (45)
sa_curso_titulacion	4 (0.2)	4 (0.3)	4 (0.6)	3.6 (0.9)
sa_titulacion_participacion	0 (0)	9.9 (29.8)	61.7 (48.6)	20.8 (40.5)
sa_titulacion_implicacion	5 (0)	5 (0.1)	4.8 (0.5)	4.9 (0.3)
il_participacion	0 (0)	0.3 (5.1)	100 (1.9)	3.4 (18.2)
il_trabajo_actual_relacionado_estudios	0 (0)	0 (0)	0.6 (0.5)	0 (0.1)

Tabla 4.1: Características de los perfiles del estudiantado.

4.3.3. Objetivo específico 3: Predecir la clasificación del estudiantado que no ha finalizado los estudios.

Según lo descrito en el apartado de metodología 3.2.1, para el desarrollo de este objetivo se aplica *10-Fold-Cross-Validation* para determinar que el **algoritmo que aporta mejores resultados es *Random forest* atendiendo a la máxima *accuracy* o exactitud**:

	Accuracy	Kappa	AccuracySD	KappaSD	Algoritmo	Valor	Métrica
1	0.82	0.67	0.01	0.01	*k-NN*	3.00	k
2	0.83	0.68	0.01	0.01	*k-NN*	5.00	k
3	0.84	0.69	0.00	0.01	*k-NN*	7.00	k
4	0.84	0.69	0.01	0.01	*k-NN*	9.00	k
5	0.84	0.70	0.01	0.01	*k-NN*	11.00	k
6	0.84	0.70	0.01	0.01	*k-NN*	13.00	k
7	0.84	0.70	0.01	0.01	*k-NN*	15.00	k
8	0.86	0.74	0.01	0.01	*Random Forest*	2.00	mtry
9	**0.87**	**0.76**	**0.00**	**0.01**	***Random Forest***	**10.00**	**mtry**
10	0.86	0.74	0.00	0.01	*Random Forest*	19.00	mtry

Tabla 4.2: Comparativa de la validación de los algoritmos de clasificación.

Las 10 variables más importantes del modelo son:

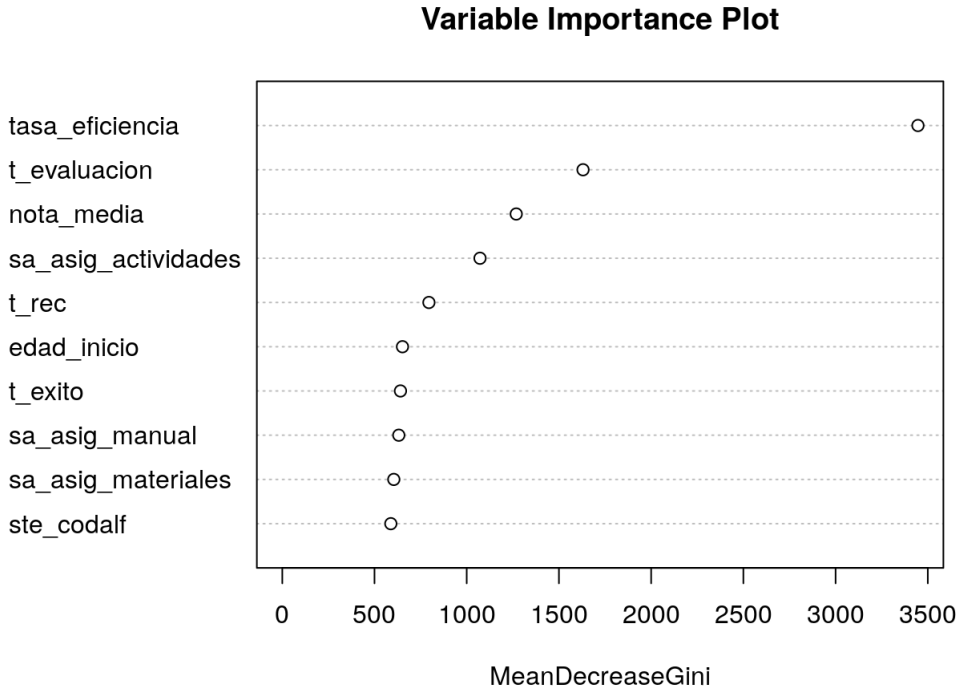

Figura 4.3: Importancia de las variables del modelo *Random Forest*

Se muestra la evolución del error *out-of-bag (OOB)* de los árboles del modelo tras aplicar *Random Forest* con el valor óptimo *mtry=10* al conjunto de entrenamiento:

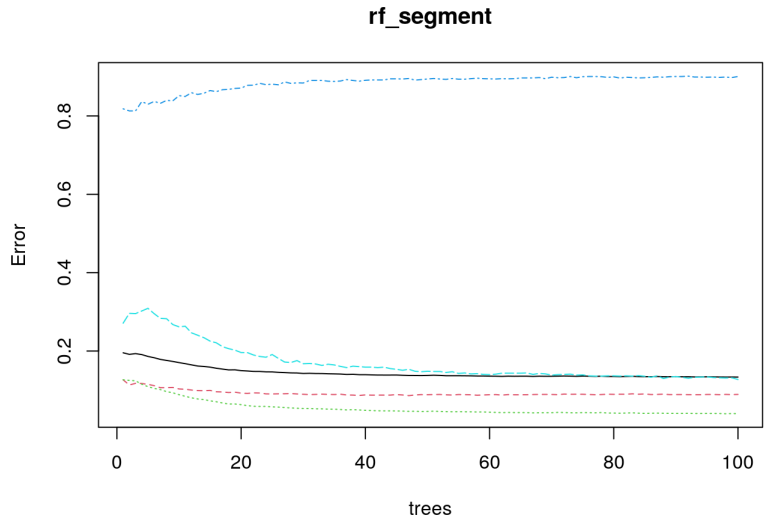

Figura 4.4: Evolución del *out-of-bag (OOB)* error de los árboles del modelo *Random Forest*

Y el detalle de *OOB estimate of error rate* del 13.31 % (media ponderada del modelo según el *class.error* de cada categoría) mostrado en la figura 4.4 a partir de la matriz de confusión por categorías es:

		Valor predicho				
		dropout	common_graduate	motivated	dissatisfied	class.error
Valor real	dropout	5633	524	9	32	0.09115844
	common_graduate	329	14148	181	61	0.03879340
	motivated	56	1598	213	298	0.90161663
	dissatisfied	22	89	88	1419	0.12299135

Tabla 4.3: Matriz de confusión del *OOB estimate of error rate* del modelo *Random Forest*.

La bondad de la predicción del conjunto de test es del 86.45 % y la matriz de confusión es:

		Valor predicho			
		dropout	common_graduate	motivated	dissatisfied
Valor real	dropout	1412	124	2	11
	common_graduate	105	3523	38	14
	motivated	12	407	49	73
	dissatisfied	6	30	15	354

Tabla 4.4: Matriz de confusión del conjunto de test.

4.3.4. Comparación y justificación de resultados: clasificación vs. predicción.

Se comparan los resultados obtenidos en la clasificación del estudiantado que ha finalizado sus estudios (egreso/abandono) y de la predicción del estudiantado en progreso.

Nota: En el caso de la proporción, se indica el valor absoluto y entre paréntesis, la proporción respecto al total del estudiantado. En el resto de variables se muestra el valor promedio y entre paréntesis, la desviación típica.

