UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Estudio sobre la búsqueda de mentor utilizando aprendizaje automático

Memoria del trabajo

Autor:

Juan García Ruiz

Director:

Antonio Rodríguez Anaya

Máster Universitario en Ingeniería Informática

Septiembre de 2023

Trabajo de Fin de Máster presentado en la Escuela Técnica Superior de Ingeniería Informática de la Universidad Nacional de Educación a Distancia para la obtención del Máster Universitario en Ingeniería Informática

Resumen

La mentoría es una relación de desarrollo personal en la cual una persona más experimentada o con mayor conocimiento ayuda a otra menos experimentada o con menor conocimiento. La persona que recibe la mentoría ha sido llamada tradicionalmente como protegido, discípulo o aprendiz.

El proceso de búsqueda de candidatos a ser tutores o mentores se entiende como la combinación deliberada de personas para formar un todo previsto. Uno de los desafíos más críticos es decidir quién sería un buen mentor de otro alumno. Es necesario una supervisión que debe completar un proceso de varios pasos, incluida la búsqueda, identificación y elección de candidatos.

El propósito de este Trabajo de Fin de Máster es el de, como su propio nombre indica, generar una propuesta de solución que recomiende mentores teniendo en cuenta factores como la información sobre las interacciones sociales de los individuos con la comunidad como un todo y con cada individuo por separado, o los criterios de agrupamiento debido a expertos. Con este fin, se utilizará análisis de redes sociales, análisis de sentimientos y técnicas de aprendizaje automático.

Como resultados de este trabajo, se proporciona una propuesta de algoritmo de generación de duplas en la que, haciendo uso de clustering, modularidad y análisis de sentimientos, se obtienen los miembros de esta, de manera que sean lo más compatible posibles. Además de esto, se presentará una propuesta de proyecto para el desarrollo de una aplicación funcional basado en este algoritmo. Finalmente, se añaden conclusiones sobre otros posibles acercamientos o consideraciones que podrían tenerse en cuenta para la mejora del modelo (como el uso de datos estáticos o estudiar más en detalle la utilidad del análisis de sentimientos).

Así, se describirá en los dos primeros capítulos (‘Introducción’ y ‘Estado de la cuestión’) las circunstancias actuales que han motivado la realización de este trabajo, así como en las diferentes hipótesis que se ha apoyado. Para comprobar la exactitud de estas cuestiones, se realizará una serie de análisis que recorrerá todo el proceso de minería de datos (extracción de datos, preprocesamiento, procesamiento y validación de los resultados) que se recogerán en el capítulo ‘Estudio’ y que se usará de base para la propuesta de solución que se presentará en la ‘Propuesta de proyecto’. Finalmente, en ‘Conclusiones y trabajos futuros’ se comentarán los resultados obtenidos en la realización de este trabajo.

Abstract

Mentoring is a personal development relationship in which a more experienced or knowledgeable person helps a less knowledgeable one. The person being mentored has traditionally been referred to as a protégé, mentee or apprentice.

The process of finding candidates to be tutors or mentors is undestood as the deliberate combination of people to an intended whole. One of the most critical challenges is deciding who would be a good mentor for another student. That requires oversight that must complete a multi-step process, including searching for, identifying and choosing candidates.

The purpose of this Master's Thesis is, as its name suggests, to generate a proposed solution that recommends mentors taking into account factors such as information about the social interactions of individuals with the community as a whole and with each individual separately, or grouping criteria due to experts. To this end, social network analysis, sentiment analysis and machine learning techniques will be used.

As a result of this work, a proposal for a pairing algorithm is provided, which utilizes clustering, modularity, and sentiment analysis to select members of pairs to be as compatible as possible. Additionally, a project proposal for the development of a functional application based on this algorithm will be presented. Finally, conclusions will be added regarding other possible approaches or considerations that could be taken into account to improve the model, such as the use of static data or a more detailed study of the utility of sentiment analysis.

Thus, the first two chapters ('Introduction' and 'State of the question') will describe the current circumstances that have motivated the realization of this work, as well as the different hypotheses that have been supported. In order to verify the accuracy of these questions, a series of analyses will be carried out that will go through the entire data mining process (data extraction, preprocessing, processing and validation of the results) which will be collected in the 'Study' chapter and will be used as a basis for the proposed solution to be presented in the 'Project proposal'. Finally, in 'Conclusions and future work' the results obtained in the realization of this work will be commented.

Conceptos clave

En este apartado se definirán conceptos clave necesarios para la correcta comprensión del contenido de este documento:

- **Algoritmo:** Conjunto de instrucciones o reglas definidas y no-ambiguas, ordenadas y finitas que permite, por lo general, solucionar un problema, realizar un cómputo, procesar datos...
- **Algoritmo de clustering:** Es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio (por lo general distancia o similitud).
- **Aprendizaje colaborativo:** Es un enfoque educativo en el que los estudiantes trabajan juntos en grupos para lograr un objetivo común, compartiendo conocimientos, experiencias y responsabilidades.
- **Aprendizaje automático:** Es una rama de la inteligencia artificial que se enfoca en desarrollar algoritmos y modelos que permiten a las computadoras aprender y mejorar su rendimiento en tareas específicas a partir de datos.
- **Análisis de sentimientos:** Es una técnica de procesamiento de lenguaje natural que se utiliza para determinar la actitud o emoción expresada en un texto, generalmente relacionada con opiniones de personas sobre un tema.
- **Aristas:** En el contexto de la teoría de grafos, las aristas son las líneas que conectan los nodos o vértices. Representan las relaciones o conexiones entre los elementos del grafo.
- **Blog:** Es un sitio web o plataforma en línea donde una persona o grupo de personas publican regularmente contenido, como artículos, opiniones, noticias o diarios personales, organizados cronológicamente.
- **Dataset:** Es un conjunto de datos organizados y estructurados que se utiliza en análisis, investigación o aprendizaje automático. Puede contener información de diferentes tipos, como texto, números o imágenes.
- **E-learning:** Es una modalidad de educación que se realiza en línea a través de internet. Los estudiantes pueden acceder a cursos, materiales de estudio y actividades desde cualquier lugar con conexión a la red.

- **Foro:** Es una plataforma en línea donde las personas pueden discutir temas específicos, hacer preguntas, compartir información y participar en conversaciones con otros usuarios.
- **Grafo:** En matemáticas y ciencias de la computación, un grafo es una estructura que representa relaciones entre objetos mediante nodos (vértices) y aristas (conexiones entre nodos).
- **Nodo:** En un grafo, un nodo es un punto o vértice que representa un elemento individual. En redes, como las redes sociales, un nodo puede representar a una persona o entidad.
- **Pagerank:** Es un algoritmo desarrollado por Google que se utiliza para calcular la relevancia de las páginas web en los resultados de búsqueda. Se basa en la estructura de enlaces de la web.
- **Plan de trabajo:** Es un documento que describe las tareas, objetivos y plazos para llevar a cabo un proyecto o actividad de manera organizada.
- **Plan de explotación:** Es una estrategia que describe cómo se aprovecharán los resultados o activos producidos para obtener beneficios o lograr objetivos específicos.
- **Preprocesamiento:** Es una etapa importante en el análisis de datos y el aprendizaje automático donde se realizan tareas como limpieza, transformación y selección de datos para prepararlos adecuadamente antes de su análisis.
- **Procesamiento:** Se refiere a las acciones realizadas para analizar, transformar o manipular los datos de acuerdo con los objetivos específicos de un proyecto o tarea.
- **Redes sociales:** Son plataformas en línea que permiten a las personas conectarse, interactuar y compartir contenido con amigos, familiares y otros usuarios. Ejemplos incluyen Facebook, Twitter e Instagram.
- **Wiki:** Es un tipo de sitio web colaborativo que permite a los usuarios crear, editar y colaborar en la creación y modificación de contenido de manera colectiva. Wikipedia es un ejemplo conocido de una wiki.
- **Wireframe:** Es una representación visual simple y esquemática de una página web, aplicación móvil u otra interfaz de usuario.

Índice de figuras

Figura 1 : Tasa de abandono global por tipo de universidad y año de abandono (2012-2013)2

Figura 2 : Ejemplo de invariancia por escala (Sancho, 2020)..... 17

Figura 3 : Ejemplo de consistencia (Sancho, 2020)..... 17

Figura 4 : Ejemplo de riqueza (Sancho, 2020)..... 17

Figura 5 : Diagrama de flujo del proceso de clasificación de alumnos.....23

Figura 6 : Diagrama de flujo del proceso de generación de duplas mentor-estudiante.....24

Figura 7 : Grafo inicial *dataset* Alfa.....30

Figura 8 : Grafo inicial *dataset* Beta..... 31

Figura 9 : Valores medidas (*dataset* Alfa)..... 33

Figura 10 : Valores normalizados (*dataset* Alfa)..... 34

Figura 11 : Resultados cluster en *dataset* Alfa con algoritmo EM (líder).....37

Figura 12 : Resultados cluster en *dataset* Beta con algoritmo EM (líder)..... 38

Figura 13 : Resultados cluster en *dataset* Alfa con algoritmo EM (periférico)..... 39

Figura 14 : Resultados cluster en *dataset* Beta con algoritmo EM (periférico)..... 40

Figura 15 : Resultados cluster en *dataset* Alfa con algoritmo *k-means* (líder)..... 44

Figura 16 : Resultados cluster en *dataset* Beta con algoritmo *k-means* (líder)..... 44

Figura 17 : Resultados cluster en *dataset* Alfa con algoritmo *k-means* (periférico)..... 45

Figura 18 : Resultados cluster en *dataset* Beta con algoritmo *k-means* (periférico).....46

Figura 19 : Modularidad en escenario Alfa.....51

Figura 20 : Modularidad en escenario Beta.....53

Figura 21 : Sentimientos en la conversación entre Estu-49 y Estu-20 del ejemplo Alfa..... 54

Figura 22 : Sentimientos en las conversaciones de Estu-4 con Estu-13 y Estu-16 del ejemplo Beta.....55

Figura 23 : Sentimientos en las conversaciones de Estu-34 con Estu-10 del ejemplo Beta.55

Figura 24 : Sentimientos en las conversaciones de Estu-1 con Estu-36 y Estu-37 del ejemplo Beta.....56

Figura 25 : Sentimientos en las conversaciones de Estu-5 con Estu-26 del ejemplo Beta...56

Figura 26 : Sentimientos en las conversaciones de Estu-5 con Estu-27 del ejemplo Beta...56

Figura 27 : Sentimientos en las conversaciones de Estu-6 con otros usuarios con perfil "*periférico*" del ejemplo Beta.....57

Figura 28 : Sentimientos en las conversaciones de Estu-7 con Estu-25 del ejemplo Beta...57

Figura 29 : Sentimientos en las conversaciones de Estu-11 con otros usuarios con perfil "*periférico*" del ejemplo Beta.....58

Figura 30 : Sentimientos en las conversaciones de Estu-9 con Estu-31 del ejemplo Beta...58

Figura 31 : Sentimientos en las conversaciones de Estu-9 con Estu-50 del ejemplo Beta...58

Figura 32 : Sentimientos en las conversaciones de Estu-42 con Estu-12 del ejemplo Beta.59

Figura 33 : Distribución de sentimientos en el escenario Alfa.....62

Figura 34 : Distribución de sentimientos en el escenario Beta..... 62

Figura 35 : Mensaje "positivo" entre Estu-31 y Estu-9 en el escenario Beta.....62

Figura 36 : Mensaje "negativo" entre Estu-13 y Estu-4 en el escenario Beta.....63

Figura 37 : Mensaje "positivo" entre dos estudiantes en el escenario Beta..... 63

Figura 38 : Distribución por grupo de edad de los estudiantes egresados en Grado por rama de enseñanza en el curso 2020-2021 (SIIU, 2022).....64

Índice de tablas

Tabla 1 : Objetivos del proyecto..... 5

Tabla 2 : Distribución mentor/alumno en escenario Alfa..... 41

Tabla 3 : Distribución mentor/alumno en escenario Beta..... 43

Tabla 4 : Distribución mentor/alumno en escenario Alfa..... 47

Tabla 5 : Distribución mentor/alumno en escenario Alfa..... 49

Tabla 6 : Distribución en comunidades en escenario Alfa.....51

Tabla 7 : Distribución en comunidades en escenario Beta.....53

Tabla 8 : Resultados de la generación de duplas en el escenario Beta.....60

Tabla 9 : Gráfica del total de alumnos matriculados según el rango de edad..... 65

Tabla 10 : Cronograma del proyecto según los distintos paquetes de trabajo.....70

Índice general

Resumen.....I

Abstract.....III

Conceptos clave.....V

Índice de figuras.....VII

Índice de tablas.....IX

Introducción.....2

1.1 Motivación y contexto.....2

1.2 Objetivos.....4

1.3 Desarrollo del proyecto.....5

Estado de la cuestión.....8

2.1 Aprendizaje colaborativo.....8

Sistemas de aprendizaje colaborativo asistido por ordenador.....9

Teoría de la colaboración.....10

Sistemas educativos basados en redes sociales.....10

2.2 Análisis de redes.....12

2.3 Aprendizaje automático.....15

Algoritmos de clustering.....16

2.4 Análisis de sentimientos.....18

2.5 Otras propuestas de mentoría: similitudes y diferencias.....19

Estudio.....22

3.1 Herramientas de estudio.....25

3.2 Parámetros de entrada.....26

3.3 Preprocesamiento.....27

Fase I..... 28

Fase II..... 30

3.4 Procesamiento.....34

Clasificación de alumnos.....35

Modularidad..... 50

Análisis de sentimientos.....53

3.5 Análisis de los resultados..... 59

3.6 Conclusiones del estudio..... 60

El tamaño sí importa.....60

La búsqueda de sentimientos en un mundo de neutrales.....61

La (posible) revolución de lo estático..... 63

Propuesta de proyecto.....66

4.1 Contenido y alcance del proyecto.....66

Alcance del proyecto..... 66

4.2 Plan de trabajo..... 66

WP1 - Gestión del proyecto..... 66

WP2 - Planificación y diseño..... 67

WP3 - Algoritmo de generación de duplas.....67

WP4 - Visualización e interacción de datos.....68

WP5 - Pruebas y depuración..... 68

WP6 - Documentación y despliegue..... 68

WP7 - Mantenimiento y mejora continua..... 69

Cronograma del proyecto..... 69

4.3 Plan de explotación.....71

Análisis del mercado y demanda.....71

Competencia.....71

Actividades de promoción y comercialización previstas..... 72

Conclusiones y trabajo futuro.....74

5.1 Conclusiones del trabajo.....74

5.2 Trabajo futuro.....76

Bibliografía.....78

Apéndice A: Recursos..... 84

Apéndice B: Siglas..... 86

Capítulo 1

Introducción

1.1 Motivación y contexto

Aunque la educación a distancia es, probablemente, el área de la educación que crece con mayor rapidez a nivel internacional, aún sufre de una debilidad fundamental: la alta tasa de abandono que experimentan sus estudiantes en comparación con la tasa de abandono que se podría encontrar en la educación convencional. Esta amplia diferencia se puede encontrar plasmada en la 7ª edición del informe U-Ranking, realizado por la Fundación BBVA (Fundación BBVA, 2019). En universidades de carácter no presencial (como es el caso de la UNED), la tasa de abandono en grados durante el primer año es mayor del 30% y, si nos centramos en la tasa de abandono global, se puede observar que este valor aumenta por encima del 50% (Figura 1).

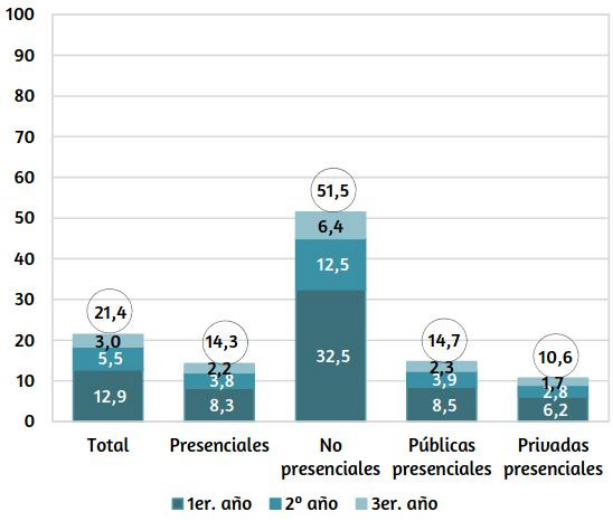

Figura 1: Tasa de abandono global por tipo de universidad y año de abandono (2012-2013)

Sin embargo, este problema no es aislado de la UNED, sino que afecta a todas las universidades de este carácter. En (Pulgar et al., 2014, p. 211) se expone que "*en la UNED, la pauta de abandono es similar a la de universidades comparables, como la Open University británica, en el cuánto, el cuándo y el cómo*". Además, se añade que según estudios realizados en la Unidad Técnica de Investigación del IUED "*no es 'la distancia' el mayor factor causante del abandono de los estudios en la UNED, sino más bien el no poder compatibilizar las*

condiciones personales, familiares y laborales, lo que restringe las proyecciones de éxito en el egreso" (Pulgar et al., 2014, p. 212). En (Mi Educación Online, 2023) se señalan 4 desafíos comunes en la educación a distancia:

- **Falta de apoyo:** Causada por la ausencia física de un docente que guíe o apoye al estudiante durante el aprendizaje. La comunicación entre ambas partes queda reducida al email (o canales similares) lo cual, para algunos estudiantes, puede resultar especialmente complicado.

- **Sentimientos de aislamiento:** La colaboración entre estudiantes es una parte fundamental en el aprendizaje. Las discusiones y los grupos de trabajo ayudan a desarrollar y afianzar los conocimientos y, aunque estas colaboraciones aun son posibles en el aprendizaje a distancia, no tienen el mismo impacto social que trabajar juntos en persona.

- **Disciplina académica:** A diferencia de los modelos presenciales, en el modelo a distancia el aprendizaje es asíncrono, trabajando con el material de clase y aprendiendo de forma independiente. Esto exige una alta autodisciplina que se traduce en la necesidad de unas habilidades de funcionamiento ejecutivo como la organización, la priorización o la administración del tiempo.

- **Tecnología:** Los avances en este campo han permitido que la educación a distancia sea más viable, pero también obliga al estudiante a tener acceso a toda la tecnología que necesita para poder usar todas las herramientas necesarias de manera efectiva.

Son por estos factores, por lo que es imprescindible centrarse en superar esta debilidad a través del apoyo a los estudiantes a distancia. Muchos de los esfuerzos en la investigación en este campo se han centrado en desarrollos de *e-learning* para uso de los estudiantes. Aquí encontramos, por ejemplo, plataformas de aprendizaje, blogs, *wikis*, foros y redes sociales (que han tenido un creciente interés en los últimos años debido a su popularidad) (Boyle et al., 2010).

Sin embargo, algunos estudios (Simpson, 2005) sugieren que el uso de estas técnicas no se ha traducido en avances en la retención de estudiantes. Es por ello por lo que se plantea un acercamiento más tradicional y que se ha usado desde hace mucho tiempo en la educación convencional, la mentoría estudiante-estudiante.

Como ejemplo de este tipo de programas en universidades de carácter presencial, se encuentra ETSII Orienta de la Universidad de Sevilla (ETSII, 2023) cuyo objetivo principal es mejorar las tasas de rendimiento y de éxito especialmente en las asignaturas de primer curso y, como consecuencia, reducir la tasa de abandono. Los mentores, que son estudiantes de últimos cursos, también se benefician de este programa al permitirles desarrollar diversas habilidades sociales (comunicación y empatía).

Este programa presenta, al igual que los correspondientes de otras universidades, una desventaja, el proceso de creación de las duplas de estudiantes. En la mayoría de casos estas se generan al azar, sin tener en cuenta parámetros que puedan indicar cómo de compatibles son dichos estudiantes. Esto puede dar como resultados duplas menos eficaces que proporcionen resultados limitados o, incluso, contraproducentes.

Otra característica común de los programas de este estilo es que los posibles mentores surgen de un sistema de voluntariado. Este método nos puede asegurar la disposición de dichos alumnos a ayudar a otros; sin embargo, no nos asegura que estos tengan las capacidades sociales suficientes para ser considerados buenos candidatos. De la misma manera, aquellos alumnos que tomen el papel de "*protegido*" lo harán, también, voluntariamente (debido a que, por su situación, comprendan que lo necesitan) o por la recomendación de algún docente. Sin embargo, este método presenta un gran peligro, que un alto número de alumnos que necesiten de mentoría no sean siquiera considerados inicialmente. Es por estos dos motivos por lo que una correcta clasificación de los estudiantes, en mentores y "*protegidos*" (lo que más adelante denominaremos como los perfiles "*líder*" y "*periférico*" respectivamente) para la creación de estas duplas es esencial para alcanzar un alto porcentaje de resultados exitosos.

La principal motivación de este proyecto, más allá del propio estudio de la red, es de carácter pedagógico. Siendo conscientes de la alta tasa de abandono en los grados de las universidades no presenciales, se busca proponer un sistema de mentoría que, por medio del análisis de redes, sea capaz de generar duplas de estudiantes compatibles que maximicen los resultados exitosos.