	dropout	common_graduate	motivated	dissatisfied
método	clasificación	clasificación	clasificación	clasificación
proporción	7747 (25.09 %)	18399 (59.59 %)	2706 (8.76 %)	2023 (6.55 %)
edad_inicio	32.6 (9.2)	29.9 (7.5)	31.1 (8.4)	31.3 (8)
t_rec	3.9 (10.5)	7.9 (14.3)	9.3 (16.3)	11.2 (19.3)
nota_media	7.6 (0.9)	8 (0.7)	8 (0.7)	8 (0.8)
tasa_eficiencia	41.4 (37.5)	97.5 (6.8)	97.1 (7.2)	94 (15.9)
t_exito	58.5 (43.5)	99.2 (3.3)	99 (3.7)	97.1 (11.8)
t_evaluacion	50.6 (38.2)	98.2 (5.5)	98.1 (5.6)	95.9 (12.6)
sa_asig_participacion	10.5 (25.4)	15.1 (30.1)	37.3 (40.6)	62 (33.4)
sa_asig_docente	4.5 (0.2)	4.5 (0.2)	4.4 (0.5)	3.3 (0.9)
sa_asig_aprendiendo	90.3 (7.1)	90.2 (8.4)	85.3 (18.6)	72.3 (26.5)
sa_curso_participacion	3.7 (17.6)	10.6 (28.3)	34.7 (43.4)	42.1 (45)
sa_curso_titulacion	4 (0.2)	4 (0.3)	4 (0.6)	3.6 (0.9)

Tabla 4.5: Comparativa de resultados: **clasificación**

	dropout	common_graduate	motivated	dissatisfied
método	predicción	predicción	predicción	predicción
proporción	4761 (59.38 %)	2464 (30.85 %)	89 (1.11 %)	675 (8.69 %)
edad_inicio	31.9 (9.3)	31.7 (8.8)	36.1 (11.3)	33.4 (9)
t_rec	4.4 (11.4)	12.3 (17.3)	12.3 (16.1)	10.5 (15.5)
nota_media	7.4 (0.8)	7.7 (0.8)	7.8 (0.8)	7.7 (0.8)
tasa_eficiencia	55.6 (32.9)	90.6 (10.8)	93.1 (8.7)	84.2 (19.2)
t_exito	73.3 (35.1)	96 (6.9)	96.8 (5.9)	92.7 (14.1)
t_evaluacion	66.6 (32.2)	94.3 (8.7)	96.2 (6.9)	89.8 (15.5)
sa_asig_participacion	14.4 (28.7)	26.2 (35.4)	57.8 (35.8)	57.1 (31.7)
sa_asig_docente	4.5 (0.3)	4.5 (0.3)	4.4 (0.4)	3.3 (0.8)
sa_asig_aprendiendo	89.7 (10.8)	89.9 (12.2)	75.1 (27)	68.3 (25)
sa_curso_participacion	14 (31)	33.2 (41.8)	73.7 (39.7)	48.9 (42.7)
sa_curso_titulacion	4 (0.5)	4.1 (0.6)	4 (0.8)	3.5 (0.9)

Tabla 4.6: Comparativa de resultados: **predicción**

Capítulo 5

Conclusiones

5.1. Conclusiones obtenidas.

Una vez obtenidos y analizados los resultados de la predicción de la clasificación (4.3.4), se concluye que las predicciones del estudiantado en progreso frente al estudiantado que ha egresado/abandonado experimenta:

- la máxima variación en los perfiles ***dropout***, incrementándose (del 25.09 % al 59.38 %) y ***common_graduate***, disminuyendo (del 59.59 % al 30.85 %);
- mientras que el perfil ***dissatisfied*** se asemeja más a la clasificación (del 6.55 % al 8.69 %);
- y se identifican, en menor medida, ***motivated*** (del 8.76 % al 1.11 %).

Estas variaciones resultan coherentes, pues la metodología a distancia empleada en la Udima experimenta un mayor abandono que las cifras más recientes que se han publicado por el INE (13).

La principal diferencia que aporta este proyecto frente a los ya utilizados en la Udima para predecir el abandono es la integración de información procedente de los estudios de satisfacción e inserción laboral realizados al estudiantado. Este nuevo modelo aporta información subjetiva sobre el estudiantado, siendo información que puede relacionarse con:

- los modelos de adaptación y psicopedagógico ya descritos por Álvarez Ferrándiz (2021) en (12);
- y el enfoque de segmentación basado en la integración de factores objetivos (racionales) y subjetivos (emocionales) sugerido por Story (2021) en (20)

ambos referenciados en 2.1.1.2.

Actualmente, el Departamento de Atención y Orientación al Estudiante (DAOE) de la Udima dispone de datos facilitados por el Servicio de Prevención del Abandono (SPA) para personalizar el seguimiento del estudiantado que se encuentra en riesgo de abandono. La información obtenida por el modelo propuesto en el presente proyecto podría complementar la que ya tiene disponible el DAOE para personalizar aún más el seguimiento.

5.2. Reflexión crítica sobre la consecución de los objetivos.

Según 4.3, se considera que los objetivos específicos planteados en el proyecto se han conseguido.

Antes de valorar la implementación real del proyecto, se debe evaluar junto a un equipo competente la adecuación de los resultados obtenidos en la predicción, principalmente del % del estudiantado en riesgo de abandono. Además, se debe atender al trabajo futuro planteado y a tratar de resolver las limitaciones planteadas en el apartado correspondiente 5.5.

De manera análoga a Rovira, S. et al. (2017) en (23), y según sugiere el estudio realizado por Gairín, J. et al. (2014) en (24), el presente proyecto se centra en analizar el abandono de primer año, es decir, según las observaciones realizadas en 2.1.2 sobre la metodología de cálculo del abandono adaptado al caso de la Udima (no limitándolo a la población óptima por la definición del perfil del estudiantado).

Como ha quedado justificado en 5.1, es coherente que el abandono de la enseñanza a distancia sea superior al de la enseñanza presencial, por ello, los valores obtenidos en el presente estudio que preciden un riesgo de abandono del 59.38 % son coherentes si se comparan con los resultados obtenidos por Rovira, S. et al. (2017) en su estudio (23). Se debe tener en cuenta que sus resultados (23) están referidos a la Universidad de Barcelona, con metodología de enseñanza presencial, y precisados a nivel de titulación, reportando una predicción de abandono en el primer año del 11.6 % en estudios de Derecho, 22.9 % en Informática y 33.1 % en Matemáticas.

En un estudio similar de Coussement, K. et al. (2020) en (25) se predice un 55 % de abandono en entornos de aprendizaje en línea aplicando *Logit Leaf Model*. Este resultado es muy similar al porcentaje reportado en el presente proyecto (59.38 %) con la aplicación de *Random Forest*.

5.3. Seguimiento de la planificación y metodología.

La planificación propuesta para el proyecto (1.5) sufrió un retraso de dos semanas en la fecha fin de las fases II y III (1.4) debido a dificultades en la obtención de datos de partida para la configuración del juego de datos definitivo. Este retraso no afectó a la fecha fin de las fases IV a VI de la planificación (1.5).

La metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) prevista (1.4) ha sido suficientemente adecuada para dessarrollar el proyecto y conseguir los objetivos planteados (4.3). Como se ha detallado en 3.2.1, tras la evaluación de los primeros resultados (predicción del abandono del 1.56 % y accuracy del 98.54 %) ha sido necesario realizar ajustes en los datos (eliminando las variables relacionadas con el egreso/abandono) y repetir parte de las fases III a V, obteniendo así resultados más coherentes (el 59.38 % del estudiantado en progreso se encuentra en riesgo de abandono, con *accuracy* del test del 86.45 %).

Las decisiones justificadas en el apartado 3.2.2 han garantizado el éxito del trabajo, no siendo necesario introducir cambios sobre el planteamiento inicial del mismo.

5.4. Impactos previstos en sostenibilidad, ético-social y de diversidad.

Se evalúan los impactos ético-sociales, de sostenibilidad y de diversidad previstos en 1.3 de manera positiva en el supuesto de que el producto/servicio se implementase:

- **Sostenibilidad**:
  - ODS 12 - *Responsible consumption and production*: mejorando el consumo responsable de los recursos público-privados (Udima e instituciones públicas).
- **Comportamiento ético y responsabilidad social (RS)**:
  - ODS 8 - *Decent work and economic growth*: aumentando el crecimiento económico de la Udima en relación a la retención del estudiantado y reducción del abandono.
  - ODS 16 - *Peace, justice and strong institutions*: aumentando la confianza de los grupos de interés de la Udima, en especial, del estudiantado, consolidando su reputación y prestigio en el panorama universitario español.
- **Diversidad (género entre otros) y derechos humanos**:
  - ODS 5 - *Gender equality*: el servicio/producto aplicaría a todo el estudiantado, con independencia de su género.