1.2 Objetivos

Centrémonos ahora en los objetivos concretos del proyecto. El primero de ellos se enfoca en plantear un algoritmo que permita generar duplas de estudiantes (un mentor y un alumno) a partir del análisis de redes centrado en las interacciones de estos entre ellos en un foro universitario. Debemos tener en cuenta que este análisis deberá ser capaz, por un lado, de clasificar a los participantes del foro según determinadas medidas de la red como su capacidad de sociabilidad, su popularidad o el grado de acceso que tenga este a la información. Y, por otro lado, deberá ser capaz de detectar como de compatibles son dos estudiantes en base a su actividad y las emociones presentes en sus mensajes en el foro. Este análisis se llevará a cabo en profundidad en la sección 'Estudio' de este documento.

El segundo de los objetivos se centra en, a partir de los resultados obtenidos en el estudio anterior y con la finalidad de reducir la alta tasa de abandono en los grados de universidades no

presenciales, desarrollar una propuesta de solución basada en un sistema de mentoría en la que se expondrá el contenido y alcance de este, un plan de trabajo, un presupuesto y un plan de explotación. Esto se puede encontrar en la sección 'Propuesta de proyecto' de este documento.

En la Tabla 1 podemos ver los objetivos del proyecto:

Nombre	Descripción	Código	Subobjetivos
Algoritmo de generación de duplas	Planteamiento de un algoritmo que permita generar duplas mentor-estudiante.	OBJ. 1	OBJ. 1.1 - Deberá ser capaz de clasificar los participantes de un foro según diversas medidas de la red
			OBJ. 1.2 - Deberá ser capaz de detectar como de compatibles son dos participantes en base a la actividad y las emociones proyectadas en esta.
Desarrollo de propuesta de solución	Desarrollar una propuesta de solución a partir de las conclusiones obtenidas en OBJ. 1.	OBJ. 2	OBJ. 2.1 - Deberá incluir un plan de trabajo.
			OBJ. 2.2 - Deberá incluir un presupuesto.
			OBJ. 2.3 - Deberá incluir un plan de explotación.

Tabla 1: Objetivos del proyecto

1.3 Desarrollo del proyecto

En esta última sección de la introducción, se resumirán los contenidos de los sucesivos capítulos que compondrán el trabajo:

- **Estado de la cuestión:** En este se explicará el estado en el que se encuentran las investigaciones de cada uno de los puntos que se tratará en este proyecto, indicando las bases por las que están formulados. Nos centraremos, especialmente, en el aprendizaje

colaborativo, el análisis de redes, el aprendizaje automático, el análisis de sentimientos y expondremos algunas de las propuestas de mentoría y las diferencias con la nuestra.

- **Estudio:** En este capítulo se realizará un recorrido por el propuesto proyecto para el algoritmo de detección de duplas mentor-estudiante (especificación de los parámetros de entrada y desarrollo de las fases de preprocesamiento y procesamiento). Al final de este capítulo, se analizarán los resultados y se expondrán algunas de las conclusiones alcanzadas durante el mismo.

- **Propuesta de proyecto:** Aquí se expondrá todo lo relativo a la propuesta del proyecto incluyendo su contenido, alcance, plan de trabajo y plan de explotación.

- **Conclusiones y trabajos futuros:** Finalmente, para este último capítulo, se expondrán las conclusiones generales del TFM centrándonos en los objetivos propuestos en la sección anterior y si estos se han cumplido correctamente. Además, se propondrá una línea de posibles trabajos futuros que se puedan realizar utilizando este TFM como base.

Capítulo 2

Estado de la cuestión

En este capítulo se explicará el estado en el que se encuentran las investigaciones de cada uno de los puntos que se tratará en este proyecto, así como las bases por las que están formulados. Con este fin, dividiremos este apartado en varias secciones según la línea de investigación.

2.1 Aprendizaje colaborativo

Con el fin de evitar que el aprendizaje en los centros educativos sea excesivamente individualista, es habitual la realización de tareas en grupos. Este aprendizaje colectivo se puede llevar acabo de diversas maneras y, es por ello, por lo que es imprescindible precisar las diferencias que existen entre aprendizaje cooperativo y colaborativo.

El aprendizaje cooperativo tiene por objetivo trabajar juntos para lograr objetivos compartidos y que den lugar a resultados que sean beneficiosos tanto individualmente como para el resto de miembros del equipo (Johnson y Johnson, 2005). En este escenario, los participantes dividen el trabajo, resuelven las tareas individualmente y luego unen los trabajos parciales en el producto final. En la colaboración, sin embargo, un grupo de personas trabaja en apoyo de los objetivos de otros (Moseley, 2020).

La colaboración, por lo tanto, es un proceso que requiere que los individuos negocien y compartan significados relevantes a una tarea de resolución de problemas. La colaboración es una actividad coordinada, sincrónica, que es el resultado de un intento continuo de construir y mantener una concepción compartida de un problema.

Según (Boxtel et al., 2000) las actividades de este tipo de aprendizaje permiten ofrecer al estudiante una explicación de su entendimiento, que puede permitirles elaborar y reorganizar su conocimiento. Las interacciones sociales estimulan la elaboración del conocimiento conceptual intentando que sea entendible por todos los miembros del grupo. La investigación demuestra que fomentar la generación de explicaciones elaboradas mejora la comprensión de conceptos por parte del estudiante. Una vez que el entendimiento conceptual se ha hecho visible por medio

del intercambio verbal, los estudiantes pueden negociar para alcanzar una convergencia o un entendimiento compartido.

Las ventajas del aprendizaje colaborativo son, por tanto, bastantes numerosas. En (Cornell University, 2023) encontramos las siguientes:

- Desarrollo de diversas habilidades como pensamiento de alto nivel, comunicación oral, autogestión y liderazgo.
- Fomento de la interacción entre estudiantes y profesores.
- Aumento de la retención, la autoestima y la responsabilidad de los estudiantes.
- Exposición a diversas perspectivas y aumento de la comprensión de las mismas.
- Preparación para situaciones sociales y laborales del día a día.

De todos los puntos anteriores, nos fijaremos especialmente en el tercero. En este se especifica que el aprendizaje colaborativo es capaz de aumentar la retención, la autoestima y la responsabilidad de los estudiantes. Un estudiante que no emplee este tipo de aprendizaje, por lo tanto, será más propenso a tener una menor autoestima y responsabilidad, lo que provocará que se genere la creencia de que no sea capaz de conseguir los objetivos marcados por los docentes y, finalmente, acabe abandonando los estudios.

Evidentemente, al igual que el aprendizaje colaborativo puede fomentar la interacción entre estudiantes y profesores (segundo punto), también lo puede hacer entre estudiantes en un sistema de mentoría como el que proponemos en este proyecto.

Sistemas de aprendizaje colaborativo asistido por ordenador

Dado que este proyecto se centra en la enseñanza a distancia, como la que proporciona la UNED, nos centraremos en sistemas de aprendizaje colaborativo desarrollados para estos ambientes, como el CSCL.

Según (Anaya, 2019a), "*los sistemas de aprendizaje colaborativo asistido por ordenador (CSCL en sus siglas en inglés) estudian cómo se puede mejorar la interacción y el trabajo en grupos y cómo la tecnología puede facilitar el compartir, la distribución del conocimiento y la experiencia entre los miembros de una comunidad. Para que la colaboración se garantice deben cumplirse una serie de requisitos y la tecnología puede ayudar a garantizarlos teniendo en cuenta tanto el análisis de las interacciones realizadas como los objetivos y el planteamiento establecido*".

La visión del CSCL está basada en la idea de "*intentar desarrollar nuevos productos y aplicaciones software que les brinden a los usuarios actividades creativas de exploración intelectual compartida y de interacción, destinadas a favorecer el aprendizaje*" (Anaya, 2019a).

Teoría de la colaboración

Esta teoría fue postulada por (Stahl, 2004) quién sugiere que el conocimiento no se restringe a las habilidades, memorias y esfuerzos del individuo, sino que se afirma que este es construido en interacciones sociales dentro de las comunidades de personas.

El objetivo de la teoría de colaboración es preparar a los estudiantes para que puedan desempeñar un rol efectivo en la sociedad del futuro. Y es que mientras que en 1980, solo el 20% del trabajo estaba basado en dinámicas de equipo, en 2010 este porcentaje subió hasta el 80% (Hollenbeck et al., 2012). El conocimiento que se transmite a estos, por lo tanto, no puede depender exclusivamente de conocimiento ya existente que los estudiantes asimilan pasivamente. Se les debe guiar para desarrollar sus habilidades sociales y personales que les permitirá encontrar información relevante a problemas no anticipados y que promuevan, al mismo tiempo, la participación en procesos de investigación.

Sistemas educativos basados en redes sociales

Dentro de la etiqueta de sistemas de aprendizaje colaborativo asistidos por ordenador, nos centraremos en aquellos basados en redes sociales debido a que el estudio que se realiza en este proyecto utiliza como conjunto inicial de datos las redes generadas en los foros universitarios de distintas asignaturas.

En (RD Station, 2023) se definen las redes sociales como "*estructuras formadas en Internet por personas u organizaciones que se conectan a partir de intereses o valores comunes. A través de ellas, se crean relaciones entre individuos o empresas de forma rápida, sin jerarquía o límites físicos*".

Las redes sociales pueden clasificarse según sus características, pero, por lo general, podemos dividirlas en dos grupos (Armando, 2017):

- **Redes sociales horizontales:** Se caracterizan por no haber sido creadas enfocada en ningún tipo de usuario en específico, sino que están pensadas para que en ellas pueda participar cualquier tipo de individuo (con la posibilidad de crear sus propias comunidades dentro de la red social). En este grupo encontramos las redes sociales más famosas como Twitter o Facebook.

- **Redes sociales verticales:** Estas, a diferencia de las anteriores, si están dirigidas a un público determinado. Dependiendo de este tipo de público, podemos subdividir esta categoría en otras como por ejemplo:
  - **Redes sociales profesionales:** En la que los participantes son profesionales que interactúan con objetivos laborales.
  - **Redes sociales de ocio:** La temática de este tipo de red social gira en torno a su temática: deporte, música, videojuegos, cine...
  - **Redes sociales mixtas:** Combinan tanto temáticas profesionales como de ocio haciéndolas menos formales.
  - **Redes sociales universitarias:** Destinada a un público universitario. Los estudiantes pueden desde hablar entre ellos, y conocerse, hasta, en algunos casos, incluso descargar apuntes.

Dentro de nuestro escenario, sería en este último tipo de redes en el que nos interesaría centrarnos. Este se ha comenzado a emplear en la educación a distancia (aunque aún sin establecer patrones o metodologías que puedan potenciar esta herramienta) y, debido a ello, se han realizado numerosas investigaciones que han señalado tanto ventajas como inconvenientes del uso de las redes sociales en entornos educativos.

Dentro de las ventajas encontramos:

- Animan a los estudiantes a interactuar entre ellos, compartir ideas y expresarse con creatividad.
- Ayudan a establecer relaciones duraderas con personas reales.
- Estas relaciones creadas en un contexto social pueden ser redimensionadas a un contexto de comunidad.
- Una comunidad agrupa a personas con un interés u objetivo común. Esto les permite ser más apropiadas para realizar tareas en conjunto.
- Debido a la gran cantidad de información en los sitios de redes sociales (en Internet en general) los estudiantes aprenden a discernir fácilmente entre lo que es útil de lo que no lo es.
- Porque son fáciles de usar y acceder en cualquier lugar y a cualquier tiempo, las redes sociales mejoran la comunicación entre los estudiantes y, además, con los docentes.

- Las redes sociales preparan a los estudiantes para la vida profesional o para encontrar trabajo.

Por otro lado, entre las desventajas encontramos:

- Las redes sociales ofrecen tanta información que el alumno puede verse superado y no atender a lo importante.
- Los procesos de razonamiento de alto orden (focalización en un objeto de pensamiento, concentración, persistencia) de los estudiantes pueden verse trastornados, siendo éstos necesarios para el pensamiento críticos debido a la enorme cantidad de datos, opiniones, mensajes que se pueden recibir sobre un determinado tema y que requerirá más tiempo y esfuerzo por parte del estudiante
- Se ha investigado la correlación entre el uso excesivo de Internet y la mayor impulsividad, menor paciencia, menor tenacidad y más débil capacidad para el pensamiento crítico.
- El uso prolongado de Internet expone al alumno a estímulos interactivos, repetitivos y adictivos que producen cambios permanentes en la estructura cerebral y daña las habilidades de aprendizaje.
- Investigaciones sugieren que un mayor uso de Internet y las redes sociales pueden degradar la capacidad de concentración.

Cabe indicar que estas características negativas son más aplicables a las redes sociales y a Internet en general que a su uso en la educación.

Aún así, lo que es una realidad es que en los últimos años, y potenciado especialmente por la pandemia global del COVID-19, se ha experimentado un tremendo auge en la educación a distancia y se espera que este aumento continúe creciendo en los años venideros (Prieto, 2021).

2.2 Análisis de redes

(Dans, 2010) nos proporciona una definición sobre las redes sociales algo más técnica de la que encontramos en la sección anterior: "*Las redes sociales son una estructura social que se puede representar en forma de uno o varios grafos, en los cuales, los nodos representan a individuos (a veces denominados actores) y las aristas, relaciones entre ellos. Las relaciones pueden ser de distinto tipo, como intercambios financieros, amistad, relaciones sexuales o rutas aéreas. También es el medio de interacción de distintas personas, como por ejemplo, juegos en línea, chats, foros, spaces, entre otros*".

El análisis de redes sociales (SNA, por el término en inglés, *social network analysis*) "*es un paradigma de investigación que permite tener una visión de la realidad, basada en las relaciones sociales que los actores establecen generando un entramado que da respuesta a problemas de la sociedad, problemas que se estudian en conjunto con múltiples disciplinas*" (López e Ibarra, 2021). Este paradigma se basa en las implicaciones o consecuencias que existen entre los actores y tiene por objetivo encontrar el impacto de estas y las causas que las originan.

Los fundamentos teóricos del SNA se remontan entre 1919 y 1960 y "*tienen su origen en disciplinas como la ciencia del comportamiento, la psicología, la sociología y la antropología, con autores como Wellman, Bott, Almark, Comte, Moreno, Durkheim y Barnes*" (López e Ibarra, 2021). Sin embargo no fue hasta las décadas de 1960 y 1970 cuando los desarrollos matemático y de software (encargado de procesar los datos) provocaron un auge en la investigación.

Dentro de los entornos educativos, como el que nos concierne en este proyecto, el análisis de las redes sociales de estudiantes puede ayudarnos a identificar como métodos educativos mejorados (como el aprendizaje colaborativo descrito en la anterior sección) pueden ser usados para que el aprendizaje sea más efectivo e inclusivo, tanto a nivel universitarios como en etapas pre-universitarias. El objetivo final es lograr el desarrollo integral de los estudiantes, mediante la expansión de sus redes sociales, así como controlar la propagación de posibles conductas delictivas (Saxena et al. 2019).

Analizar la estructura de la red no solo indica el grado de beneficio del estudiante de las interacciones entre ellos, sino que también muestra la importancia de un actor o grupo concreto en los entornos de aprendizajes modelado por las redes. Según (Saxena et al. 2019) existen 6 aspectos principales relacionados con la educación que el análisis de las redes de estudiantes destacan:

- **La efectividad de los entornos de aprendizaje colaborativo:** Esta efectividad en entornos educativos se puede medir haciendo uso del SNA, especialmente en aquellos sistemas basados en CSCL. Al analizar las interacciones y las redes de los estudiantes en este entorno, es posible determinar el nivel de participación de los estudiantes, identificar posibles mejoras en la comunicación y fomentar una mayor interacción entre los estudiantes. Además, es posible examinar cómo la estructura de una red afecta al trabajo en equipo y cómo factores, como la resolución de dudas, el intercambio de información y la influencia de los compañeros, influye en el aprendizaje colaborativo.

- **Estudios sobre el trabajo en equipo entre estudiantes:** El trabajo en equipo es fundamental entre los estudiantes para realizar actividades como proyectos grupales o discusiones, y construir conexiones a largo plazo. La comunicación y la cohesión dentro de los equipos son factores clave para lograr resultados positivos. Además, la presencia de

intermediarios que transfieren conocimientos entre grupos y la inclusión de todos los estudiantes son importantes para el éxito del equipo. El análisis de redes sociales ayuda a comprender estas interacciones y a identificar estudiantes desconectados.

- **Rendimiento académico frente a las redes sociales:** Analizar la correlación entre el rendimiento académico de los estudiantes y su posición en la red social revela que los estudiantes de alto rendimiento tienden a ocupar posiciones centrales en la red y establecen vínculos persistentes con sus compañeros. Además, se ha encontrado que la similitud en el comportamiento de los estudiantes se correlaciona con similitud en las calificaciones, y la participación en grupos de estudio se relaciona con un mayor número de vínculos en la red.

- **Divulgación de conocimientos y apoyo entre iguales:** La difusión del conocimiento en las redes de estudiantes se ve facilitada por la presencia de líderes, una densidad óptima de la red y la existencia de subgrupos con conexiones inter-subgrupo adecuadas. La formación de grupos de proyectos está influenciada inicialmente por las amistades preexistentes, pero los estudiantes buscan ayuda entre aquellos que tienen conocimiento sobre el tema (los estudiantes de alto rendimiento) los cuales, una vez que las relaciones entre los estudiantes se estabilizan, acaban recibiendo más atención de sus compañeros. La implementación de foros de discusión en línea fortalece, además, las redes sociales.

- **Estudio sobre subculturas y homofilia entre estudiantes:** La homofilia, que se refiere a la tendencia de las personas a formar relaciones con aquellos que comparten rasgos similares, como el rendimiento académico y el género, se ha observado en las interacciones entre estudiantes de secundaria. Los estudios han demostrado que los estudiantes tienden a buscar amistades con otros estudiantes que tienen un rendimiento académico similar, lo que indica una fuerte homofilia académica. Sin embargo, la homofilia también puede tener desventajas, como la formación de grupos entre estudiantes con bajo rendimiento, lo que puede afectar negativamente su desempeño académico. Superar la homofilia puede promover mejores relaciones interraciales y étnicas entre los estudiantes.

- **Estudios sobre el comportamiento adolescente:** La amistad entre estudiantes puede influir en su comportamiento de diversas maneras, como mejorar el rendimiento académico o adoptar conductas antisociales. Existe una tendencia a establecer conexiones con compañeros que tienen comportamientos similares, incluyendo la agresividad. Además, el comportamiento delictivo y los hábitos como fumar, consumir alcohol, marihuana y comida chatarra también se ven influenciados por el círculo social. Estos patrones de comportamiento pueden ser comprendidos a través del análisis de las redes sociales y utilizados para diseñar políticas educativas y estrategias de prevención. Además, la amistad

también está relacionada con la prevalencia del engaño académico, y la disposición de asientos en los exámenes puede ser optimizada para evitar el fraude.

(Saxena et al. 2019) cierra con la siguiente conclusión: "*Sugerimos que investigaciones futuras se centren en la recopilación automática de datos utilizando plataformas en línea como [...] foros de discusión, que ya han sido utilizados por algunos estudios, junto con la integración de las conexiones de los estudiantes en sitios de redes sociales en línea. Se debería desarrollar software para analizar las redes construidas a partir de los datos con el fin de examinar la participación de los estudiantes e identificar a aquellos que están desconectados, para que el profesorado pueda tomar medidas correctivas apropiadas*".

2.3 Aprendizaje automático

En este apartado nos centraremos en los algoritmos de aprendizaje automático (ML, por el término en inglés, *Machine Learning*) que se emplearán en la fase de estudio de este proyecto para clasificar los diferentes nodos de la red según unas características concretas.

El aprendizaje automático puede definirse como "*el proceso mediante el cual se usan modelos matemáticos de datos para ayudar a un equipo a aprender sin instrucciones directas. [...] El aprendizaje automático usa algoritmos para identificar patrones en los datos, y esos patrones luego se usan para crear un modelo de datos que puede hacer predicciones*" (Azure, 2023). Es, en concreto, su adaptabilidad lo que lo convierte en una solución perfecta en escenarios en los que los datos sufren cambios constantemente.

Este enfoque computacional para el aprendizaje está destinado a ser ampliamente utilizado en las ciencias de la educación y se perfila como una herramienta analítica importante (como ya lo es en muchos otros campos). Los diversos casos de usos del ML en la educación son evidentes en algunas áreas como los sistemas inteligentes de tutorías (Conati et al., 2018) o la predicción del rendimiento de los estudiantes (Đambić et al., 2016), y subestimados en otros, como la validación de modelos estadísticos latentes (Bleidorn y Hopwood, 2019). Según (Hilbert et al., 2021), más que proporcionar técnicas analíticas, el ML puede ayudar a los investigadores educativos a cambiar la cultura de modelado hacia una ciencia más confiable con un mayor enfoque en la predicción real de datos novedosos, como ya se ha propuesto para el campo de la psicología, pero que aún no se ha incorporado completamente en las ciencias de la educación. En particular, el enfoque en modelos robustos con predicciones confiables es lo que se debería utilizar para avanzar en la ciencia educativa.

De los casos de uso descritos en el párrafo anterior nos interesa aquel centrado en la predicción del rendimiento de los estudiantes (Đambić et al., 2016). En este estudio se aborda la

dificultad que enfrentan algunos estudiantes de educación superior al comprender los conceptos básicos de programación. Se propone utilizar el aprendizaje automático para predecir, al comienzo del semestre, qué estudiantes pueden tener dificultades para aprobar el curso de *Introducción a la Programación*. Con esta información, se seleccionarán los estudiantes que probablemente necesiten actividades de aprendizaje adicionales. Se describe cómo se ha implementado este enfoque en una universidad, con clases adicionales y tutores para ayudar a los estudiantes con dificultades. El objetivo es mejorar la tasa de aprobación y proporcionar un apoyo temprano a los estudiantes en riesgo.