- ODS 10 - *Reduced inequalities*: el servicio/producto aplicaría a todo el estudiantado, con independencia de su raza, etnia, origen, orientación sexual, ideología, religión, diversidad funcional o posición social.

No se han detectado impactos no previstos previamente en 1.3.

5.5. Trabajo futuro y limitaciones del proyecto.

Se detallan a continuación las líneas de trabajo detectadas que serían susceptibles de mejora de resolver en determinadas fases del proyecto:

5.5.1. Fase de ETL.

En los juegos de datos utilizados para construir el juego de datos del proyecto se han encontrado limitaciones que han dado lugar a una definición no homogénea del perfil del estudiantado. Son las siguientes:

- Las variables de utilización de la plataforma de formación *Moodle* (*dias_laborable, dias_festivo, horario_madrugada, horario_mañana, horario_tarde, horario_noche*) están limitadas a los cursos '2012-13' a '2020-21' (no al período del proyecto comprendido entre 2008-09 y 2021-22);

- Las variables que provienen de los estudios de satisfacción o de inserción laboral no son constantes en el tiempo:

  - El histórico de resultados de satisfacción del estudiantado con las asignatura integra datos desde el curso '2012-13' (no desde 2008-09), además de la variación experimentada por el formato;
  - El histórico de resultados de satisfacción del estudiantado con el curso integra datos desde el curso '2017-18', que es cuando se inició el estudio (no desde 2008-09);
  - El histórico de resultados de satisfacción de egresados con la titulación integra datos desde el curso '2009-10', que es cuando se inició el estudio. Se considera una limitación menor, ya que el primer curso en el que se impartió docencia fue '2008-09';
  - El histórico de resultados de inserción laboral de egresados integra datos desde el curso '2017-18', que es cuando se inició el estudio. Se considera una limitación menor, ya que el primer curso en el que se impartió docencia fue '2008-09'.

- Mantener la variable *plan_estudios* excluida y el valor original de las variables transformadas a dicotómicas (*acceso, il_motivo_estudios, il_consecuencia_estudios*) puede que aporte información sustancial de comportamientos del estudiantado que pueden ser evaluados en futuros proyectos.

Las técnicas de procesamiento del lenguaje natural NLP probadas (tanto con el diccionario de léxico NRC como AFFIN, como se indica en 3.2.2, **para extraer la connotación negativa o positiva de los comentarios expresados por los usuarios en las distintas encuestas de satisfacción o estudios de inserción laboral no siempre aportan un valor adecuado** asociado al rango (donde valores negativos se corresponden con connotaciones negativas y valores positivos con connotaciones positivas). Aunque se ha utilizado el léxico de sentimientos NRC, ambos léxicos calculan el valor numérico asociado a los unigramas (*tokens*) del comentario, evaluándolos de manera aislada por lo que algunos términos pueden ser calificados como positivos cuando en conjunto tienen una carga negativa (p.e. la técnica evaluaría 'bien desilusionado' como neutro al analizar por separado la connotación de las palabras: bien, positiva y desilusionado, negativa cuando bien está enfatizando la negatividad), o por ejemplo, las ironías o metáforas que pueden realizarse en este tipo de comentarios no son correctamente evaluadas, obteniendo valoraciones que no se corresponden con el contenido o connotación real del comentario. **Se considera interesante explorar la obtención de la connotación de los comentarios a partir de bigramas o trigramas** y contrastarlo con los resultados actuales a partir de unigramas.

Se considera que la baja participación en las encuestas es una limitación del juego de datos del proyecto, lo que ha conllevado a la imputación de valores ausentes según lo descrito en 3.2.2. Pese a que se ha decidido imputar la mediana (métrica robusta a los *outliers*) en las variables de las encuestas, principalmente de asignatura y curso, se identifica como una penalización a la hora de clasificar al estudiantado, pues la no respuesta a las encuestas puede ser un factor identificador del perfil de los mismos. Actualmente, se está imputando datos de la misma manera en una variable que cuenta con valores ausentes por dos motivos distintos:

- los históricos de resultados de satisfacción no cuenten con datos desde el curso '2008-09', donde parece lógico imputar la mediana;
- los históricos de resultados de cursos sí recogidos en el mismo cuentan con valores ausentes porque el estudiantado ha decidido no manifestar su opinión al respecto o porque en el curso de referencia, el formato aplicado no contemplaba cierto ítem.