Dentro del aprendizaje automático, existen diversas técnicas que pueden ser clasificadas en dos categorías (Anaya, 2019c):

- **Aprendizaje supervisado:** Los conjuntos de datos son abordados con etiquetas. Estos algoritmos son entrenados con un subconjunto de estos datos con el valor de la clase para que el sistema aprenda como se han clasificado estás instancias. Una vez que este proceso de aprendizaje ha finalizado, el algoritmo es capaz de clasificar el resto de instancias dando a cada una de ellas un valor de clase. El aprendizaje supervisado puede ser dividido en clasificación (si toma valores discretos) o regresión (si, por el contrario, toma valores lineales).
- **Aprendizaje no supervisado:** Se caracterizan porque el conjunto de datos carece de un atributo de clase (o no se quiere usar) y la agrupación de las instancias se realiza teniendo en cuenta su similitud.

Dentro de nuestra propuesta de solución se ha decidido optar por algoritmos de aprendizaje no supervisado, concretamente por los algoritmos de clustering. Sobre el porqué de esta decisión se hablará más en profundidad en la sección de "Estudio", aquí nos centraremos en definir y desarrollar este tipo de algoritmo.

Algoritmos de clustering

Los algoritmos de clustering "*tiene como objetivo agrupar los objetos de un dataset según su similaridad, de forma que los objetos que hay dentro de un grupo (cluster) sean más similares que aquellos que caen en grupos distintos*" (Sancho, 2020). Intuitivamente, el objetivo de este tipo de algoritmos es muy claro: agrupar adecuadamente un conjunto de datos no etiquetados. Sin embargo, la falta de precisión en la definición de "cluster" ha provocado la existencia de un amplio rango de algoritmos de este tipo.

A pesar de esto, existen puntos en común para todos los algoritmos de clustering. (Kleinberg, 2002) propone 3 axiomas que pueden cumplirse, independientemente del algoritmos empleado para encontrar la solución:

1. **Invariancia por escala:** Al aplicar una escala al conjunto de puntos, un algoritmo de clustering no debe dar resultados distintos.

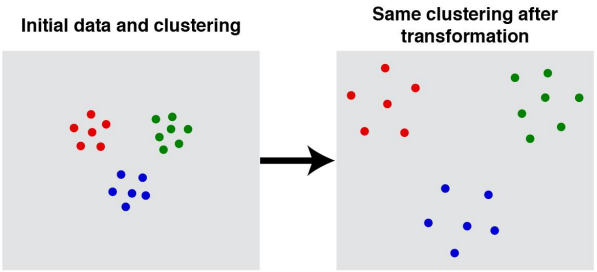

Figura 2: Ejemplo de invariancia por escala (Sancho, 2020)

2. **Consistencia:** Un algoritmo de clustering no debe variar sus resultados si las distancias dentro de cada cluster se reduce (o si la distancia entre clusters se aumenta).

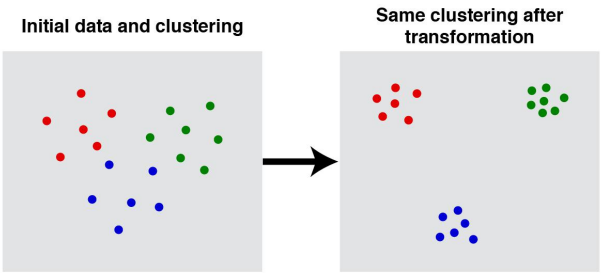

Figura 3: Ejemplo de consistencia (Sancho, 2020)

3. **Riqueza:** La función de agrupación debe ser lo suficientemente flexible como para ser capaz de producir cualquier agrupación arbitraria del conjunto de datos de entrada.

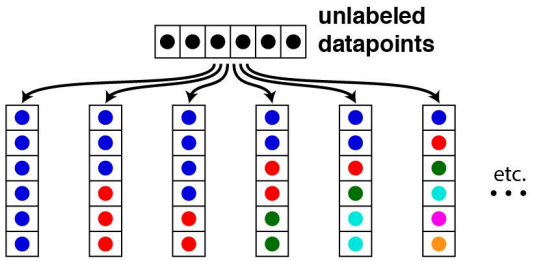

Figura 4: Ejemplo de riqueza (Sancho, 2020)

Lo más interesante sobre estos axiomas propuestos por (Kleinberg, 2002) es que, a pesar de lo intuitivos que son, es imposible construir una función que sea capaz de verificar los 3 simultáneamente. Esto da como resultado que sea posible diseñar algoritmos de clustering que viole uno de los axiomas mientras cumple los otros dos.

El procedimiento normal de los algoritmos de clustering consta de los siguientes pasos (Anaya, 2019c):

- **Selección de características o atributos:** Estos deben ser seleccionado correctamente para maximizar la cantidad de información útil para los intereses de la tarea a realizar (minimizando, al mismo tiempo, la redundancia de información).
- **Medida de proximidad:** Estos algoritmos agrupan las instancias según su semejanza con medidas de proximidad (hipótesis de mayor proximidad, mayor semejanza) usando distancias métricas. Según el tipo de distancia que se use, es importante que los atributos estén normalizados.
- **Criterios de agrupación:** Estos dependen del algoritmo de cluster seleccionado y pueden ser compactos (cada instancia solo pertenece a un cluster), probabilístico (cada instancia pertenece a distintos clusters con diferente probabilidad) o borrosos (cada instancia pertenece a distintos clusters según su grado de ambigüedad).
- **Algoritmos de clustering:** Esta elección dependerá de las decisiones tomadas en los anteriores puntos.
- **Validación de los resultados:** Con el uso de tests apropiados, se puede llevar a cabo la corrección de los resultados.
- **Interpretación de los resultados:** Un experto del campo, en muchas ocasiones, puede sacar conclusiones correctas estudiando la agrupación.

2.4 Análisis de sentimientos

(Shivanandhan, 2020) define el análisis de sentimientos como una técnica a través de la cual puedes analizar un fragmento de texto para determinar el sentimiento detrás de él. Combina el aprendizaje automático y el procesamiento del lenguaje natural (NLP, por el término en inglés, *natural language processing*) para lograr esto. Utilizando un análisis básico de sentimientos, un programa puede entender si el sentimiento detrás de un fragmento de texto es positivo, negativo o neutral. Por ejemplo, se puede utilizar el análisis de sentimientos para analizar los comentarios de los clientes. Después de recopilar esos comentarios a través de diferentes medios como Twitter y Facebook, se puede ejecutar algoritmos de análisis de

sentimientos en esos fragmentos de texto para comprender la actitud de tus clientes hacia tu producto.

En el entorno educativo, el análisis de sentimientos tiene el potencial de extraer las opiniones de los estudiantes a nivel de documento, nivel de oración, nivel de entidad y nivel de aspecto, junto con su orientación de sentimiento (Shaik et al., 2023). A nivel de documento, se analiza un comentario y se determina el sentimiento general de los comentarios hacia un curso, ya sea positivo, negativo o neutral. A nivel de oración, se extrae el sentimiento de cada oración y ayuda a calcular los aspectos positivos y negativos de un curso. La extracción de sentimientos a nivel de entidad combina el análisis de entidad y sentimiento para proporcionar la opinión de los estudiantes sobre una entidad como tutor, curso y tarea. El análisis de sentimiento basado en aspectos es un análisis detallado de diferentes categorías de datos en un comentario e identifica la orientación del sentimiento en cada categoría de datos. Según una aplicación educativa, los datos de retroalimentación de los estudiantes se analizarán en diferentes niveles. Por ejemplo, una aplicación de toma de decisiones consideraría el análisis de sentimientos a nivel de documento, y para comprender la participación de los estudiantes, el análisis se realizaría a nivel de aspecto.

2.5 Otras propuestas de mentoría: similitudes y diferencias

Los resultados positivos de la aplicación de los sistemas de mentoría en entornos educativos ha provocado que la popularización de este tipo de sistemas haya crecido progresivamente en los últimos años.

En la introducción de esta memoria ya expusimos el caso de la Universidad de Sevilla, pero no es el único. En (Boyle et al., 2010) se exponen 3 estudios, de 3 universidades de distintos países, que buscan también reducir el abandono universitario en la educación a distancia y que proponen, al igual que en nuestro caso, un sistema mentor-estudiante que permita darle soporte a estos segundos.

Estas 3 universidades son la Open University UK (Asbee et al., 1999), la Korean National Open University (Lee et al., 2009) y la Open Polytechnic of New Zealand (Crosling et al., 2008) (Earle, 2007) (Gibbs et al., 2007) (Zepke et al., 2003).

En todos estos escenarios se realiza un recorrido por todo el proceso (incluyendo resultados y conclusiones), pero, resulta interesante resaltar los criterios al agrupar los estudiantes, es decir, que consideraciones se han tenido en cuenta al generar las duplas mentor-estudiante:

- **Open University UK:** Se ha considerado el curso, la localización geográfica, la situación doméstica (tiene hijos, padre soltero...), genero y edad (en este caso de acuerdo a preferencias previamente expresadas).
- **Korean National Open University:** En este caso se ha considerado especialmente el curso y las asignaturas.
- **Open Polytechnic of New Zealand:** Por la naturaleza de este proyecto (en el que se busca integrar a estudiantes indígenas en institutos de tecnología), se ha considerado especialmente la cultura.

Un punto en común de estos 3 programas es que ninguno de ellos emplea sistemas automatizados de ningún tipo. Investigando un programa que presente esta características, encontramos el ejemplo de la Universidad Complutense de Madrid, expuesto en (Gómez-Flechoso et al., 2019). Este programa de mentoría hace uso de la automatización para la gestión de "*los datos referentes al cumplimiento y satisfacción tanto de los mentores como de los telémacos*". Estos datos serán usados posteriormente para evaluar el funcionamiento del programa y el efecto que este puede tener en el proceso de aprendizaje. Sin embargo, esta automatización no es empleada para generar las duplas.

Como se puede observar, y a diferencia de lo que proponemos en este proyecto, en todos los casos expuestos se emplean criterios estáticos al generar las duplas mentor-estudiante, que suele ser el criterio común seguido en la mayoría de los programas de este tipo. Además, estos sistemas siguen procedimientos manuales (no automatizados) para el proceso de generación de duplas y ya en anteriores secciones observamos como (Saxena et al. 2019) recalcaba la importancia del uso de procedimientos automáticos que realicen análisis de la red para mejorar los resultados.

Son estos los dos puntos que diferencian nuestra propuesta de la mayoría, el uso de datos dinámicos al generar las duplas mentor-estudiante y el uso del aprendizaje automático.

Capítulo 3

Estudio

En este capítulo se procederá, por un lado, a especificar el proceso desarrollado para la generación de duplas mentor-estudiante, en entornos universitarios, usando aprendizaje automático, y, por el otro, se describirá todo el procedimiento seguido durante el estudio que dio como resultado dicho proceso, justificando los motivos de la toma de las diferentes decisiones, señalando los problemas encontrados y proponiendo anotaciones sobre posibles mejoras.

Dicho proceso de generación de duplas mentor-estudiante, puede verse representado en forma de diagrama de flujo en la Figura 6. A continuación, se explica detalladamente el procedimiento descrito en dicho diagrama:

- **Datos foro:** Estos datos se proporcionarán como parámetros de entrada para el estudio realizado. Más información acerca de los diferentes atributos (columnas) de las instancias puede encontrarse en la sección **3.2 Parámetros de entrada**.
- **Preprocesamiento:** Aquí se realizará una división en dos fases. En la primera se realizarán distintas operaciones de preprocesamiento en los parámetros de entrada de la anterior sección que serán empleados en la segunda fase, encargada de calcular las medidas de la red.
- **Transformación en grafo:** Con el preprocesamiento realizado, podemos transformar esos datos en un grafo en el que cada nodo se corresponde con un usuario y las aristas (direccionales) representan las conversaciones entre dichos usuarios.
- **Procesamiento:** Los cálculos de las medidas de la red serán empleados en esta última sección para el procesamiento de los datos. Para comprender esto, lo dividiremos en 3 fases:
  - **Clasificación de alumnos:** En la Figura 5 podemos observar un extracto del diagrama de flujo de la Figura 6 que corresponde a esta fase. Los usuarios serán listados y se procesarán uno a uno comprobando si este es "*lider*" (mentor) o "*periférico*" (estudiante), añadiendo cada usuario a su respectiva lista (*listaMentor* o *listaEstudiante*) hasta que todos hayan sido comprobados. En caso de no encajar en ninguno de estos perfiles, se descartará de ulteriores procesamientos.

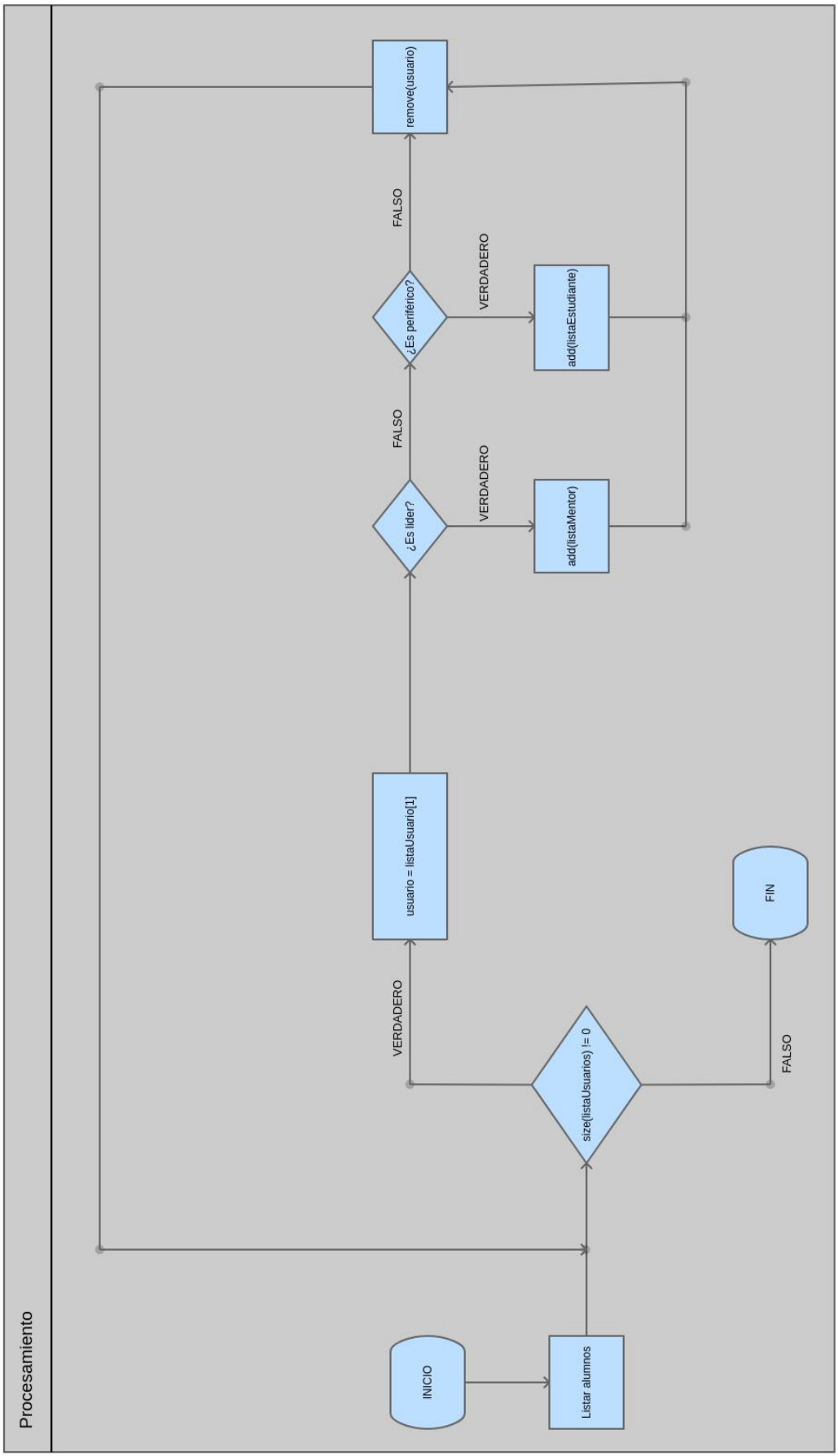

Figura 5: Diagrama de flujo del proceso de clasificación de alumnos

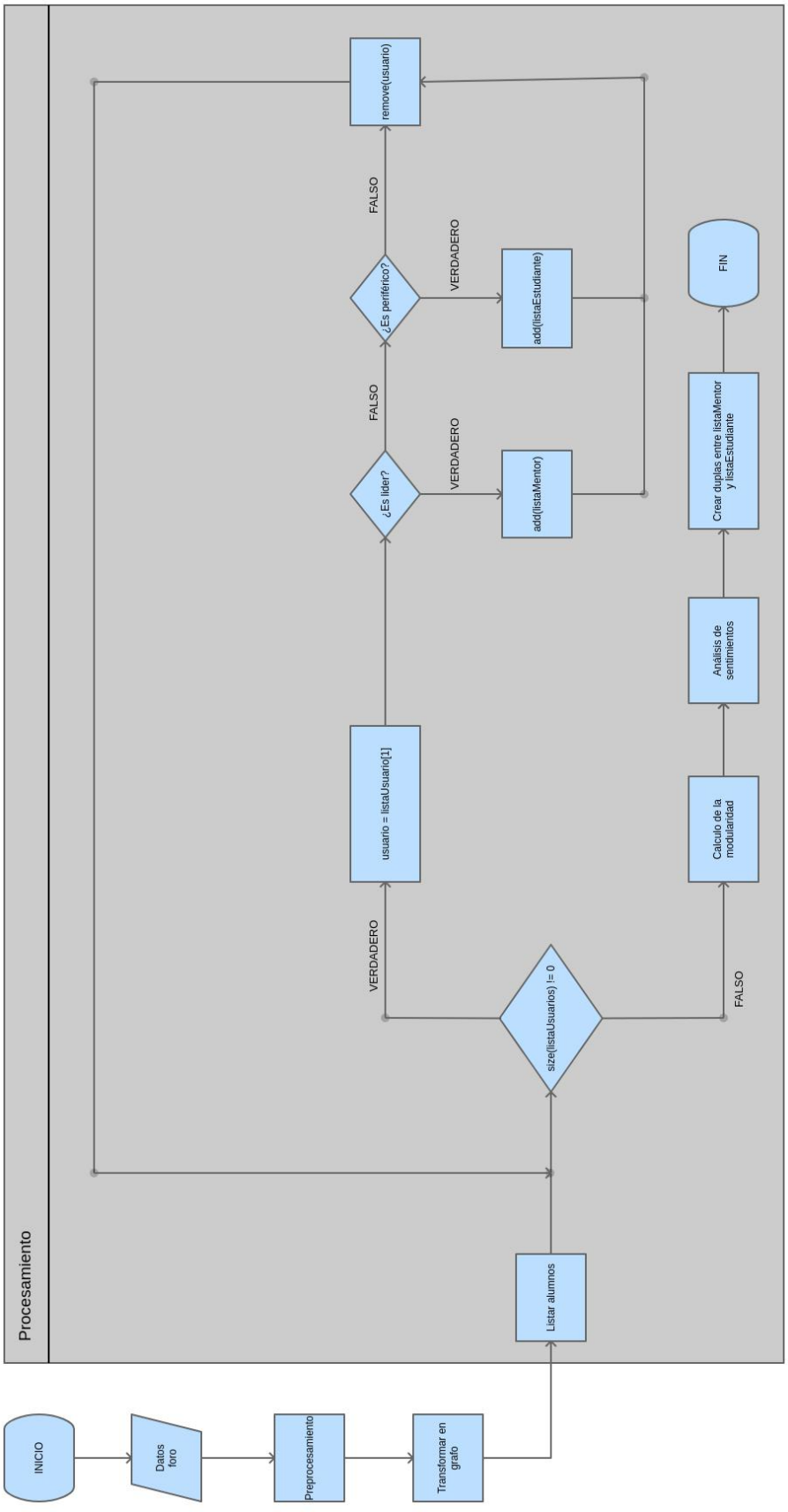

Figura 6: Diagrama de flujo del proceso de generación de duplas mentor-estudiante

- **Modularidad:** Realizará una búsqueda de comunidades y comprobará si existen duplas de mentores-estudiantes que se encuentren en una de estas. Al igual que en el caso anterior, aquellos usuarios (ya sean mentores o estudiantes) que se encuentren en una comunidad aislada serán descartados.

- **Análisis de sentimientos:** Por último, se comprobará entre los usuarios coincidentes en una misma comunidad, de perfiles distintos, si se han intercambiado mensajes y, en caso afirmativo, se comprobarán los sentimientos detectados en estas conversaciones, buscando establecer las duplas de mentor-estudiantes entre aquellos cuyas conversaciones hayan sido positivas.

- **Crear duplas entre listaMentor y listaEstudiante:** A partir de los resultados del procesamiento anterior, generaremos las duplas mentor-estudiante que hayan sido sugeridas por el sistema.

Dado la extensión de la explicación, y para mayor claridad, se ha decidido dividir este capítulo en varias secciones según las diferentes fases anteriormente expuestas.