En fases futuras de revisión del proyecto se explorará imputar 0 (valor no contenido en la escala *Likert* utilizada 1-5) en los ítems de las encuestas con valores ausentes donde sí ha tenido

opción el estudiante de participar y la mediana, en datos no incluidos en la construcción del histórico. De este modo, la técnica de imputación se corresponderá con el motivo de la ausencia: diferenciando el tratamiento de los valores ausentes intencionados (el estudiante decide no responder) del no intencionado (derivado de la construcción del histórico desde '2012-13').

5.5.2. Objetivos específicos.

Si se resuelven las limitaciones identificadas en el juego de datos según 5.5.1 se deberá proceder de nuevo con la resolución de los objetivos específicos, identificando los perfiles del estudiantado, sus características y prediciendo de nuevo la clasificación del estudiantado que no ha finalizado los estudios.

5.5.3. Evaluación de los resultados obtenidos y futuras consideraciones.

Como ya concluyeron Norambuena et al. (2022) (4) en su estudio sobre los modelos predictivos basados en uso de analíticas de aprendizaje en educación superior, en el futuro, los modelos predictivos utilizados para el rendimiento académico del estudiantado deberían tener en cuenta el último método de valoración de las evaluaciones basado en un sistema educativo moderno que haga hincapié en las habilidades blandas, las habilidades interpersonales y las capacidades de pensamiento de alto nivel.

Otras posibles vías de exploración serían:

- Si se resuelven las limitaciones identificadas en las técnicas del procesamiento del lenguaje natural NLP identificadas en 5.5.1, **se podría analizar la connotación de las intervenciones del estudiantado en cualquier foro del campus virtual *Moodle*, no solo en los comentarios de las encuestas o estudios**;
- **Explorar el concepto de abandono por titulación**, como ya defienden Josep Grau-Valldosera & Minguillón, 2014 en (15), referenciado en 2.1.2 como '(...) cuando el estudiante toma N *breaks* o más, se asume que ha abandonado el programa, siendo N diferente para cada programa de estudios'.
- **Si no es posible explorar el concepto de abandono por titulación**, como se identifica en el párrafo anterior, **al menos aplicarlo a nivel de rama de conocimiento y nivel de estudios**, diferenciando así por Grado y Máster de cada rama de conocimiento.

Capítulo 6

Glosario

A continuación se muestra la relación y definición de los términos y acrónimos más relevantes utilizados en la Memoria:

- **AA**: Analíticas de Aprendizaje. Las Analíticas de Aprendizaje (AA) son una herramienta tecnológica considerada como una práctica de minería de conjunto de datos de las instituciones educativas universitarias para obtener inteligencia procesable a partir de técnicas de modelado estadístico y predictivo con el propósito de mejorar la toma de decisiones, el resultado y el éxito del estudiantado.

- ***Accuracy* (o exactitud)**: métrica de evaluación de los modelos predictivos que determina el porcentaje de acierto de la predicción. Su fórmula de cálculo es:

$$Accuracy = \frac{VN + VP}{VP + FP + VN + FN} * 100$$

  Siendo VP (verdaderos positivos), FP(falsos positivos), VN(verdaderos negativos) y FN (verdaderos negativos). Una matriz de confusión muestra los recuentos de VP, FP, VN y FN.

- **CCEG**: Competencia de compromiso ético y global.

- **CRISP-DM (*Cross Industry Standard Process for Data Mining*)**: Método probado para orientar trabajos de minería de datos que incluye descripciones de las fases del proyecto, las tareas de cada fase y las relaciones entre las tareas.

- **ECTS (*European Credit Transfer and Accumulation System*)**: Sistema europeo de transferencia y acumulación de créditos.

- **EEES**: Espacio Europeo de Educación Superior.

- **LMS (*Learning Management System*)**: Un LMS o sistema de gestión de aprendizaje es un paquete de *software* basado en la web que está diseñado para planificar, implementar y evaluar el aprendizaje, facilitando la interacción del estudiantado, dar retroalimentación sobre el rendimiento y gestionar actividades. Los LMS más comunes son: de código abierto (Moodle, Sakai y ATutor) o comerciales (Blackboard, SuccessessFactors y Sum Total).
- **ODS**: Objetivos de Desarrollo Sostenible.
- **ONU**: Organización de las Naciones Unidas.
- **OOB**: el *Out-of-Bag* (OOB) error de un modelo *Random Forest*, permite obtener una estimación del error de test sin recurrir a validación cruzada. Esta característica, combinada con una estrategia de *early stopping*, se emplea para encontrar eficazmente los hiperparámetros óptimos. El *OOB-error* es un estimador del error de test. Si el número de árboles es alto, el *OOB-error* es equivalente o similar al *leave-one-out cross-validation* error.
- **RS**: Responsabilidad social.
- **SIGC**: Sistema Interno de Garantía de Calidad.
- **SIIU**: El Sistema Integrado de Información Universitaria (SIIU) es una plataforma de recogida, procesamiento, análisis y difusión de datos del Sistema Universitario Español (SUE) que inició su actividad en el curso 2010-11. Constituye una herramienta primordial para la obtención de las estadísticas universitarias oficiales recogidas en el Plan Estadístico Nacional, permitiendo disponer de información homogénea y comparable.
- **SUE**: Sistema Universitario Español. Conjunto de enseñanzas e instituciones universitarias de España.
- **TIC**: Tecnologías de la Información y la Comunicación (TIC). Conjunto de recursos y herramientas utilizados para el proceso, administración y distribución de la información a través de la tecnología. Posibilitan el acceso a la información a través de la digitalización de la información y de manera inmediata, permitiendo la comunicación bidireccional.
- **TF o TFM**: Trabajo Fin de Máster.
- **Udima**: Universidad a Distancia de Madrid.
- **UNESCO (*United Nations Educational, Scientific and Cultural Organization*)**: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

con el objetivo de contribuir a la paz y a la seguridad en el mundo mediante la educación, la ciencia y la cultura.

- **UTC**: Unidad Técnica de Calidad.

Bibliografía

[1] N. Soni and A. Ganatra, "Categorization of several clustering algorithms from different perspective: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 8, pp. 63–68, 2012. [Online]. Available: https://bit.ly/3ab5IbH

[2] Wikipedia, "Proceso de bolonia." [Online]. Available: https://es.wikipedia.org/wiki/Proceso_de_Bolonia

[3] miformacion, "Estas son las 8 mejores universidades online en 2022." [Online]. Available: https://miformacion.eu/mejores-universidades-espana/online/

[4] J. M. Norambuena and Y. L. Badilla-Quintana, María Graciela and Angulo, "Modelos predictivos basados en uso de analíticas de aprendizaje en educación superior: una revisión sistemática," *Texto Livre*, vol. 15, 2022. [Online]. Available: https://doi.org/10.35699/1983-3652.2022.36310

[5] UDIMA, "Misión, visión y valores." [Online]. Available: https://www.udima.es/es/vision-mision-udima.html

[6] UNESCO, "La unesco y la educación: toda persona tiene derecho a la educación," Tech. Rep., 2011. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000212715_spa

[7] UOC, "Guía transversal sobre la competencia Ética y global," 2022. [Online]. Available: https://drive.google.com/file/d/1sJWP6UzG8a5gPbV0-NLSNAj0JE5NjVBJ/view?usp=sharing

[8] ONU, "Objetivos de desarrollo sostenible 2030." [Online]. Available: https://www.un.org/sustainabledevelopment/

[9] UOC, "Impacto global agenda2030." [Online]. Available: https://www.uoc.edu/portal/es/compromis-social/index.html

[10] Subdirección General de Actividad Universitaria Investigadora de la Secretaría General de Universidades, *Catálogo oficial de indicadores universitarios del Sistema Integrado de Información Universitaria (SIIU)*, Ministerio de Universidades, Enero 2022.