3.1 Herramientas de estudio

Antes de proceder con el análisis de todo el proceso realizado en el estudio, sería conveniente enumerar las herramientas que han sido utilizadas en el proyecto:

- Pencil Project: Es una herramienta de software de creación de prototipos y diseño de interfaces de usuario. Es una aplicación de código abierto que permite a los diseñadores crear bocetos interactivos y *wireframes* de forma rápida y sencilla. Para el escenario que nos ocupa será utilizada para la realización del diagrama de flujo presentado al principio de este capítulo.

- Gephi: Herramienta de visualización y análisis de redes, diseñada especialmente para explorar y representar visualmente datos complejos de redes. Permite importar y analizar datos de redes, identificar patrones, calcular métricas y crear visualizaciones interactivas de redes. Esta herramienta será empleada durante la fase de preprocesamiento, para la visualización de los grafos y para el calculo de diferentes medidas de la red, y en la fase de procesamiento, para calcular la modularidad.

- Weka: Es una herramienta de software de aprendizaje automático y minería de datos. Es una plataforma de código abierto que proporciona una amplia gama de algoritmos de aprendizaje automático y herramientas de análisis de datos. Weka permite a los usuarios realizar tareas como clasificación, regresión, agrupamiento, asociación y selección de

características. Se empleará, por lo tanto, en la fase de procesamiento, concretamente en la clasificación de alumnos por medio de algoritmos de aprendizaje no supervisados.

- Análisis de Foros UNED: Herramienta desarrollada por Félix Adame Toledano para el Trabajo de Fin de Máster, para el curso 2022/2023, de *Análisis de entornos colaborativos con tecnologías de análisis de redes y Learning Analytics* (Director: Dr. Antonio Rodríguez Anaya) del Máster Universitario en Ingeniería Informática de la UNED. Esta herramienta realiza una serie de análisis sobre datos de conversaciones en foros. Para nuestro caso, la hemos usado, exclusivamente, por su herramienta de análisis de sentimiento. Esta última herramienta se empleará también en la fase de procesamiento para el análisis de sentimientos en los mensajes intercambiados entre los distintos usuarios de la red.

3.2 Parámetros de entrada

Para el proceso se tomará como parámetros de entrada diferentes *datasets* de foros de asignaturas de la UNED proporcionados por mi tutor, Dr. Antonio Rodríguez Anaya, quién se ha encargado previamente de anonimizarlos y eliminar cualquier rastro de información personal según dicta la GDPR (por sus siglas en inglés, *General Data Protection Regulation*).

Cada fila de los *datasets* hace referencia a una instancia que, en este caso, se corresponde con un mensaje del foro; mientras que las columnas, por su parte, representa cada una un atributo. Dichos atributos son los siguientes:

- **idAsignatura:** Código alfanumérico que hace la función de identificador de una determinada asignatura.
- **Asignatura:** Nombre de la asignatura.
- **idForo:** Código alfanumérico que hace la función de identificador de un determinado foro.
- **Foro:** Título del foro.
- **idHilo:** Código alfanumérico que hace la función de identificador de un determinado hilo de mensajes.
- **Hilo:** Título del hilo de mensajes.
- **idMensaje:** Código alfanumérico que hace la función de identificador de un determinado mensaje.

- **Responde a idMensaje:** Código alfanumérico que hace la función de identificador del mensaje al que está respondiendo. Este campo estará vacío en caso de que se esté iniciando el hilo.
- **Source:** Código alfanumérico que hace la función de identificador del autor del mensaje. Es importante considerar que este autor puede ser un estudiante o, incluso, un miembro del equipo docente.
- **Dia:** Nombre del día de la semana en el que el mensaje fue enviado.
- **Fecha:** Fecha en la que el mensaje es enviado. Para identificar la fecha se emplea el formato "DD/MM/YYYY".
- **Hora:** Hora en la que el mensaje es enviado. Para identificar la hora se emplea el formato "hh/mm/ss".
- **Titulo mensaje:** Título dado a un mensaje concreto.
- **Mensaje:** Cuerpo del mensaje.
- **Caracteres mensaje:** Número total de caracteres por los que está formado el cuerpo de un mensaje.

Para el estudio se han usado varios *datasets* de distintos tamaños para probar que el proceso funciona, independientemente del número de usuarios y de interacciones que haya entre estos. Para esta memoria se realizará el proceso con dos *datasets* de diferente tamaño:

- **Conjunto de datos Alfa:** Referente al *dataset* de menor tamaño, compuesto por 229 mensajes.
- **Conjunto de datos Beta:** Referente al *dataset* de mayor tamaño, compuesto por 527 mensajes.

3.3 Preprocesamiento

Esta sección se dividirá en dos fases, para mejorar la comprensión de la misma. En la primera, se describirán las operaciones realizadas en los *datasets* antes de que estos sean representados como grafos. A continuación, se especificaran que medidas de la red se realizarán sobre dicho grafo para su uso en el procesamiento posterior.

Fase I

A partir de los parámetros de entrada descritos en la anterior sección, en esta fase realizaremos determinadas operaciones de preprocesamiento sobre dichos datos, previo a ser representados como grafos. El preprocesamiento "*se encarga de la limpieza de datos, su integración, transformación y reducción para la siguiente fase de minería de datos*" (García et al., 2016). Este paso es esencial en el procedimiento debido a que el uso de datos de baja calidad, por una falta de técnicas de preprocesamiento, implica, por lo general, un proceso de minería de datos con pobres resultados. Después de aplicarse estas técnicas, el conjunto resultante "*puede verse visto como una fuente consistente y adecuada de datos de calidad para la aplicación de algoritmos de minería de datos*" (García et al., 2016).

Para medir la calidad de los datos, hay 4 casos que hay que tener en cuenta (Anaya, 2019c):

- **El ruido es una distorsión de los datos:** Este "ruido" se debe eliminar si no se quiere que los algoritmos posteriores sean afectados.
- **Valores erráticos o atípicos:** Son instancias consideradas diferentes de otras instancias del conjunto de datos.
- **Valores perdidos:** Algunas de las instancias presentan atributos sin valor, es decir, los datos están incompletos.
- **Datos duplicados:** Ocurre cuando distintas instancias tienen el mismo valor en todos sus atributos.

Para el escenario que nos ocupa, nos centraremos en el primer y tercer caso (al no presentar los *datasets* proporcionados ni valores erráticos ni duplicados).

Es importante tener en cuenta que, para el primer caso, aunque los datos requieren de un extensivo preprocesamiento para la eliminación del ruido, la definición de este "ruido" es relativa a la tarea de procesamiento posterior y un preprocesamiento estricto podría eliminar una gran cantidad de datos. Es decir, el objetivo es eliminar la suficiente cantidad de ruido para que la calidad de los datos mejore, pero, sin que esto conlleve a una perdida significativa de información.

El primer proceso que se realizará será agregar un nuevo atributo a los *datasets* a partir de los datos de dos de las otras columnas (*Responde a idMensaje* y *source*):

- **Target:** Código alfanumérico que funciona como identificador del autor del mensaje al que se está respondiendo. Al igual que con el atributo *Responde a idMensaje*, este campo

podría aparecer vacío en caso de que se esté iniciando el hilo. Además, como sucedía con *source*, el autor puede ser un estudiante o un miembro del equipo docente.

Este atributo identificará al usuario destinatario de cada mensaje lo que nos permitirá, después, poder generar los respectivos grafos. Sin embargo, antes de eso, realizaremos un par más de operaciones sobre el conjunto de *datasets*:

- **Eliminación de mensajes sin destinatario:** Debido a que nuestro objetivo principal es medir y evaluar las interacciones entre dos estudiantes distintos, aquellos mensajes que carezcan de destinatario no tienen, por lo tanto, ningún tipo de interés para nuestro estudio. Este proceso nos permite, además, deshacernos de todos aquellos hilos que no han recibido ningún tipo de respuesta.

  En última instancia, este método no modificará el grafo resultante (ni en su número de nodos, ni de aristas), pero nos permitirá reducir considerablemente la potencia computacional necesaria en caso de, a futuro, usar *datasets* de un tamaño mayor a los que empleamos aquí. Por ejemplo, en el caso del *dataset* Beta, emplear este procedimiento reduce el número de instancias casi un 25% (de 527 a 396 mensajes).

- **Eliminación de auto-respuestas:** Tal y como definimos en el punto anterior, el objetivo principal es medir y evaluar las interacciones entre dos usuarios distintos. Aquellos mensajes en los que el emisor y el destinatario son el mismo estudiante, tampoco nos es de interés para el estudio. Este proceso, además de permitirnos reducir la potencia computacional (evitando que el sistema evalúe el intercambio de información de un usuario consigo mismo), elimina la posibilidad (poco probable) de que se recomienden duplas mentor-estudiante formadas en ambos lados por el mismo usuario.

Además de estas dos operaciones, se consideró y probó una tercera que finalmente se terminó por descartar:

- **Eliminación del docente:** Aunque nuestro objetivo final es que el sistema recomiende duplas de estudiantes, eliminar al docente del grafo, para que este no sea considerado, provocaría que muchos de los estudiantes quedaran aislados y los resultados distaran de la realidad.

  La parte positiva es que, tal y como se explicará en la sección de "Procesamiento", el docente es un usuario con unos valores tan altos en las diferentes medidas de la red consideradas que se reserva un cluster exclusivamente para dicho usuario. Por lo tanto, bastaría con no considerarlo una vez realizada la clasificación.

Tras esto, ahora sí, procederemos a generar el grafo. Un grafo, de manera simplificada, se puede definir como una colección de vértices (nodos) y sus conexiones entre ellos (aristas), que pueden ser dirigidas o no-dirigidas. Para este escenario, los nodos son usuarios y las aristas (dirigidas) los mensajes entre ellos, donde el origen representa al remitente (*source*) y el final, al destinatario (*target*). Para la generación de este grafo emplearemos la herramienta de Gephi, anteriormente presentada.

Los grafos de ambos *datasets*, tras esta primera fase de preprocesamiento, quedarían tal y como se muestran en la Figura 7 (para Alfa) y en la Figura 8 (para Beta):

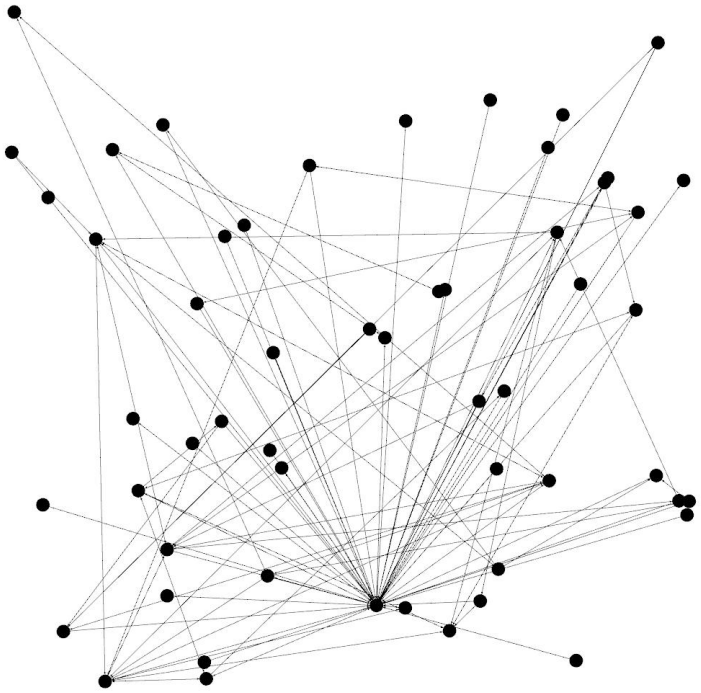

Figura 7: Grafo inicial *dataset* Alfa

Fase II

En esta fase, una vez el grafo ya ha sido generado, se especificarán y calcularán las diferentes medidas de la red que serán empleadas, posteriormente, en el procesamiento por parte del algoritmo.

Según (Saqr et al., 2018) el análisis de redes sociales, junto con indicadores cuantitativos, tienen el potencial de ofrecer ideas tanto sobre la cantidad como sobre la calidad de la colaboración, así como el papel de los colaboradores. En cuanto a nivel individual, encontramos 3 constructos esenciales:

- **El nivel de actividad (parámetros de participación cuantitativos):** El nivel de actividad de un colaborador se puede medir mediante 3 medidas de centralidad:

- **Grado de salida (*out-degree*):** Indica la cantidad de interacciones de los participantes y puede interpretarse como una medida de sociabilidad.

- **Grado de entrada (in-degree):** Es el número de interacciones que recibe un participante y mide su influencia. Un usuario, por lo general, recibe una interacción cuando, por ejemplo, aporta conocimiento más allá de lo que han aportado otros o un punto de vista que merece ser discutido. En contextos de intercambio de conocimientos, un valor más alto en esta medida puede considerarse como un signo de experiencia, popularidad o liderazgo.

- **Centralidad de grado:** Esta medida engloba el número total de interacciones a las que un usuario ha contribuido (grado de salida) o ha recibido (grado de entrada). La centralidad de grado es otro indicador de interactividad que tiene en cuenta ambos sentidos de las interacciones.

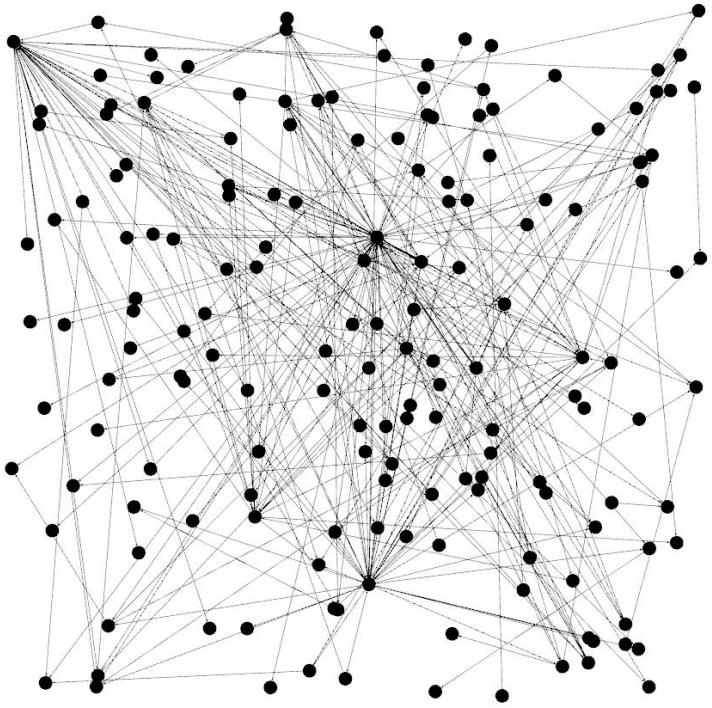

Figura 8: Grafo inicial *dataset* Beta

- **La posición en el intercambio de información:** El rol en la transferencia de la información se puede medir haciendo uso de 3 medidas:

  - **Centralidad de intermediación (*betweenness centrality*):** Esta centralidad cuantifica la frecuencia o el número de veces que un nodo actúa como un puente a lo largo del camino más corto entre otros dos nodos. A nivel social, hace referencia al número de veces que un participante actuó como coordinador en las interacciones entre colaboradores que, de otra manera, no estarían conectados.

Un valor alto en esta medida se traduce en un signo de creatividad y prominencia en la difusión de información e ideas en una red; mientras que un valor bajo, es un signo de dificultad o indiferencia para comunicarse con otros miembros del grupo sin un intermediario o mediador.

- **Centralidad de cercanía (closeness centrality):** El grado de que un individuo está cerca de todos los otros individuos en una red (directamente o indirectamente), es decir, como de fácil es alcanzarlos e interactuar con ellos. Valores más altos en esta medida reflejan una posición accesible en el intercambio de información y, valores más bajos, pueden ser vistos como un signo de aislamiento social y una comunicación deficiente.

- **Centralidad de la información (*information centrality*):** Esta medida engloba la cantidad de información que fluye a través de un participante en una red social. Tener una posición a través de la cual fluye la información es un activo privilegiado durante el intercambio de información.

- **El rol en el grupo:** Según (Marcos-García et al., 2015) los estudiantes pueden clasificarse en distintos roles según su nivel de actividad: líder (*leader*), coordinador (*coordinator*), animador (*animator*), activo (*active)*, periférico (*peripheral*), silencioso (*quiet*) y ausente (*missing*). Dado nuestro escenario, el objetivo es encontrar cual de estos roles (y, por consecuente, que usuarios pertenecientes a estos) serían más apropiados para ser asignados a cada uno de los perfiles de la dupla mentor-estudiante (Saqr et al., 2018):

  - **Mentor:** Para este perfil es imprescindible que el usuario posea un alto nivel de actividad (altos valores de grado de salida, grado de entrada y centralidad de grado). Además, este debe poseer una buena posición en la transferencia de la información (que se traduce en valores altos de centralidad de intermediación y centralidad de cercanía).

    Debemos, por lo tanto, buscar a aquellos usuarios que destaquen por encima de todos los demás, es decir, para este perfil nos centraremos en aquellos usuarios que posean un rol de líder.

    Sería aconsejable, también, considerar valores altos de *pagerank*. Esta medida permite mitigar el inconveniente que surge cuando, ante la presencia de un nodo de autoridad, sus vecinos pasan a tener, automáticamente, una gran importancia. En nuestro escenario, esto puede resultar un problema aún mayor debido a que tenemos un usuario que tendrá una autoridad mucho más alta que la del resto, el docente. Por lo tanto, esta medida nos permitirá suavizar este problema, dividiendo el valor de la

centralidad por el número de aristas salientes al nodo, de tal modo que, para un nodo conectado, se obtiene una fracción de la centralidad del nodo fuente.

- **Alumno:** Para este perfil, el usuario debe presentar un bajo nivel de actividad (bajo grado de entrada), así como un rol limitado en el intercambio de información (valores bajos de centralidad de intermediación y centralidad de cercanía). Sin embargo, esperamos que el usuario sea algo participativo (bajo grado de salida y centralidad de grado), es decir, no consideraremos usuarios con roles silenciosos o ausentes debido a que seria imposible, posteriormente, evaluar la comunicación de estos con otros tipos de usuario al ser esta inexistente.

  Con esto en cuenta, buscaremos a aquellos usuarios que, a diferencia del caso anterior, destaquen negativamente por debajo del resto pero que tengan cierto nivel de participación. Por ello, nos centraremos en aquellos usuarios que posean un rol de periférico.

Id	In-Degree	Out-Degree	Degree	Eccentricity	Closeness Centrality	Betweenness Centrality	PageRank
039d7061d552c18dfd67...	1	2	3	5.0	0.386667	0.000815	0.006093
0646079ddd6754d787db...	1	1	2	5.0	0.371795	0.0	0.009103
0a8755c98b57a70b4aa6...	0	3	3	5.0	0.336957	0.0	0.002727
0aad920a63c8e397ae53...	8	8	16	3.0	0.537037	0.05902	0.063008
0c95e4a48870ebbba540...	2	2	4	4.0	0.408451	0.0	0.026632
186218d358c431d7fe74...	0	1	1	5.0	0.37037	0.0	0.002727
1899c64c739a9679355f...	0	1	1	5.0	0.37037	0.0	0.002727
1b16defcfde3f6516a197...	1	1	2	5.0	0.37037	0.001724	0.0035
1cd12dffc337d28f9dc07...	0	1	1	5.0	0.37037	0.0	0.002727
1d2d912a30306ac87409...	0	1	1	5.0	0.37037	0.0	0.002727
1ed75217e05b580a3e51...	3	4	7	4.0	0.453125	0.050116	0.029998
1f90e256295523248cba...	0	1	1	6.0	0.277778	0.0	0.002727
208b3f090bea349b7e1d...	0	1	1	5.0	0.37037	0.0	0.002727
24fb77d6ef839d751679f...	6	4	10	4.0	0.460317	0.010966	0.029463
28be17e1cd3b2aebba5a...	0	1	1	5.0	0.37037	0.0	0.002727
2d72d85663bdcc41821c...	0	1	1	5.0	0.37037	0.0	0.002727
33b8cbadae67bffcc02cf...	0	1	1	5.0	0.37037	0.0	0.002727
37ac418bf9558984fbcda...	2	2	4	5.0	0.371795	0.000175	0.025711
39bf7786bc5a019e2c6d...	0	1	1	5.0	0.37037	0.0	0.002727
3ad6f72cd9b5208dfe50d...	6	6	12	4.0	0.5	0.077632	0.033941
3b56b4a31259250e60e6...	2	4	6	4.0	0.426471	0.008473	0.027177
3c01319490457214c466...	0	1	1	5.0	0.37037	0.0	0.002727
45c2fa9a13586c13ee4ef...	3	1	4	5.0	0.367089	0.001724	0.03164

Figura 9: Valores medidas (*dataset* Alfa)

Teniendo en cuenta estos 3 constructos y, especialmente, considerando los roles tomados para cada uno de los perfiles, calcularemos (haciendo uso, de nuevo, de la herramienta de Gephi) las siguientes medidas de red de ambos grafos obtenidos en la primera fase de esta sección para su uso en el posterior procesamiento:

- Centralidad de grado (general, grado de salida y grado de entrada).
- Centralidad de intermediación.
- Centralidad de cercanía.

- *PageRank*.

Realizadas las medidas, nos quedaría una tabla como la que aparece en la Figura 9.

Tras esto, y justo antes de proceder con el procesamiento, se debería realizar, además, una normalización estadística de los valores de las medidas realizadas. La normalización "*ocurre cuando los posibles rangos de los valores para cada atributo toman un valor entre un determinado intervalo, siendo este intervalo el mismo para cada atributo*"(Anaya, 2019c). Es muy común que los atributos de cada instancia de los *datasets* tengan un rango de valores muy variado y, para algunos algoritmos de aprendizaje automático (como el que usaremos en este caso) conviene normalizar los valores para que los rangos de los atributos sean compatibles.

La tabla, una vez normalizados las medidas dentro del rengo [0...1], nos quedaría como a continuación:

Id	indegree	outdegree	degree	eccentricity	closnesscentrality	betweenesscentrality	pageranks
039d7061d552c18dfd6763d...	0.020408	0.076923	0.032258	0.666667	0.374383	0.001829	0.01199
0646079ddd6754d787db744...	0.020408	0.0	0.016129	0.666667	0.32325	0.0	0.022712
0a8755c98b57a70b4aa6154...	0.0	0.153846	0.032258	0.666667	0.20347	0.0	0.0
0aad920a63c8e397ae53ee3...	0.163265	0.538462	0.241935	0.0	0.891387	0.132419	0.214729
0c95e4a48870ebbba5408ae...	0.040816	0.076923	0.048387	0.333333	0.449281	0.0	0.085153
186218d358c431d7fe7481d...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
1899c64c739a9679355f886...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
1b16defcfde3f6516a197afd6...	0.020408	0.0	0.016129	0.666667	0.318351	0.003868	0.002754
1cd12dffc337d28f9dc077cf5...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
1d2d912a30306ac87409504...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
1ed75217e05b580a3e51b67...	0.061224	0.230769	0.096774	0.333333	0.60288	0.112441	0.097143
1f90e256295523248cba662...	0.0	0.0	0.0	1.0	0.0	0.0	0.0
208b3f090bea349b7e1d091...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
24fb77d6ef839d751679f9c2...	0.122449	0.230769	0.145161	0.333333	0.627607	0.024604	0.095237
28be17e1cd3b2aebba5a390...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
2d72d85663bdcc41821c60a...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
33b8cbadae67bffcc02cfd43...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
37ac418bf9558984fbcdae5f...	0.040816	0.076923	0.048387	0.666667	0.32325	0.000393	0.081872
39bf7786bc5a019e2c6d64c...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
3ad6f72cd9b5208dfe50d2b8...	0.122449	0.384615	0.177419	0.333333	0.764046	0.174177	0.111188
3b56b4a31259250e60e63a9...	0.040816	0.230769	0.080645	0.333333	0.511238	0.01901	0.087094
3c01319490457214c466853...	0.0	0.0	0.0	0.666667	0.318351	0.0	0.0
45c2fa9a13586c13ee4eff16...	0.061224	0.0	0.048387	0.666667	0.30707	0.003868	0.102992
4da45a54f500eb6d2b7e721...	0.020408	0.076923	0.032258	0.333333	0.511238	0.0	0.008439
53716f5d75b4da1aa1fd6a41...	0.020408	0.076923	0.032258	0.666667	0.410804	0.0	0.021601
73114c6ea4639b6149f70c1...	0.061224	0.307692	0.112903	0.666667	0.511238	0.047376	0.060841
7486f5cb94460937b718542...	0.061224	0.153846	0.080645	0.666667	0.429776	0.005331	0.051512
7c95be4c421ca601382baf8...	0.040816	0.0	0.032258	0.666667	0.30707	0.0	0.095006

Figura 10: Valores normalizados (*dataset* Alfa)

3.4 Procesamiento

En esta sección se describirá, en detalle, el procesamiento seguido. Para ello la dividiremos en 3 bloques, claramente diferenciados según las diferentes fases de este procedimiento.

Clasificación de alumnos

Como ya se especificó en la anterior sección, nuestro primer paso en esta fase será clasificar los alumnos según el rol al que pertenezcan. Para ello emplearemos un algoritmo de aprendizaje no supervisado, concretamente un algoritmo de clustering.

Los algoritmos de aprendizaje no supervisado nos permite poder agrupar en clusters conjuntos de datos sin etiquetar, por medio del descubrimiento de patrones ocultos, sin necesidad de intervención humana. Este tipo de algoritmos es ideal para nuestro escenario por los siguientes motivos:

- **Naturaleza de los datos:** La propia naturaleza desestructurada de los datos, del tipo de los que normalmente se manejan en una web social (como los *datasets* usados aquí, extraídos directamente de un foro universitario), hace que etiquetar estos sea una tarea prácticamente imposible. El aprendizaje no supervisado nos permite realizar tareas más complejas sobre un conjunto de datos en crudo para convertirlos en información de utilidad.

- **Objetivos de la solución:** Mientras que los algoritmos de aprendizaje supervisado, generalmente, se usan para clasificar datos o hacer predicciones (haciendo uso de unos datos de entrenamiento), los algoritmos no supervisados se emplean para comprender las relaciones dentro de los conjuntos de datos.

  En nuestro escenario nuestro objetivo es, a partir del conjunto de mensajes en los *datasets*, entender las relaciones existentes entre los distintos usuarios y detectar los roles que tienen cada uno de estos según la naturaleza de estas relaciones. Es, por lo tanto, imprescindible el uso de algoritmos de aprendizaje no supervisado para alcanzar nuestro objetivo.

Sin embargo, debemos considerar también que este tipo de algoritmos, al no disponer de un conjunto de entrenamiento, son bastante más complejos (lo que afecta a potencia computacional necesaria) y los resultados pueden ser inexactos, siempre y cuando no haya intervención humana de expertos que realicen una validación de las variables de salida.

La clasificación de los usuarios se realizará con cada perfil por separado, ejecutando el algoritmo de clustering en dos ocasiones usando características o atributos diferentes según el perfil de usuario que queramos detectar. Para cada uno de estos perfiles se buscará los "extremos", es decir, los usuarios con los valores más altos y más bajos en sus respectivos atributos; aunque se tendrán en consideración algunas excepciones que se especificarán más adelante.

Para esta clasificación se probarán dos algoritmos y se comprobará cuál de ellos es más eficiente para el problema que se presenta, el algoritmo de esperanza-maximación (EM) y el

algoritmo de *k-means*. Ambos son algoritmos de agrupamiento no supervisado que tienen como objetivo agrupar puntos de datos similares juntos y ambos implican un algoritmo iterativo que busca optimizar un objetivo de agrupamiento (FC, 2021). Sin embargo, aunque ambos busquen el mismo objetivo, son dos acercamientos distintos con ciertas diferencias.

EM

Comenzaremos por el algoritmo de EM que es un tipo de algoritmo de agrupamiento probabilístico que asume que los puntos de datos son generados por una mezcla de varias distribuciones Gaussianas. El algoritmo estima de manera iterativa los parámetros desconocidos de las distribuciones Gaussianas y las probabilidades de asignación de cada punto de datos a cada distribución. El algoritmo converge cuando la verosimilitud de los datos bajo los parámetros del modelo estimado deja de aumentar significativamente (FC, 2021). Esto dará como resultado grupos compactos, es decir, cada instancia solo puede pertenecer a un cluster. Esto es lo que nos interesa puesto que nuestro objetivo es agrupar los miembros del *dataset* en diversos grupos para identificar distintos perfiles de usuarios.

Para este proceso de clasificación de usuarios en los diferentes roles, líder o periférico, se empleará, en este estudio, la herramienta de Weka. Primero, detectaremos los usuarios "líderes" para cada uno de los *datasets* usando los atributos de centralidad de grado, grado de salida, grado de entrada, centralidad de intermediación, centralidad de cercanía y *pagerank* (de acuerdo al análisis realizado en la sección de "Preprocesamiento").

Los usuarios que tomarán el perfil de mentor deben tener valores altos en todos los atributos indicados. Si nos fijamos en el caso del estudio realizado sobre el *dataset* Alfa en la Figura 11, vemos como el cluster con los valores más altos es el número 4. Este cluster contiene, de hecho, un solo usuario cuyos valores son mucho más altos que los del resto, este se trata del docente. Dado que nuestro objetivo, en última instancia, es analizar a los estudiantes, lo descartaremos.

El siguiente cluster con los valores más altos sería el número 5. Los 3 usuarios que componen este cluster presentan, en los diferentes atributos, unos valores bastante superiores a los del resto. Por lo tanto, son buenos candidatos de "lideres" y, consecuentemente, de posibles mentores.

```
EM
==

Number of clusters: 6
Number of iterations performed: 2


                        Cluster
Attribute                    0      1      2      3      4      5
                          (0.09) (0.44) (0.15) (0.25) (0.02) (0.05)
=============================================================================
indegree
  mean                   0.0082 0.0047 0.0714 0.0347      1 0.1293
  std. dev.              0.0163 0.0092  0.027 0.0212 0.1363 0.0255

outdegree
  mean                   0.0463      0   0.25 0.0722      1 0.4359
  std. dev.              0.0614 0.0001 0.0334 0.0453 0.1757 0.0725

Degree
  mean                   0.0162 0.0037 0.1089 0.0425      1 0.1935
  std. dev.              0.0144 0.0073 0.0225  0.016 0.1401 0.0348

closnesscentrality
  mean                   0.1455  0.317 0.5663 0.4088      1 0.7695
  std. dev.              0.1386 0.0041 0.0844 0.0854 0.1808 0.0973

betweenesscentrality
  mean                   0.0001 0.0001 0.0632 0.0043      1 0.1491
  std. dev.              0.0002 0.0007 0.0397 0.0078 0.1387  0.018

pageranks
  mean                   0.0053 0.0055 0.0257 0.0146 0.2835 0.0493
  std. dev.              0.0052 0.0064 0.0069 0.0097 0.0389 0.0119


Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       7 ( 13%)
1      24 ( 44%)
2       8 ( 15%)
3      12 ( 22%)
4       1 (  2%)
5       3 (  5%)


Log likelihood: 19.16573
```

Figura 11: Resultados cluster en *dataset* Alfa con algoritmo EM (líder)

Para el escenario Beta que aparece en la Figura 12, observamos un escenario muy parecido al que teníamos en Alfa por lo que seguiremos el mismo proceso. Descartamos el cluster 5 (por ser el de los docentes) y tomaremos el número 2 (formado por 12 usuarios) por ser el que tiene los valores más altos dentro del conjunto de estudiantes.

Para los usuarios periféricos, por el contrario, emplearemos los atributos de grado de entrada, grado de salida, centralidad de grado, centralidad de intermediación y centralidad de cercanía (de acuerdo, de nuevo, al análisis realizado en la sección de “Preprocesamiento”).

```
EM
==

Number of clusters: 6
Number of iterations performed: 1


                     Cluster
Attribute                 0      1      2      3      4      5
                       (0.25) (0.35) (0.07) (0.22) (0.09) (0.02)
=============================================================================
indegree
  mean                0.0122 0.0676 0.3092 0.0782 0.0194   0.71
  std. dev.           0.0195 0.0381 0.0965 0.0343 0.0354 0.1759

outdegree
  mean                0.0202 0.0003 0.1065  0.034 0.0202  0.607
  std. dev.           0.0049 0.0023 0.0802  0.015 0.0048 0.2679

Degree
  mean                0.0047 0.0076 0.1568 0.0347 0.0069 0.6333
  std. dev.           0.0064 0.0117 0.0602 0.0174 0.0119 0.2313

closnesscentrality
  mean                0.2687 0.0074 0.3832 0.3289 0.9665 0.4806
  std. dev.           0.0385 0.0673 0.0417 0.0441 0.1003 0.0684

betweenesscentrality
  mean                0.0004 0.0002 0.1231 0.0165 0.0008 0.6554
  std. dev.           0.0011 0.0019 0.0695 0.0135  0.003 0.2032

pageranks
  mean                0.0026 0.0046 0.0163 0.0051 0.0031 0.0478
  std. dev.           0.0006 0.0017  0.006 0.0018 0.0021 0.0174


Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       46 ( 28%)
1       57 ( 34%)
2       12 (  7%)
3       33 ( 20%)
4       15 (  9%)
5        4 (  2%)


Log likelihood: 18.1483
```

Figura 12: Resultados cluster en *dataset* Beta con algoritmo EM (líder)

Los usuarios que tomarán el perfil de estudiante deben tener valores bajos en todos los atributos indicados. Si nos fijamos en el caso del estudio realizado sobre el *dataset* Alfa en la Figura 13, vemos como el cluster con los valores más bajos es el número 1. Sin embargo, tal y como comentamos en la anterior sección, buscamos usuarios que sean algo participativos y, como se ve en el atributo de grado de salida, este es un requisito que no se cumple para este cluster.

El siguiente cluster con los valores más bajos sería el cluster 0 formado por 6 usuarios que, aunque algo más participativos que los del anterior cluster, continúan presentando unos valores

lo suficientemente bajos como para considerarlos unos buenos candidatos al rol de "*periféricos*" y, con ello, al perfil de estudiante.

Para el escenario Beta que aparece en la Figura 14, observamos un escenario muy parecido al que teníamos en Alfa por lo que seguiremos el mismo proceso. Descartamos el cluster 0 (por no ser lo suficientemente participativos) y tomaremos el número 4 (formado por 41 usuarios) por ser el que tiene los valores en global más bajos.

```
EM
==

Number of clusters: 6
Number of iterations performed: 2


                     Cluster
Attribute                 0       1       2       3       4       5
                       (0.09)  (0.46)  (0.15)  (0.23)  (0.02)  (0.05)
=============================================================================
indegree
  mean                  0.009   0.007  0.0714  0.0325       1  0.1293
  std. dev.            0.0168  0.0135   0.027  0.0218  0.1363  0.0255

outdegree
  mean                 0.0438  0.0001    0.25  0.0805       1  0.4359
  std. dev.            0.0625  0.0026  0.0334    0.04  0.1757  0.0725

Degree
  mean                 0.0163  0.0055  0.1089  0.0426       1  0.1935
  std. dev.             0.015  0.0107  0.0225  0.0167  0.1401  0.0348

closnesscentrality
  mean                 0.1341  0.3165  0.5663    0.42       1  0.7695
  std. dev.            0.1295  0.0045  0.0844  0.0818  0.1808  0.0973

betweenesscentrality
  mean                 0.0001  0.0002  0.0633  0.0045       1  0.1491
  std. dev.            0.0002  0.0009  0.0397  0.0081  0.1387   0.018


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 11%)
1      25 ( 45%)
2       8 ( 15%)
3      12 ( 22%)
4       1 (  2%)
5       3 (  5%)


Log likelihood: 14.17033
```

Figura 13: Resultados cluster en *dataset* Alfa con algoritmo EM (periférico)

```
EM
==

Number of clusters: 6
Number of iterations performed: 1


                        Cluster
Attribute                    0       1       2       3       4       5
                          (0.34)  (0.12)  (0.17)  (0.02)  (0.28)  (0.07)
=============================================================================
indegree
  mean                   0.0671  0.0207   0.084   0.711  0.0217  0.3085
  std. dev.              0.0371  0.0326  0.0382  0.1751  0.0293  0.0973

outdegree
  mean                   0.0002  0.0199  0.0365  0.6078  0.0214  0.1067
  std. dev.              0.0023  0.0042  0.0157  0.2678  0.0065  0.0803

Degree
  mean                   0.0074   0.007  0.0382  0.6342  0.0085  0.1567
  std. dev.              0.0114  0.0109  0.0184  0.2309  0.0105  0.0604

closnesscentrality
  mean                   0.0061  0.7993  0.3428  0.4808   0.269  0.3836
  std. dev.              0.0601  0.2987  0.0713  0.0683  0.0323  0.0413

betweenesscentrality
  mean                   0.0002  0.0003  0.0183  0.6564  0.0026  0.1231
  std. dev.              0.0022  0.0007  0.0138  0.2027  0.0064  0.0697


Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       57 ( 34%)
1       22 ( 13%)
2       31 ( 19%)
3        4 (  2%)
4       41 ( 25%)
5       12 (  7%)


Log likelihood: 12.72571
```

Figura 14: Resultados cluster en *dataset* Beta con algoritmo EM (periférico)

Por lo tanto, al ejecutar el algoritmo nos quedaría la siguiente distribución en los dos *datasets* estudiados:

- **Alfa:** 3 mentores y 6 estudiantes.

Perfil	ID Autor	Nombre
Mentor	0aad920a63c8e397ae53ee34c0ae076a1561dbd0	Estu-155
	a8b41b15be0686c2d494f90641de310474f9910d	Estu-49

	3ad6f72cd9b5208dfe50d2b8634fa5f6132e7645	Estu-82
Estudiante	1f90e256295523248cba6621e730126fc3c05327	Estu-92
	87233b9bba88511662174e16b6d7b86cc313e8cb	Estu-19
	841ae72cf8f7a036743ef302697fdc1205a5c7c1	Estu-110
	0a8755c98b57a70b4aa6154e740e55fe9f52c060	Estu-24
	acc09481ce1151147976e69fcd209f4a7b8184f0	Coor-36
	809aae900144ebff84da6ad36445a9dd3bb4072a	Estu-20

Tabla 2: Distribución mentor/alumno en escenario Alfa

Podemos observar que uno de los usuarios con rol periférico no es un estudiante sino un coordinador y, por lo tanto, será descartado para posteriores análisis.

- **Beta:** 12 mentores y 41 estudiantes.

Perfil	ID Autor	Nombre
Mentor	2dc78a0929f266496f96144c29f3cd22804b5c1e	Estu-1
	74cf09e81e75ec98dc1212d933737417656536ec	Estu-2
	2cab7655699fa464b94aba35dc4a5c057b23899d	Estu-3
	9c594aa7e30bfbe0b219bf433e79d1a247290cf8	Estu-4
	71736b0e5b9943903a8a33dac5a19501727debf9	Estu-5
	39c2d1e0b5a9e388ecda28eb1b778a7c01a911e4	Estu-6
	bd727f3c7a104eab070f398b2e10a8da2b207221	Estu-7
	c4e8c05baf64ab276357f51e42b0e71268678e2f	Estu-8
	9c4743ee0633578b10a0535a1e0f6fb771adc7c6	Estu-9
	63bb9b5f521460377bb06a65c28a366042911bcd	Estu-10
	9e4817c1eaacbf9f65ad4a952c3bb15d41efb1e3	Estu-11
	48e6baec8283839a354eff890d5816a4a446a0fe	Estu-12
Estudiante	2f4ce8a0803c214d4e6a307f29f6b75b66fcec7b	Estu-13

	75ad6b19f6a25a046f62b046ae7364483a1a63de	Estu-14
	5c74c84de1eee4f5dc178288dc44eaea7d5a7de1	Estu-15
	8c133465b34d69180da5f54bc64922fc9ca7b5c5	Estu-16
	7f496c216f69aca021c80b060603e62ff2973faa	Estu-17
	1db2ab0329523f70c1a53cdc1f67363fa88ae56c	Estu-18
	ece3e9a281905a24b560b6ad712bb99d32178671	Estu-19
	599fb05778bcacf06fe82e06dafdc76b97e11c32	Estu-20
	db531d817355304f9bfa16221c933cfd3e8921b9	Estu-21
	ccbb88c407120f84bb295e91592d27fab2df6d84	Estu-22
	85cbb1bde5f80b1c7b7a9d207a64a48de7cdb513	Estu-23
	db8d0ece288fe3968822e92546f642ab55ba657d	Estu-24
	3ffee227aeded42187685d13965e56f0d915af31	Estu-25
	0ea07c4c210e890117a4e452a2af75af27100166	Estu-26
	c0c16eefd5cb6e12ab9910cf105d51041398cc74	Estu-27
	7d2cc0fac12ccacf7c892a227a5057bf4e383433	Estu-28
	0b292695c44c0b1f7bd186ea40c596ac90be4c18	Estu-29
	7625518550c235a10f9dc2e24594e1288277222a	Estu-30
	db296ebbfd3c7bc1ed5fbf5496d360b1e15eb4fd	Estu-31
	10531de53288f03304548028b249a7408cf813c7	Estu-32
	c07e3916668fe448dfa49457094bd24fc9514f06	Estu-33
	b263a7c7465683b54166779825c7ad3e3fef96f3	Estu-34
	c78b9eced612629c62580ea493e8a3dea05b5948	Estu-35
	1e6b93d288eb05428a9a13584d764299828647af	Estu-36
	6bc97b5916acb9f1d653a59520b45c07dd03ef5b	Estu-37
	b69837550c52660f74f279876b64862b30c2e7ea	Estu-38

	05e5f5dacbf171c58ef75b619d9a1149ffa5e01d	Estu-39
	63c8192913c6167f6143f7a97ff223f93698d8fb	Estu-40
	6234076827de5c00590346e0450b11c4db33c378	Estu-41
	489b5a1b7c473dadbaf5015d69b180f45b7a4664	Estu-42
	a1392c54e1670ffd703b218435660f9abc02d8cb	Estu-43
	15ff2ed69bbd46111342716f20294fbcc1e83f42	Estu-44
	68a2bcc80194c975615b3818f19dc75211250d39	Estu-45
	52a386b7e22b2ef80ba1c58693f2082f0203f975	Estu-46
	3ea6c3a689c40ecc24f556345445d3746e7486d1	Estu-47
	0766d083371ef85477d06309807a55a7c314ec90	Estu-48
	c780e04e4046b10c634497cd94ba211a3a9b6cb1	Estu-49
	f1ee2b6acbd6dbc1e8dee75d4097661587ee956e	Estu-50
	b6c7341adc243743fc13df67d1eae45b24dd5ec2	Estu-51
	f4601992af376bd5638f5b909f1fd91d0a899ec1	Estu-52
	dcb531fccb2dd980aa43834df7ee5fb7690b8fda	Estu-53

Tabla 3: Distribución mentor/alumno en escenario Beta

El resto de usuarios (compuestos por los docentes, estudiantes con baja o nula participación, y aquellos que se encuentran con valores entre los dos perfiles) serán descartados al no ser de interés para nuestro objetivo.