[11] Subdirección General de Actividad Universitaria Investigadora, in *Datos y cifras del Sistema Universitario Español*, Secretaría General Técnica del Ministerio de Universidades ed., P. E. del Ministerio de Universidades, Ed. Publicación 2021-2022, 2022. [Online]. Available: https://www.universidades.gob.es/stfls/universidades/Estadisticas/ficheros/DyC_2021_22.pdf

[12] D. Álvarez Ferrandiz, "Análisis del abandono universitario en España: un estudio bibliométrico," *PUBLICACIONES*, vol. 51, no. 2, pp. 241–261, dic. 2021. [Online]. Available: https://revistaseug.ugr.es/index.php/publicaciones/article/view/23843

[13] INE, "Indicadores de rendimiento académico universitario (datos publicados en 2022)," Tech. Rep., 2022. [Online]. Available: https://www.ine.es/dyngs/IOE/es/operacion.htm?id=1259946001133

[14] M. Fernández-Mellizo, in *Análisis del abandono de los estudiantes de grado en las universidades presenciales en España*, Secretaría General Técnica del Ministerio de Universidades ed., P. E. del Ministerio de Universidades, Ed. Marzo 2022, 2022. [Online]. Available: https://www.universidades.gob.es/stfls/universidades/Estadisticas/ficheros/DyC_2021_22.pdf

[15] B. G. Troche de Trevisan, "Estudio del rendimiento académico del estudiante en línea como variable predictiva del abandono en educación superior: el caso de la Universitat Oberta de Catalunya," Ph.D. dissertation, Universitat Oberta de Catalunya (UOC), 2019. [Online]. Available: https://openaccess.uoc.edu/handle/10609/100346

[16] JSTOR. [Online]. Available: https://www.jstor.org/

[17] J. Kleinberg, "An impossibility theorem for clustering," vol. 15, 2002. [Online]. Available: https://proceedings.neurips.cc/paper/2002/file/43e4e6a6f341e00671e123714de019a8-Paper.pdf

[18] J. A. Muffo, "Market segmentation in higher education: A case study," *Journal of Student Financial Aid*, vol. 17(3), pp. , 31–40, 1987. [Online]. Available: https://ir.library.louisville.edu/jsfa/vol17/iss3/3/

[19] F. Angulo, A. Pergelova, and J. Rialp, "A market segmentation approach for higher education based on rational and emotional factors," *Journal of Marketing*

for Higher Education, vol. 20, no. 1, pp. 1–17, 2010. [Online]. Available: https://doi.org/10.1080/08841241003788029

[20] J. Story, "Unique challenges of segmentation and differentiation for higher education," 2021. [Online]. Available: https://doi.org/10.1080/08841241.2021.1874589

[21] W. R. Wynd and C. S. Bozman, "Student learning style: A segmentation strategy for higher education," *Journal of Education for Business*, vol. 71, no. 4, pp. 232–235, 2010. [Online]. Available: https://doi.org/10.1080/08832323.1996.10116790

[22] L. Chavez and J. Salinas, "Segmentación de los alumnos ingresantes a una universidad pública aplicando el algoritmo k-prototype," *Tierra nuestra*, vol. 15, no. 2, pp. 10–21, 2021. [Online]. Available: https://doi.org/10.21704/rtn.v15i2.1825

[23] S. Rovira, E. Puertas, and L. Igual, "Data-driven system to predict academic grades and dropout," February 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0171207

[24] J. Gairín, X. Triado, M. Feixas, P. Figuera, P. Aparicio-Chueca, and M. Torrado, "Student dropout rates in catalan universities: profile and motives for disengagement." 2014. [Online]. Available: https://doi.org/10.1080/13538322.2014.925230

[25] K. Coussement, M. Phan, A. De Caigny, D. F. Benoit, and A. Raes, "Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model," *Decision Support Systems*, vol. 135, p. 113325, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923620300804