K-means

K-means es un algoritmo de agrupamiento basado en centroides que divide los puntos de datos en K clusters al minimizar la suma de las distancias al cuadrado entre cada punto de datos y su centro de cluster asignado. El algoritmo inicializa de manera aleatoria K centros de cluster y actualiza repetidamente las asignaciones de cluster y los centros de cluster hasta que se alcance la convergencia (FC, 2021).

Para este algoritmo se empleará el mismo procedimiento y los mismos atributos que en el escenario con EM para evaluar ambos métodos en las mismas condiciones. Empecemos, pues, por la búsqueda de *líderes*.

```
Initial starting points (random):

Cluster 0: 0,0,0,0.318351,0,0.002727
Cluster 1: 0.040816,0.230769,0.080645,0.6795,0.125142,0.024745
Cluster 2: 0.020408,0.076923,0.032258,0.533122,0,0.007987
Cluster 3: 0,0,0,0,0,0.002727
Cluster 4: 0.061224,0,0.048387,0.30707,0.003868,0.03164
Cluster 5: 0.020408,0.076923,0.032258,0.410804,0,0.008791

Missing values globally replaced with mean/mode

Final cluster centroids:
                               Cluster#
Attribute               Full Data          0          1          2          3          4          5
                           (55.0)     (23.0)      (1.0)      (9.0)      (2.0)     (10.0)     (10.0)
==========================================================================================================
indegree                    0.0471     0.0027          1     0.0907          0     0.0367     0.0347
outdegree                   0.1007     0.0067          1     0.3162          0     0.0308     0.1231
Degree                      0.0584     0.0035          1      0.138          0     0.0355     0.0532
closnesscentrality          0.3978     0.3224          1     0.6595          0     0.2861     0.4666
betweenesscentrality        0.0367     0.0002          1     0.1018          0      0.005     0.0044
pageranks                   0.0182     0.0032     0.2835     0.0332     0.0027     0.0203     0.0136



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      23 ( 42%)
1       1 (  2%)
2       9 ( 16%)
3       2 (  4%)
4      10 ( 18%)
5      10 ( 18%)
```

Figura 15: Resultados cluster en *dataset* Alfa con algoritmo *k-means* (líder)

```
Initial starting points (random):

Cluster 0: 0,0.018868,0,0.364985,0,0.002243
Cluster 1: 0.086957,0.018868,0.026667,0.364179,0.02223,0.005107
Cluster 2: 0,0.018868,0,0.284722,0,0.002243
Cluster 3: 0.695652,0.415094,0.493333,0.394822,0.482329,0.042807
Cluster 4: 0.086957,0,0.013333,0,0,0.005218
Cluster 5: 0,0.018868,0,0.280822,0,0.002243

Missing values globally replaced with mean/mode

Final cluster centroids:
                               Cluster#
Attribute               Full Data          0          1          2          3          4          5
                          (167.0)     (15.0)     (12.0)     (40.0)      (4.0)     (57.0)     (39.0)
==========================================================================================================
indegree                    0.0851     0.0203     0.3116      0.075     0.7174     0.0679     0.0111
outdegree                   0.0369     0.0201     0.1085      0.033     0.6132          0     0.0203
Degree                      0.0389     0.0071     0.1589      0.033       0.64     0.0075     0.0044
closnesscentrality          0.2655     0.9556      0.384     0.3326     0.4825          0     0.2606
betweenesscentrality        0.0288     0.0014     0.1261     0.0149     0.6624          0     0.0008
pageranks                    0.006     0.0031     0.0165      0.005     0.0482     0.0046     0.0026



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      15 (  9%)
1      12 (  7%)
2      40 ( 24%)
3       4 (  2%)
4      57 ( 34%)
5      39 ( 23%)
```

Figura 16: Resultados cluster en *dataset* Beta con algoritmo *k-means* (líder)

Como sucedía en el caso del algoritmo de EM, en los resultados obtenidos en la Figura 15 para el escenario Alfa, descartaremos el cluster 1 al ser el que contiene al docente. Tomaremos, por lo tanto, el siguiente con los valores más altos, es decir, el cluster 2 que agrupa a 9 estudiantes.

Por otro lado, para los resultados obtenidos en la Figura 16 para el *dataset* Beta, descartamos el cluster 3 (por ser el de los docentes) y tomaremos el número 1 (formado por 12 usuarios) por ser el que tiene los valores más altos dentro del conjunto de estudiantes.

```
Initial starting points (random):

Cluster 0: 0,0,0,0.318351,0
Cluster 1: 0.040816,0.230769,0.080645,0.6795,0.125142
Cluster 2: 0.020408,0.076923,0.032258,0.533122,0
Cluster 3: 0,0,0,0,0
Cluster 4: 0.061224,0,0.048387,0.30707,0.003868
Cluster 5: 0.020408,0.076923,0.032258,0.410804,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3          4          5
                           (55.0)     (26.0)      (1.0)      (9.0)      (2.0)      (6.0)     (11.0)
=============================================================================================================
indegree                  0.0471     0.0055          1     0.0907          0     0.0442     0.0334
outdegree                 0.1007      0.003          1     0.3162          0     0.0513     0.1189
Degree                    0.0584      0.005          1      0.138          0     0.0457     0.0513
closnesscentrality        0.3978      0.318          1     0.6595          0     0.2722     0.4582
betweenesscentrality      0.0367     0.0001          1     0.1018          0     0.0084     0.0042



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      26 ( 47%)
1       1 (  2%)
2       9 ( 16%)
3       2 (  4%)
4       6 ( 11%)
5      11 ( 20%)
```

Figura 17: Resultados cluster en *dataset* Alfa con algoritmo *k-means* (periférico)

Para el procedimiento con los usuarios con perfil periférico, seguimos, también, con el mismo proceso que con el algoritmo de EM. Descartaremos, para el caso del escenario Alfa presente en la Figura 17, el cluster 3 (al tener valores demasiado bajos) y nos quedaremos con el cluster 4.

Para el escenario Beta que aparece en la Figura 18, observamos un escenario muy parecido al que teníamos en Alfa por lo que seguiremos el mismo proceso. Descartamos el cluster 0 (por no ser lo suficientemente participativos) y tomaremos el número 4 (formado por 41 usuarios) por ser el que tiene los valores en global más bajos.

```
Initial starting points (random):

Cluster 0: 0,0.018868,0,0.364985,0
Cluster 1: 0.086957,0.018868,0.026667,0.364179,0.02223
Cluster 2: 0,0.018868,0,0.284722,0
Cluster 3: 0.695652,0.415094,0.493333,0.394822,0.482329
Cluster 4: 0.086957,0,0.013333,0,0
Cluster 5: 0,0.018868,0,0.280822,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3          4          5
                          (167.0)     (15.0)     (12.0)     (38.0)      (4.0)     (57.0)     (41.0)
=============================================================================================================
indegree                  0.0851     0.0203     0.3116     0.0744     0.7174     0.0679     0.0148
outdegree                 0.0369     0.0201     0.1085     0.0323     0.6132          0     0.0216
Degree                    0.0389     0.0071     0.1589     0.0323       0.64     0.0075     0.0065
closnesscentrality        0.2655     0.9556      0.384     0.3375     0.4825          0     0.2597
betweenesscentrality      0.0288     0.0014     0.1261     0.0148     0.6624          0     0.0015



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       15 (  9%)
1       12 (  7%)
2       38 ( 23%)
3        4 (  2%)
4       57 ( 34%)
5       41 ( 25%)
```

Figura 18: Resultados cluster en *dataset* Beta con algoritmo *k-means* (periférico)

Por lo tanto, al ejecutar el algoritmo nos quedaría la siguiente distribución en los dos *datasets* estudiados:

- **Alfa:** 9 mentores y 6 estudiantes.

Perfil	ID Autor	Nombre
Mentor	24fb77d6ef839d751679f9c2db497de78a3cd7c8	Estu-96
	3ad6f72cd9b5208dfe50d2b8634fa5f6132e7645	Estu-82
	a8b41b15be0686c2d494f90641de310474f9910d	Estu-49
	e4fc98fc073fca23a32301cceb2f74a355fdbbd7	Estu-33
	0aad920a63c8e397ae53ee34c0ae076a1561dbd0	Estu-155
	1ed75217e05b580a3e51b67db2ac57374229868e	Estu-69
	be273ceabe66f0b0ef2f29336b16ca411731a218	Estu-3

	73114c6ea4639b6149f70c19f5cf28259edd698b	Estu-79
	aa8c6473812dd80c8bb037015585c51978a30355	Estu-151
Estudiante	37ac418bf9558984fbcdae5f44007800808df879	Estu-88
	a7a27d2220b27a8472e1e9c8b251a31a9935538f	Estu-30
	fc0e05ee720422b76129dc4f2ed8561b0e6d21d5	Estu-77
	0a8755c98b57a70b4aa6154e740e55fe9f52c060	Estu-24
	45c2fa9a13586c13ee4eff1669425d88a7a443fc	Coor-126
	809aae900144ebff84da6ad36445a9dd3bb4072a	Estu-20

Tabla 4: Distribución mentor/alumno en escenario Alfa

Para este caso, podemos observar que este algoritmo toma los mentores detectados en el algoritmo de EM y añade unos cuantos más. Además, también hay varios estudiantes que se repiten con respecto al anterior método.

- **Beta:** 12 mentores y 41 estudiantes.

Perfil	ID Autor	Nombre
Mentor	2dc78a0929f266496f96144c29f3cd22804b5c1e	Estu-1
	74cf09e81e75ec98dc1212d933737417656536ec	Estu-2
	2cab7655699fa464b94aba35dc4a5c057b23899d	Estu-3
	9c594aa7e30bfbe0b219bf433e79d1a247290cf8	Estu-4
	71736b0e5b9943903a8a33dac5a19501727debf9	Estu-5
	39c2d1e0b5a9e388ecda28eb1b778a7c01a911e4	Estu-6
	bd727f3c7a104eab070f398b2e10a8da2b207221	Estu-7
	c4e8c05baf64ab276357f51e42b0e71268678e2f	Estu-8
	9c4743ee0633578b10a0535a1e0f6fb771adc7c6	Estu-9
	63bb9b5f521460377bb06a65c28a366042911bcd	Estu-10
	9e4817c1eaacbf9f65ad4a952c3bb15d41efb1e3	Estu-11
	48e6baec8283839a354eff890d5816a4a446a0fe	Estu-12

Estudiante	403f51fb2118d605acb50fa8fc7e62652a3cee7b	Estu-54
	75ad6b19f6a25a046f62b046ae7364483a1a63de	Estu-14
	5c74c84de1eee4f5dc178288dc44eaea7d5a7de1	Estu-15
	8c133465b34d69180da5f54bc64922fc9ca7b5c5	Estu-16
	7f496c216f69aca021c80b060603e62ff2973faa	Estu-17
	1db2ab0329523f70c1a53cdc1f67363fa88ae56c	Estu-18
	ece3e9a281905a24b560b6ad712bb99d32178671	Estu-19
	599fb05778bcacf06fe82e06dafdc76b97e11c32	Estu-20
	db531d817355304f9bfa16221c933cfd3e8921b9	Estu-21
	ccbb88c407120f84bb295e91592d27fab2df6d84	Estu-22
	85cbb1bde5f80b1c7b7a9d207a64a48de7cdb513	Estu-23
	db8d0ece288fe3968822e92546f642ab55ba657d	Estu-24
	3ffee227aeded42187685d13965e56f0d915af31	Estu-25
	c0c16eefd5cb6e12ab9910cf105d51041398cc74	Estu-27
	7d2cc0fac12ccacf7c892a227a5057bf4e383433	Estu-28
	0b292695c44c0b1f7bd186ea40c596ac90be4c18	Estu-29
	7625518550c235a10f9dc2e24594e1288277222a	Estu-30
	db296ebbfd3c7bc1ed5fbf5496d360b1e15eb4fd	Estu-31
	c07e3916668fe448dfa49457094bd24fc9514f06	Estu-33
	b263a7c7465683b54166779825c7ad3e3fef96f3	Estu-34
	c78b9eced612629c62580ea493e8a3dea05b5948	Estu-35
	1e6b93d288eb05428a9a13584d764299828647af	Estu-36
	6bc97b5916acb9f1d653a59520b45c07dd03ef5b	Estu-37
	b69837550c52660f74f279876b64862b30c2e7ea	Estu-38
	05e5f5dacbf171c58ef75b619d9a1149ffa5e01d	Estu-39

	63c8192913c6167f6143f7a97ff223f93698d8fb	Estu-40
	6234076827de5c00590346e0450b11c4db33c378	Estu-41
	60b85b931c6c2b7264dd46d5546f2bee66ea5517	Estu-55
	489b5a1b7c473dadbaf5015d69b180f45b7a4664	Estu-42
	a1392c54e1670ffd703b218435660f9abc02d8cb	Estu-43
	15ff2ed69bbd46111342716f20294fbcc1e83f42	Estu-44
	68a2bcc80194c975615b3818f19dc75211250d39	Estu-45
	52a386b7e22b2ef80ba1c58693f2082f0203f975	Estu-46
	2d46e396a812ad355377e2c75565cd320fbb9337	Estu-56
	3ea6c3a689c40ecc24f556345445d3746e7486d1	Estu-47
	0766d083371ef85477d06309807a55a7c314ec90	Estu-48
	c780e04e4046b10c634497cd94ba211a3a9b6cb1	Estu-49
	f1ee2b6acbd6dbc1e8dee75d4097661587ee956e	Estu-50
	b6c7341adc243743fc13df67d1eae45b24dd5ec2	Estu-51
	f4601992af376bd5638f5b909f1fd91d0a899ec1	Estu-52
	dcb531fccb2dd980aa43834df7ee5fb7690b8fda	Estu-53

Tabla 5: Distribución mentor/alumno en escenario Alfa

Para el caso del escenario Beta esta igualdad es aún mayor entre el uso de los algoritmos EM y *k-means*. Mientras que, por un lado, la diferencia entre mentores es nula (se ha detectado los mismos mentores empleando cada tipo de algoritmo), en el caso de los estudiantes de los 41 detectados, 38 de ellos ya se detectaron también al usar EM y solo 3 de ellos difieren.

Conclusiones

Como se puede observar por los resultados anteriormente obtenidos, la elección del uso del algoritmo de EM o el de *k-means* supone una diferencia mínima en los resultados. Para el caso del escenario Beta, que corresponde al *dataset* de mayor tamaño, de los 53 alumnos detectados haciendo uso de cada algoritmo, 50 de ellos coinciden.

Debido a esta similitud en los resultados, parece que deberemos basarnos en otros parámetros para tomar una decisión. Artículos como (Jung et al., 2014) y (Nathiya et al., 2010) realizan ciertas comprobaciones en las que se nos indica que *k-means* requiere de más tiempo y, por tanto, de una mayor capacidad de procesamiento que EM.

Basándonos en esto, se ha tomado la decisión de emplear para este procedimiento el algoritmo de clustering de EM y, por lo tanto, para las siguientes fases se emplearán los resultados obtenidos en el uso de este.

Modularidad

Según (Holster, 2022) la modularidad es una medida que describe hasta que extensión está presente la estructura de comunidades dentro de una red. La detección de comunidades es esencial en la minería de medios sociales ya que "*los individuos a menudo forman grupos según sus intereses*" (Anaya, 2019b).

Esta es una de las bases de la teoría de la correlación social, la homofilia social. Esta se puede definir como "*la tendencia de los individuos a asociarse y relacionarse con otros similares. Los individuos [...] comparten características comunes (creencias, valores, educación...) que hacen que la formación de la comunicación y de la relación sean más fáciles*" (Anaya, 2019b). Por lo tanto, usuarios pertenecientes a una misma comunidad son, por lo general, más compatibles entre ellos que con otros fuera de esta.

Desde el enfoque de nuestra solución, podemos comprobar si existen posibles duplas de mentor-estudiante que se encuentren en la misma comunidad. Para ello, calcularemos la modularidad de la red, detectaremos las distintas comunidades, asignaremos a que comunidad pertenece cada uno de los usuarios detectados en la fase anterior y comprobaremos si es posible formar alguna dupla mentor-estudiante con usuarios de una misma comunidad.

Por ejemplo, para el caso del escenario Alfa que tenemos en la Figura 19 (generada con la herramienta de Gephi), vemos que en total se forman 6 comunidades. Los usuarios detectados en la anterior fase quedan distribuidos de la siguiente manera:

Comunidad	Mentor	Alumno
1	-	Estu-24
2	Estu-155 Estu-49	Estu-110 Estu-20

	Estu-82	
3	-	Estu-92
4	-	Estu-19

Tabla 6: Distribución en comunidades en escenario Alfa

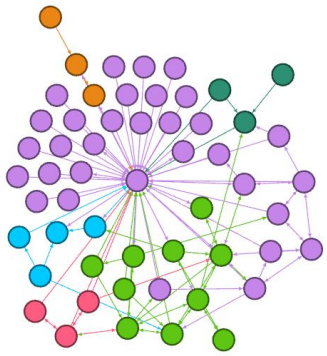

Figura 19: Modularidad en escenario Alfa

Como podemos visualizar, todos los posibles mentores detectados están en la comunidad 2. En esta, además, están *Estu-110* y *Estu-20* como posibles perfiles de alumnos. El resto de usuarios periféricos se encuentran en comunidades distintas por los que no los tendremos en cuenta por ahora.

Por otro lado, para el escenario Beta que tenemos en la Figura 20 (generada, de nuevo, con la herramienta de Gephi), vemos que en total se forman 9 comunidades. Los usuarios detectados en la anterior fase quedan distribuidos de la siguiente manera:

Comunidad	Mentor	Alumno
0	Estu-3 Estu-8	Estu-35 Estu-46
1	Estu-2 Estu-4	Estu-13 Estu-14 Estu-15 Estu-16 Estu-17

		Estu-18 Estu-19
2	Estu-10	Estu-34 Estu-44
3	Estu-1 Estu-5 Estu-6 Estu-7 Estu-11	Estu-21 Estu-22 Estu-23 Estu-24 Estu-25 Estu-26 Estu-27 Estu-28 Estu-29 Estu-36 Estu-37 Estu-39 Estu-40 Estu-41
4	-	Estu-32
5	Estu-9	Estu-30 Estu-31 Estu-50
6	-	Estu-33 Estu-38 Estu-47

		Estu-48 Estu-49 Estu-51 Estu-52
7	-	-
8	Estu-12	Estu-20 Estu-42 Estu-43 Estu-45 Estu-53

Tabla 7: Distribución en comunidades en escenario Beta

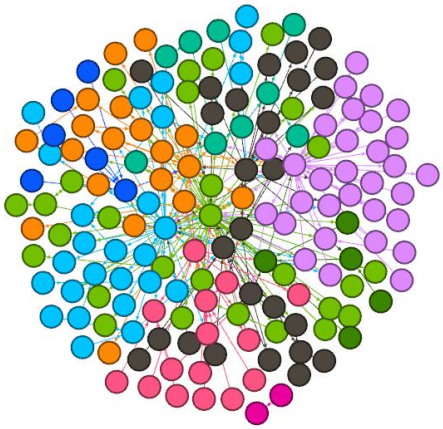

Figura 20: Modularidad en escenario Beta

Como podemos visualizar, y a diferencia de lo sucedido en el anterior escenario, los mentores se distribuyen por casi todas las comunidades. Gracias a esto, más del 80% de los alumnos se encuentran en una comunidad con al menos un mentor. Al igual que en el caso anterior, los pocos usuarios periféricos que se encuentran en comunidades distintas no los tendremos en cuenta por ahora.

Análisis de sentimientos

Una vez que se ha realizado la distribución de los distintos usuarios en sus respectivas comunidades, podemos estudiar si existe comunicación entre los lideres y periféricos de una

misma comunidad y, en caso afirmativo, analizar si estos sentimientos han sido positivos. Este procedimiento se llevará a cabo usando la herramienta de Análisis de Foros UNED (desarrollada por Félix Adame Toledano). En esta herramienta se emplean 3 métodos para el análisis de sentimientos: *N. Bayes* (que recordemos que este procedimiento no considera la neutralidad, es decir, los mensajes solo pueden ser positivos o negativos), *PySentimiento* y un valor que combina los dos algoritmos anteriores. Para este escenario, consideraremos los valores de este último campo para el análisis.

Encontrar conversaciones con sentimientos positivos entre dos miembros de una posible dupla mentor-estudiante es un indicio que nos permite deducir que dicha dupla puede funcionar correctamente en un sistema de mentoría. Del mismo modo, encontrar comentarios negativos, nos hará tratar de evitar juntar a dichos miembros y buscar otras alternativas. En caso de encontrar comentarios neutrales, y que no exista ningún otro comentario positivo, los consideraremos también como posibles candidatos.

Para los escenarios expuestos anteriormente, como Alfa, encontramos, por ejemplo, dos mensajes intercambiados entre los estudiantes Estu-49 y Estu-20. Como se puede visualizar en la Figura 21, estos mensajes tienen una naturaleza neutral.

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u7	Estu-49	Estu-20	1.3470814223	pos	NEU	0.0041666673
	u21	Estu-20	Estu-49	1.2432204803	pos	NEU	0.0041685979

Figura 21: Sentimientos en la conversación entre Estu-49 y Estu-20 del ejemplo Alfa

Si bien es cierto que en primera instancia nos interesa más relaciones en las que haya habido conversaciones con sentimientos positivos, al no haber ninguna de este estilo y siendo esta neutral (que no negativa) podemos considerar estos dos estudiantes como una posible dupla.

Para el caso del escenario Beta, al haber una mayor cantidad de interacciones entre los distintos alumnos y de comunidades, comprobamos que el numero de conversaciones entre usuarios de distintos perfiles en una misma comunidad también es mayor. A continuación, se realiza un análisis de dichas interacciones por cada comunidad detectada mostrando como se manejaría la situación ante la presencia de mensajes positivos y negativos:

- **Comunidad 0:** Para esta comunidad no existen intercambio de mensajes entre usuarios con perfiles distintos. Por lo tanto, no es posible realizar un análisis de sentimientos para comprobar la compatibilidad entre estos.

- **Comunidad 1:** En esta encontramos las interacciones del usuario "*lider*" Estu-4 (con ID 9c594aa7e30bfbe0b219bf433e79d1a247290cf8) con los usuarios "*periféricos*" Estu-16 (con ID 8c133465b34d69180da5f54bc64922fc9ca7b5c5) y Estu-13 (con ID 2f4ce8a0803c214d4e6a307f29f6b75b66fcec7b). Como se puede observar en la Figura 22, mientras la conversación entre Estu-4 y Estu-16 tiene una naturaleza neutral (siendo esta incluso negativa si consideramos exclusivamente el algoritmo de *PySentimiento*), la que mantienen Estu-4 con Estu-13 tienen unos valores positivos considerablemente altos. Esta comparación nos permite determinar que una dupla ente Estu-4 y Estu-13 tiene mayores posibilidades de resultar exitosa que una entre Estu-4 y Estu-16.

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u61	2f4ce8a080:	9c594aa7e30b	1.7290791905	pos	POS	0.0058399092
	u92	5c74c84de1e	74cf09e81e75	0.0071492048	pos	POS	0.0062561734
	u56	8c133465b3·	74cf09e81e75	1.0193160726	pos	NEG	0.0025333838
	u11	9c594aa7e3	8c133465b34c	2.6706259362	pos	NEG	0.0025359176

Figura 22: Sentimientos en las conversaciones de Estu-4 con Estu-13 y Estu-16 del ejemplo Beta

- **Comunidad 2:** En esta encontramos solo una interacción entre el usuario "*periférico*" Estu-34 (con ID b263a7c7465683b54166779825c7ad3e3fef96f3) y el usuario "*líder*" Estu-10 (con ID 63bb9b5f521460377bb06a65c28a366042911bcd). Como se puede observar en la Figura 23, esta conversación tiene una naturaleza neutral. Dado que es la única en esta comunidad y que no es negativa, se puede considerar como posible dupla.

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u71	b263a7c746!	63bb9b5f5214	0.0001678138	pos	NEU	0.0042314496

Figura 23: Sentimientos en las conversaciones de Estu-34 con Estu-10 del ejemplo Beta

- **Comunidad 3:** Esta comunidad es la que mayor número de usuarios (tanto con perfil "*líder*" como con "*periférico*") presenta y esto, además, provoca que el número de conversaciones entre usuarios de distinto perfil sea también mayor al de otras comunidades. Por ello, y para facilitar el entendimiento, analizaremos cada caso encontrado agrupándolos según el usuario "*líder*":

  - **Estu-1**: Este usuario (con ID 2dc78a0929f266496f96144c29f3cd22804b5 c1e) intercambia mensajes con Estu-36 (con ID 1e6b93d288eb05428a9a13584d76429 9828647af) y Estu-37 (con ID 6bc97b5916acb9f1d653a59520b45c07dd03ef5b). Como se observa en la Figura 24, la conversación entre Estu-1 y Estu-37 esta categorizada como negativa, mientras que la otra (entre Estu-1 y Estu-36) se considera

neutral. Teniendo en cuenta el criterio expresado en el análisis de la comunidad 1, daremos preferencia a una conversación neutral frente a una negativa y, por lo tanto, consideraremos la dupla entre Estu-1 y Estu-36 una con mayores posibilidades de resultar exitosa.

	Emisor	Alias	Destinatario	Sentim. Ar	Sentim. A	Sent. N.	PySentimi	Sent. Co
54	71736b0e5b9943903a8a3	u16	2dc78a0929f266496f9614	2.4e-23	2.4e-23	pos	NEU	0.0042
55	1e6b93d288eb05428a9a1	u136	2dc78a0929f266496f9614	4.8e-05	4.8e-05	pos	NEU	0.0042
56	544f93b201a8b70c0448c	u13	2dc78a0929f266496f9614	3.6e-22	3.6e-22	pos	NEU	0.0042
57	cdd56ae2c09543ead46bfc7	u43	2dc78a0929f266496f9614	0.00066	0.00066	pos	NEU	0.0043
58	cdd56ae2c09543ead46bfc7	u43	2dc78a0929f266496f9614	0.00028	0.00028	pos	NEG	0.0026
59	152cdc6590be0a88d35e7	u8	2dc78a0929f266496f9614	1.1e-05	1.1e-05	pos	POS	0.0059
60	152cdc6590be0a88d35e7	u8	2dc78a0929f266496f9614	0.001	0.001	neg	NEG	0.00053
61	544f93b201a8b70c0448c	u13	2dc78a0929f266496f9614	1.1e-08	1.1e-08	pos	NEU	0.0042
62	9428e0ee5cfff43a732b3a	u26	2dc78a0929f266496f9614	0.0057	0.0057	pos	NEU	0.0045
63	74cf09e81e75ec98dc1212	u10	2dc78a0929f266496f9614	8.5e-13	8.5e-13	neg	NEU	0.002
64	2cab7655699fa464b94aba	u42	2dc78a0929f266496f9614	1e-14	1e-14	neg	NEU	0.002
65	ea355aef36cc0939240773	u3	2dc78a0929f266496f9614	2.3e-06	2.3e-06	pos	NEU	0.0042
66	ea355aef36cc0939240773	u3	2dc78a0929f266496f9614	4.2e-10	4.2e-10	pos	NEU	0.0042
67	bd727f3c7a104eab070f39	u45	2dc78a0929f266496f9614	4.6e-16	4.6e-16	pos	NEU	0.0042
68	6bc97b5916acb9f1d653a5	u137	2dc78a0929f266496f9614	0.00015	0.00015	neg	NEG	0.00043
69	9428e0ee5cfff43a732b3a	u26	388d78c290292dd9adbe6	0.25	0.25	neg	NEU	0.0045

Figura 24: Sentimientos en las conversaciones de Estu-1 con Estu-36 y Estu-37 del ejemplo Beta

- **Estu-5:** Este usuario (con ID 71736b0e5b9943903a8a33dac5a19501727de bf9) intercambia mensajes con Estu-26 (con ID 0ea07c4c210e890117a4e452a2af75af2 7100166) y con Estu-27 (con ID c0c16eefd5cb6e12ab9910cf105d51041398cc74). Como se puede observar en la Figura 25 y en la Figura 26, ambas conversaciones tiene una naturaleza neutral. Sin embargo, por la puntuación otorgada en la columna "Sent. Combinado" parece que la conversación entre Estu-5 y Estu-27 tiene una puntuación más alta (y, por lo tanto, más positiva). Por ello, elegiremos esta dupla por delante de la formada entre Estu-5 y Estu-26.

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u16	71736b0e5b'	0ea07c4c210e	1.1115472798	neg	NEU	0.0020000000

Figura 25: Sentimientos en las conversaciones de Estu-5 con Estu-26 del ejemplo Beta

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u123	c0c16eefd5c	71736b0e5b9	1.8679326056	pos	NEU	0.0041735014

Figura 26: Sentimientos en las conversaciones de Estu-5 con Estu-27 del ejemplo Beta

- **Estu-6:** Este usuario (con ID 39c2d1e0b5a9e388ecda28eb1b778a7c01a91 1e4) intercambia mensajes con Estu-21 (con ID db531d817355304f9bfa16221c933cfd 3e8921b9), Estu-22 (con ID ccbb88c407120f84bb295e91592d27fab2df6d84), Estu-23 (con ID 85cbb1bde5f80b1c7b7a9d207a64a48de7cdb513), Estu-24 (con ID db8d0ece2 88fe3968822e92546f642ab55ba657d) y Estu-26 (con ID 0ea07c4c210e890117a4e452

a2af75af27100166). De acuerdo con la Figura 27, vemos como hay 3 mensajes considerados neutrales (Estu-21, Estu-22 y Estu-23) y 2 considerados negativos (Estu-24 y Estu-26). Descartaremos estos dos últimos y elegiremos entre aquellos que participan en mensajes neutrales. Sin embargo, en la misma figura se puede observar como estos 3 presentan exactamente la misma puntuación y, por lo tanto, no tenemos la suficiente información para tomar una decisión de cual es más adecuado.

	Emisor	Alias	Destinatario	Sentim. An	Sentim. A	Sent. N.	PySentimi	Sent. Co
71	0ea07c4c210e890117a4e45	u17	39c2d1e0b5a9e388ecda2	4.3e-15	4.3e-15	neg	NEG	0.00037
72	9428e0ee5cfff43a732b3aa8	u26	39c2d1e0b5a9e388ecda2	6.4e-05	6.4e-05	neg	NEU	0.002
73	2dc78a0929f266496f96144c	u25	39c2d1e0b5a9e388ecda2	0.00013	0.00013	pos	NEU	0.0042
74	db8d0ece288fe3968822e92	u104	39c2d1e0b5a9e388ecda2	6.7e-05	6.7e-05	neg	NEG	0.00041
75	71736b0e5b9943903a8a33d	u16	39c2d1e0b5a9e388ecda2	1.1e-12	1.1e-12	pos	NEU	0.0042
76	ccbb88c407120f84bb295e9	u102	39c2d1e0b5a9e388ecda2	2e-08	2e-08	pos	NEU	0.0042
77	db531d817355304f9bfa162	u101	39c2d1e0b5a9e388ecda2	5.6e-05	5.6e-05	pos	NEU	0.0042
78	f8ea261fbb0fbf0e3355ea4b	u98	39c2d1e0b5a9e388ecda2	1.8e-14	1.8e-14	pos	NEU	0.0042
79	9428e0ee5cfff43a732b3aa8	u26	39c2d1e0b5a9e388ecda2	4.7e-10	4.7e-10	pos	NEU	0.0042
80	d8991eca02bc9ea6a1703a0	u63	39c2d1e0b5a9e388ecda2	3.7e-27	3.7e-27	neg	NEG	0.00037
81	85cbb1bde5f80b1c7b7a9d2	u103	39c2d1e0b5a9e388ecda2	3.4e-10	3.4e-10	pos	NEU	0.0042
82	9428e0ee5cfff43a732b3aa8	u26	3ab8fb70e2d97745457bb	1.6e-05	1.6e-05	pos	NEU	0.0042

Figura 27: Sentimientos en las conversaciones de Estu-6 con otros usuarios con perfil "*periférico*" del ejemplo Beta

- **Estu-7:** Este usuario (con ID bd727f3c7a104eab070f398b2e10a8da2b207 221) intercambia mensajes con Estu-25 (con ID 3ffee227aeded42187685d13965e56f0 d915af31). Este mensaje se considera neutral, aunque si nos fijamos en las valoraciones de los algoritmos de *N. Bayes* y *PySentimiento* vemos que difieren. Aún así, siguiendo el criterio previamente establecido, consideraremos esta dupla.

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u106	3ffee227aed	bd727f3c7a10	1.6632040476	pos	NEG	0.0025333537

Figura 28: Sentimientos en las conversaciones de Estu-7 con Estu-25 del ejemplo Beta

- **Estu-11:** Este usuario (con ID 9e4817c1eaacbf9f65ad4a952c3bb15d41efb 1e3) intercambia mensajes con Estu-39 (con ID 05e5f5dacbf171c58ef75b619d9a1149f fa5e01d), Estu-40 (con ID 63c8192913c6167f6143f7a97ff223f93698d8fb), Estu-41 (con ID 6234076827de5c00590346e0450b11c4db33c378). Los mensajes con Estu-40 y Estu-41 tienen connotaciones negativas, mientras que el de Estu-39 es neutral. Consideraremos este último para la dupla.

	Emisor	Alias	Destinatario	Sentim. An	Sentim. A	Sent. N.	PySentimi	Sent. Co
293	71736b0e5b9943903a8a3	u16	9c594aa7e30bfbe0b219bf	8e-08	8e-08	pos	NEU	0.0042
294	9e4817c1eaacbf9f65ad4a	u31	9e4817c1eaacbf9f65ad4a	0.016	0.016	pos	NEU	0.0048
295	6234076827de5c0059034	u150	9e4817c1eaacbf9f65ad4a	6.1e-05	6.1e-05	neg	NEG	0.00041
296	544f93b201a8b70c0448c	u13	9e4817c1eaacbf9f65ad4a	6.7e-09	6.7e-09	pos	NEU	0.0042
297	63c8192913c6167f6143f7a	u149	9e4817c1eaacbf9f65ad4a	0.068	0.068	pos	NEG	0.0038
298	791f7b6cd787c28140528f5	u15	9e4817c1eaacbf9f65ad4a	0.21	0.22	neg	NEU	0.0043
299	ea355aef36cc0939240773	u3	9e4817c1eaacbf9f65ad4a	0.6	0.6	neg	NEG	0.0042
300	05e5f5dacbf171c58ef75b6	u148	9e4817c1eaacbf9f65ad4a	0.0024	0.0024	neg	NEU	0.0022
301	590fbc12fb83131dfa8020	u64	Inicia el hilo	7.6e-06	7.6e-06	pos	NEU	0.0042

Figura 29: Sentimientos en las conversaciones de Estu-11 con otros usuarios con perfil "*periférico*" del ejemplo Beta

- **Comunidad 4:** Esta comunidad presenta usuarios con perfil "*periférico*", pero no con perfil "*líder*". Por ello, al no ser posible generar duplas, descartaremos esta comunidad.

- **Comunidad 5:** En esta comunidad encontramos una situación similar a la anterior. El usuario "*lider*" Estu-9 (con ID 9c4743ee0633578b10a0535a1e0f6fb771adc7c6) tiene interacciones con los usuarios "*periféricos*" Estu-31 (con ID db296ebbfd3c7bc1ed5fbf5496 d360b1e15eb4fd) y Estu-50 (con ID f1ee2b6acbd6dbc1e8dee75d4097661587ee956e). Como se puede observar en la Figura 30, la conversación entre Estu-9 y Estu-31 es bastante negativa mientras que la que hay entre Estu-9 y Estu-50 en la Figura 31 se considera neutral. Teniendo en cuenta el mismo criterio expuesto anteriormente, daremos preferencia a una conversación neutral frente a una negativa. Por lo tanto consideraremos la dupla entre Estu-9 y Estu-50 una con mayores posibilidades de resultar exitosa que la otra.

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u70	db296ebbfd:	9c4743ee063:	5.4006406976	neg	NEG	0.0003666666

Figura 30: Sentimientos en las conversaciones de Estu-9 con Estu-31 del ejemplo Beta

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u166	f1ee2b6acbc	9c4743ee063:	5.9481106503	pos	NEU	0.0041667886

Figura 31: Sentimientos en las conversaciones de Estu-9 con Estu-50 del ejemplo Beta

- **Comunidad 6:** Aquí sucede lo mismo que en la comunidad 4, por lo tanto se descartará también.

- **Comunidad 7:** En esta comunidad no se ha detectado ningún usuario que encaje con el perfil de "*líder*" o "*periférico*". Se descartará esta comunidad.

- **Comunidad 8:** En esta encontramos solo una interacción entre el usuario "*periférico*" Estu-42 (con ID 489b5a1b7c473dadbaf5015d69b180f45b7a4664) y el usuario "*líder*" Estu-12 (con ID 48e6baec8283839a354eff890d5816a4a446a0fe). Como se puede observar en la Figura 32, y tal y como sucedía también en la comunidad 2, esta conversación tiene

una naturaleza neutral. Dado que es la única en esta comunidad, y que no es negativa, se puede considerar como posible dupla.

Icono	Alias	Emisor	Destinatario	Sent. Analys	Sentimiento	PySentimien	Sent. Combi
	u87	489b5a1b7c	48e6baec828	1.4161191757	pos	NEU	0.0041666854

Figura 32: Sentimientos en las conversaciones de Estu-42 con Estu-12 del ejemplo Beta

Como se ha podido observar, estos análisis nos permiten, una vez se han generado las diferentes comunidades, detectar que usuarios con perfil "*líder*" tienen más posibilidades de congeniar con aquellos con perfil "*periférico*" por medio de la evaluación de sus interacciones. Sin embargo, este sistema basado en el análisis de sentimientos puede resultar, en entornos educativos como en los que nos encontramos, poco eficientes. Se profundizará sobre este tema en el siguiente apartado.

3.5 Análisis de los resultados

En esta sección desarrollaremos y analizaremos los resultados obtenidos en las anteriores y expondremos las parejas de alumnos que serían propuesta por el sistema que proponemos en este trabajo.

Alfa

Para el caso del escenario Alfa, el número de resultados obtenidos se reduce a 1. Solo la dupla formada por los alumnos Estu-49 y Estu-20 ha sido presentada como una buena candidata para un sistema de mentoría (a pesar de que en el análisis de sentimiento, los mensajes intercambiado entre ambos han sido neutrales.

Beta

Por el contrario, y debido especialmente a que en este escenario el número de instancias es mucho mayor (más del doble con respecto al de Alfa), en el escenario Beta si hemos obtenido diversos resultados. Las duplas sugeridas pueden verse en la Tabla 8:

Comunidad	Mentor	Estudiante
1	Estu-4	Estu-13
2	Estu-10	Estu-34

3	Estu-1	Estu-36
	Estu-5	Estu-27
	Estu-6	Estu-21 Estu-22 Estu-23
	Estu-7	Estu-25
	Estu-11	Estu-39
5	Estu-9	Estu-50
8	Estu-12	Estu-42

Tabla 8: Resultados de la generación de duplas en el escenario Beta

Como se puede observar se han propuesto 9 duplas mentor-estudiante. Los usuarios que participan en cada dupla, además, no se repiten en las otras haciendo que cada alumno solo tenga un mentor o estudiante asignado.

Cabe destacar el caso particular que se da en la comunidad 3 donde, para el mentor Estu-6, existen 3 estudiantes que son igual de compatibles con dicho mentor. Para ese caso habría que estudiar si aplicar alguna evaluación adicional o, simplemente, tomar una decisión al azar.

3.6 Conclusiones del estudio

En esta última sección de este capítulo nos centraremos en las conclusiones alcanzadas en el estudio realizado. Para ello nos fijaremos en 3 puntos principales.

El tamaño sí importa

La mayor diferencia existente entre los dos conjuntos de datos usados durante el estudio (para el ejemplo Alfa y Beta respectivamente) es el tamaño de estos. Tal y como se muestra en la sección “Parámetros de entrada” de este mismo capítulo, el *dataset* usado en el ejemplo Alfa cuenta, inicialmente, con 229 instancias; mientras que el usado en Beta contiene 527 instancias (mas del doble con respecto al primero).

Esta diferencia de tamaños provoca variaciones desde el principio que se traslada durante todo el proceso para los dos escenarios. Esto puede verse reflejado en la clasificación de

estudiantes que se realiza al principio del procesamiento, donde para el ejemplo Alfa se obtienen únicamente 3 mentores y 6 alumnos; mientras que el ejemplo Beta detecta 12 mentores y 41 alumnos. Es decir, a pesar de que el ejemplo Beta cuenta solo con poco más del doble de instancias que el Alfa, el número de mentores detectados es 4 veces mayor y, para el caso de los periféricos, este número es casi 7 veces mayor.

Esto, como hemos dicho además, se traslada al resto del proceso. Un mayor número de mentores y alumnos detectados hace que sea más probable que al menos cada mentor tenga algún alumno en su misma comunidad (y viceversa) al calcular la modularidad. Esto, a su vez, provoca también que sea más probable que entre miembros de una misma comunidad se hayan intercambiado mensajes dentro del foro universitario pudiendo, así, analizar los sentimientos en estos mensajes.

Y es que, aunque el sistema propuesto es capaz de proporcionar resultados fiables sin importar el tamaño del *dataset*, lo cierto es que un mayor número de instancias proporciona un grafo con información menos limitada, más rica, que acaba dando como resultado duplas mentor-estudiante basadas en una información y en unos cálculos más robustos.

La búsqueda de sentimientos en un mundo de neutrales

Tal y como se indica en el capítulo 2, el análisis de sentimientos en redes sociales es de gran utilidad en la actualidad, especialmente en el mundo del marketing digital. Este análisis nos permite conocer los sentimientos que las personas transmiten a través de los textos que redactan en una red social, pero, ¿es igual de efectivo para todos los tipos de redes sociales?

El análisis realizado en los dos escenarios descritos en este capítulo (haciendo uso de la herramienta de Análisis de Foros UNED y, en concreto, los análisis con *PySentimiento* al ser el que considera la existencia de mensajes neutrales) nos muestran que en entornos educativos, como los foros universitarios, el número de mensajes "neutrales" es muy superior al resto (esto se puede observar en la Figura 33, para el escenario Alfa, y en la Figura 34, para el escenario Beta). Para Alfa este porcentaje se aproxima al 66%, pero es que para el escenario Beta (que cuenta con una mayor cantidad de datos) este porcentaje es superior al 73%.

Es en el ejemplo Alfa donde vemos más claramente este problema. Durante el análisis, y debido al pequeño tamaño del *dataset* del cual solo obtenemos la interacción entre dos usuarios, observamos que el análisis de sentimientos no aporta ninguna información adicional más allá de verificar que la interacción entre los dos usuarios no es negativa. Si bien es cierto que en el caso del escenario Beta se han conseguido encontrar ejemplos, existen otros que presentan los mismos problemas que los que encontramos en el ejemplo Alfa.

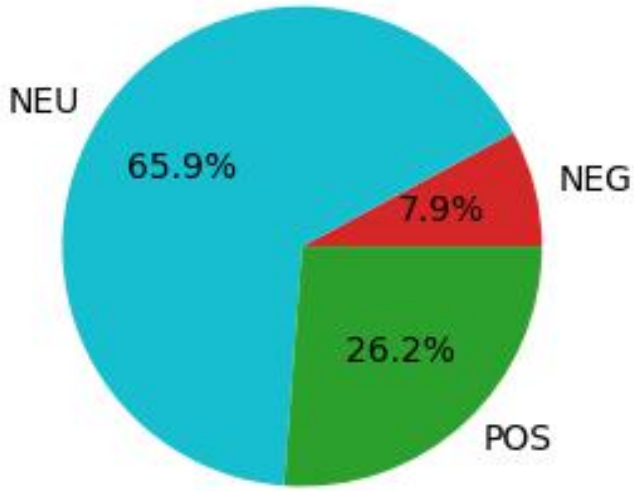

Figura 33: Distribución de sentimientos en el escenario Alfa

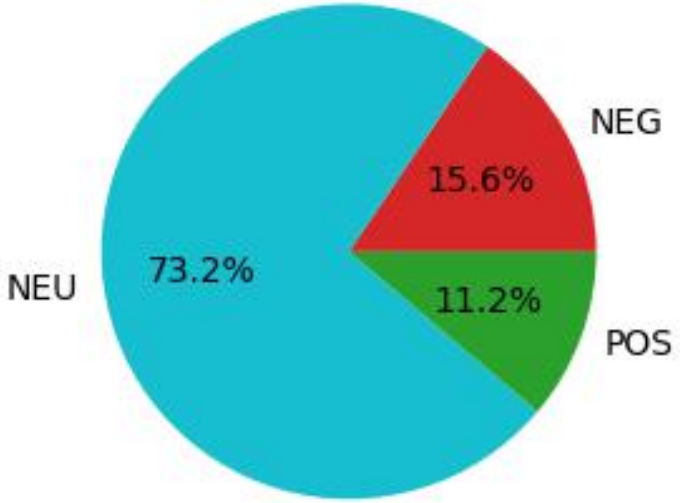

Figura 34: Distribución de sentimientos en el escenario Beta

Además, es importante analizar los mensajes que el análisis considera como "positivos" y "negativos". Los mensajes expuestos en la Figura 35 y la Figura 36 pueden entenderse de la misma manera, el emisor estando de acuerdo con lo expuesto por el receptor; sin embargo, el primero es considerado como un mensaje "positivo" y el segundo como "negativo". Es fácil ver que el segundo mensaje tiene una connotación más negativa que el primero, pero este sentimiento está enfocado a un tercero. Y es que, en muchas ocasiones, los sentimientos que emiten un mensaje no coinciden con los sentimientos entre el emisor y el receptor.

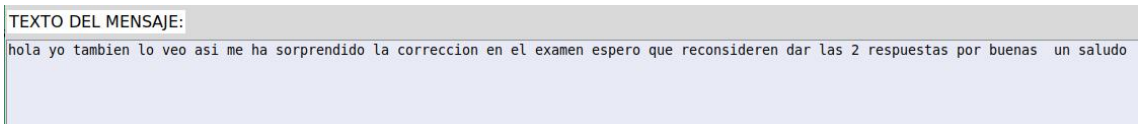

Figura 35: Mensaje "positivo" entre Estu-31 y Estu-9 en el escenario Beta

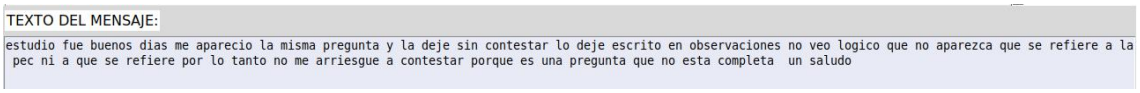

Figura 36: Mensaje "negativo" entre Estu-13 y Estu-4 en el escenario Beta

También en mensajes como el de la Figura 37 se confunde a veces sentimientos "positivos" con educación por medio del agradecimiento expresado por el emisor que, de nuevo, no nos proporciona mucha información sobre la relación entre dos usuarios.

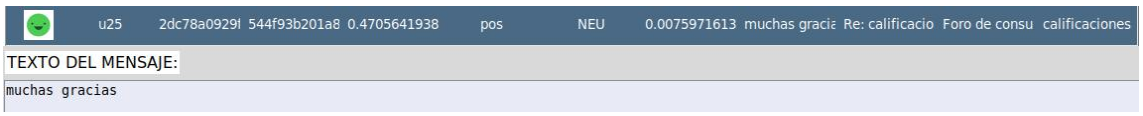

Figura 37: Mensaje "positivo" entre dos estudiantes en el escenario Beta

Este sistema de análisis de sentimientos presenta, además, un problema de base adicional, la presuposición de la existencia de conversaciones entre miembros de una misma comunidad. Para realizar el análisis de sentimientos es necesario que haya mensajes entre los dos usuarios objetos del estudio y, partiendo del hecho de que uno de los perfiles que buscamos se caracteriza por ser poco participativo, esto hace que en muchas ocasiones haya usuarios "*periféricos*" que no intercambien ningún mensaje con usuarios "*líderes*" de su misma comunidad. Esto provoca que realizar un análisis de este tipo no sea posible para muchos de estos usuarios y queden "fuera" de la evaluación.

Es por estos motivos por los que, probablemente, sea mejor buscar una alternativa al análisis de sentimientos. Quizás la respuesta y la solución a nuestro problema resida en considerar algo más allá de los datos dinámicos, los datos estáticos.

La (posible) revolución de lo estático

Mientras que los datos dinámicos son algo que están en constante cambio, a menudo debido a factores externos (como son los mensajes enviados dentro de un foro universitario); los datos estáticos hacen referencia a aquella información que no cambia (con frecuencia). Ejemplos de este tipo de datos tenemos, por ejemplo, los nombres de ciudades o países, descripción de eventos... (DCR, 2022).

El uso de estos datos estáticos, como ya se ha usado en otros estudios similares, puede ser de gran utilidad para nuestro escenario. Por ejemplo, como ya comentamos en el capítulo 2, en la Open University UK se consideraban datos como el curso, la localización geográfica, la situación doméstica (tiene hijos, padre soltero...), genero y edad (Boyle et al., 2010).

En (Abyaa, 2019), por otro lado, se considera otros datos como el conocimiento del estudiante (que comprende el nivel, competencias, conceptos erróneos...), las características cognitivas, que comprende las habilidades del estudiante para percibir, organizar, procesar y recordar información (esto incluye, entre otros, el estilo de aprendizaje); los rasgos de personalidad (que nos puede permitir desvelar como una persona afrontará determinadas situaciones); y la motivación.

Una característica fundamental de las universidades basadas en la educación a distancia es la gran variedad de perfiles que esta presenta entre su alumnado. Esto lo podemos encontrar desarrollado en (Fernández, 2022), donde se definen a los alumnos de la UNED como "*unos destinatarios extraordinariamente diversos en cada una de las situaciones o perspectivas sobre las que podemos clasificarlos. Jóvenes y adultos; solteros y con compromisos familiares; de todos los oficios, ocupaciones y profesiones; de cualquier nivel social y cultural; con estudios universitarios o sin ellos; y de muy distinta situación económica*".

Esto provoca que tanto los perfiles como las situaciones de los distintos estudiantes varíe en gran medida de uno a otro y no considerar este tipo de datos puede provocar escenarios no deseables.

Un ejemplo de esto se puede observar en los rangos de edad. Mientras que en el conjunto del sistema universitario español hay una evidente mayoría de estudiantes menores de 25 años (tal y como se observa en la Figura 38), en la UNED esta situación es distinta.

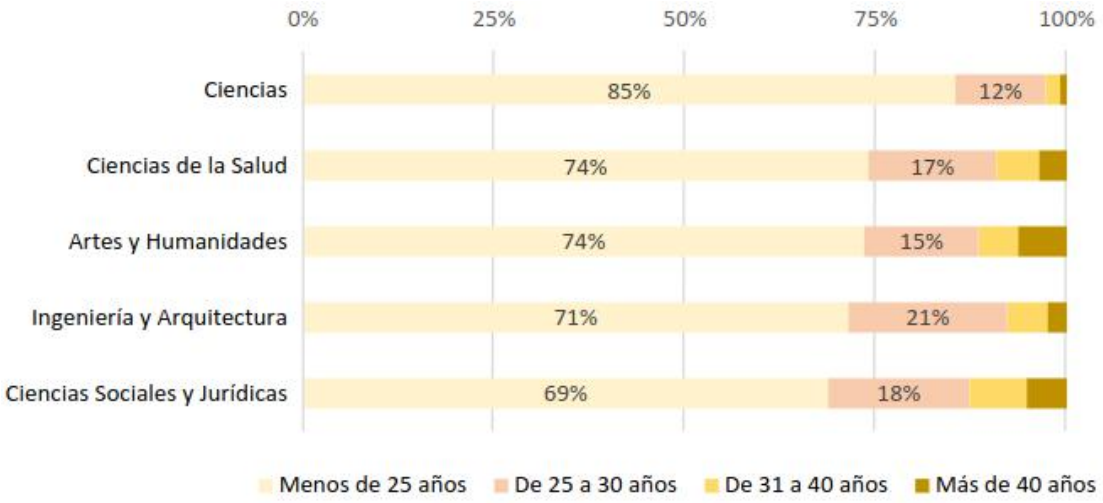

Figura 38: Distribución por grupo de edad de los estudiantes egresados en Grado por rama de enseñanza en el curso 2020-2021 (SIIU, 2022)

Según las estadísticas de matriculación anonimizadas proporcionadas (recogidas en el recurso *datos_tituls-01-2023-30_ 08_2023_21_52.csv*), y tal como se muestra en la Tabla 9; aunque los alumnos menores de 25 años siguen siendo aquel grupo de edad con mayor número

de estudiantes, la diferencia con otros es mínima. Esta escasa diferencia se mantiene hasta los 50 años, rango a partir del cual comienza a descender considerablemente el número de estudiantes. Es importante tener en consideración esta variedad en los rangos de edad porque, si no se considera (y con ello, otros parámetros relacionados) pueden surgir problemas en el desarrollo del sistema de mentoría.

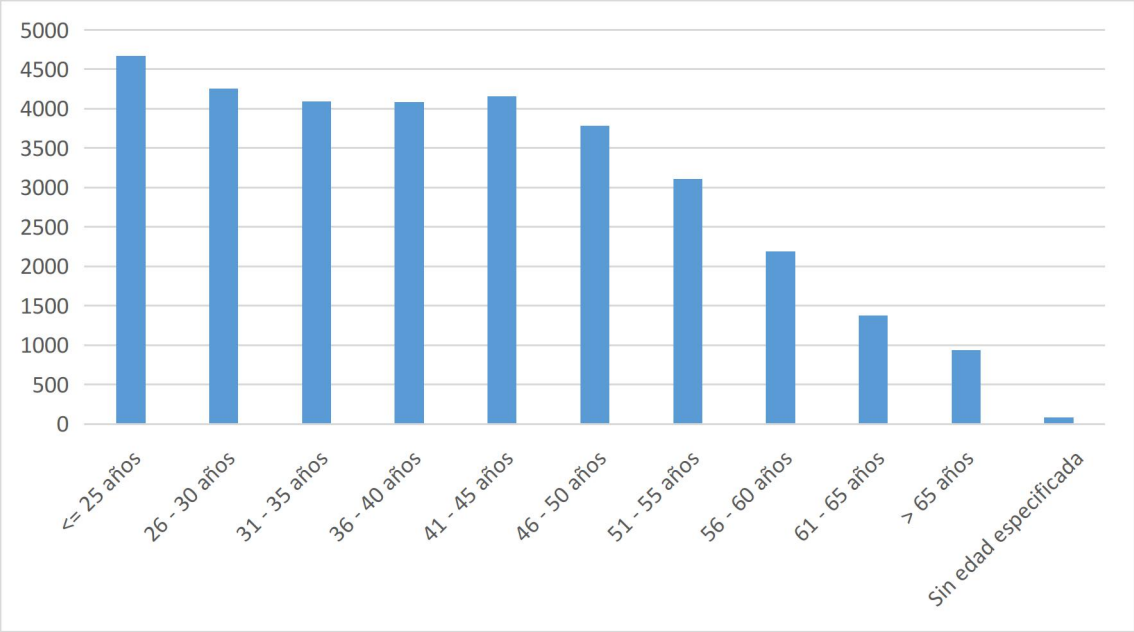

Tabla 9: Gráfica del total de alumnos matriculados según el rango de edad

Por ejemplo, imaginemos un sistema de mentoría en el que no se tenga en cuenta la situación doméstica de sus estudiantes. Podría darse el caso de que un alumno de 20 años acabara como mentor de otro de 40 cuyo mayor obstáculo al afrontar los estudios sea compaginarlo con tener un hijo. El mentor en este escenario podría verse superado por la situación, al no estar familiarizado con una situación de esa naturaleza, y el alumno, por su lado, no sentir que el programa le esté aportando nada.

Es por ello por lo que, tener en cuenta los datos estáticos, puede resultar esencial para maximizar la eficacia de un sistema de mentoría. Valores como el curso, la situación geográfica, la situación familiar o, incluso, los rasgos de personalidad y la motivación (obtenidos haciendo uso de cuestionarios) pueden ayudar a formar duplas mentor-estudiante más adecuadas para ambas partes y, como consecuencia, más robustas y exitosas frente al objetivo final de reducir la tasa de abandono.

Capítulo 4

Propuesta de proyecto

4.1 Contenido y alcance del proyecto

La finalidad del proyecto es la implementación de un sistema formado por una aplicación que, a partir de la aportación de un *dataset* que contiene las conversaciones de un foro entre diferentes alumnos, identifique a que tipo de perfil corresponde cada uno de ellos según su participación y, posteriormente, genere duplas mentor-estudiante según la compatibilidad entre estos alumnos (formada cada dupla por un miembro de cada perfil).

Al finalizar este proceso, se deberá mostrar un listado con aquellas duplas que, según los parámetros considerados, sean más compatibles para un sistema de mentoría junto con la argumentación del porqué se ha tomado dicha decisión.

Alcance del proyecto

El objetivo principal de este proyecto será reducir la tasa de abandono en las entidades educativas por medio de un sistema de mentoría que obtiene recomendaciones de posibles duplas mentor-estudiante según determinados parámetros de entrada (entre los que se incluyen los datos dinámicos generados en las conversaciones en un foro).

Estas predicciones resultantes son obtenidas aplicando técnicas de aprendizaje automático a los datos proporcionados de los foros educativos. Para ello se hará usos de algoritmos de aprendizaje no colaborativo, concretamente de algoritmos de clustering.

4.2 Plan de trabajo

En esta sección se procede a detallar el plan de trabajo. Este estará dividido en paquetes de trabajo (WP por sus siglas en inglés, *Working Package*) junto con sus respectivas actividades:

WP1 - Gestión del proyecto

Este paquete de trabajo se subdividirá en 3 actividades:

- **Administración del proyecto y monitorización de los recursos:** Esta actividad será la encargada de proporcionar soporte administrativo y financiero al proyecto mientras controla que los recursos destinados a cada una de las otras actividades es el necesario.
- **Control de calidad y monitorización de los planes de trabajo:** Esta actividad se centrará en asegurar la calidad del proceso y que los distintos planes de trabajo cumplan el calendario impuesto.
- **Coordinación técnica:** Esta última actividad del WP1 está enfocada en proporcionar apoyo técnico para alcanzar los objetivos durante todo el proceso de desarrollo.

WP2 - Planificación y diseño

Este paquete de trabajo se subdividirá en 2 actividades:

- **Análisis de requisitos:** Realiza un análisis en profundidad de los requisitos funcionales y no funcionales de la aplicación. Además, en esta actividad se identificarán las tecnologías y herramientas necesarias para el desarrollo.
- **Diseño de la interfaz de usuario (IU):** Por medio de *wireframes* y prototipos de la interfaz de usuario, se diseña la estructura de la aplicación, incluyendo la forma en la que se cargarán los datos de los foros universitarios y como se mostrarán los resultados finales.

WP3 - Algoritmo de generación de duplas

Este paquete de trabajo se subdividirá en 4 actividades:

- **Interpretación de los datos de los foros universitarios:** Estos datos serán la base para el funcionamiento del algoritmo de generación de duplas. La aplicación deberá ser capaz de de interpretar correctamente los datos proporcionados de los foros universitarios y realizar el preprocesamiento necesario para que este pueda ser usado correctamente por el algoritmo que se desarrollará.
- **Clasificación de alumnos:** Esta actividad se centrara en el desarrollo de algoritmo para que, a partir de los datos preprocesados de la anterior actividad, se realice una clasificación de los alumnos de acuerdo a si su perfil corresponde al de un mentor o al de un estudiante (o a ninguno de los dos).
- **Detección de duplas:** A partir de la clasificación de la anterior actividad, se desarrollará el algoritmo para que pueda generar duplas mentor-estudiante según diversos parámetros que muestren su compatibilidad.

- **Herramientas de generación de reportes:** La tarea final de este plan de trabajo se encargará de desarrollar un proceso que sea capaz de generar informes que resuman los datos obtenidos, las decisiones tomadas y las acciones ejecutadas; y que permita a los ingenieros de la instalación comprobar el correcto funcionamiento de la herramienta.

WP4 - Visualización e interacción de datos

Este paquete de trabajo se subdividirá en 2 actividades:

- **Visualización de datos:** El objetivo de esta actividad es investigar métodos actuales de representación de datos, sus problemas y su influencia con la interacción de usuarios con el fin de analizar nuevos paradigmas de visualización para la presentación de datos y encontrar uno que se adapte a las exigencias del proyecto.
- **Diseño de la interacción y de la interfaz gráfica del usuario:** Esta actividad se enfocará en del diseño de la interacción del usuario y en implementar, y probar, distintas interfaces gráficas de usuario hasta encontrar aquella que proporcione el equilibrio adecuado entre personificación y cumplimiento de los estándares de usabilidad.

WP5 - Pruebas y depuración

Este paquete de trabajo se subdividirá en 2 actividades:

- **Pruebas funcionales:** Realiza pruebas exhaustivas para garantizar que la aplicación cumple con los requisitos obtenidos en WP2.
- **Pruebas de compatibilidad:** Evalúa el rendimiento de la aplicación al agrupar usuarios en duplas mentor-estudiante y ajusta el algoritmo según sea necesario.

WP6 - Documentación y despliegue

Este paquete de trabajo se subdividirá en 2 actividades:

- **Documentación:** Crea documentación técnica y de usuario para la aplicación.
- **Despliegue:** Prepara la aplicación para su despliegue en sistemas de escritorio (Windows, macOS o Linux). Además, proporciona instrucciones de instalación y configuración.

WP7 - Mantenimiento y mejora continua

Este paquete de trabajo se subdividirá en 2 actividades:

- **Mantenimiento post-lanzamiento:** Establece un proceso de seguimiento y solución de problemas para el período post-lanzamiento.
- **Mejoras y actualizaciones:** Continúa mejorando la aplicación según la retroalimentación de los usuarios y las necesidades cambiantes.

Cronograma del proyecto

A continuación, en la Tabla 10, se expone el cronograma del proyecto según los diferentes paquetes de trabajos especificados anteriormente.

En este, solo se exponen los 6 primeros paquetes de trabajo (desde WP1 hasta WP6) que son los que corresponden con el proceso de planificación, desarrollo y evaluación del producto. Las actividades englobadas en WP7, relativas al mantenimiento y la mejora del producto, se desarrollará durante los años posteriores a la finalización del proyecto.

Como se observa en el cronograma, y sin contar el WP1 (que se realiza durante todo el proceso implicado en el desarrollo del proyecto), los diferentes planes de trabajo se realizarán escalonadamente en orden y, alguno de ellos, se realizarán de manera concurrente. El tiempo previsto para la realización del proyecto es de dos años, estando el primer año centrado en la planificación, el diseño y el desarrollo del algoritmo de generación de duplas; mientras que el segundo está enfocado en desarrollar la interfaz de usuario, las pruebas, la documentación y el despliegue.

Tabla 10: Cronograma del proyecto según los distintos paquetes de trabajo

	Año 1												Año 2											
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
WP1: Gestión del proyecto																								
ACT.1.1: Administración del proyecto y monitorización de los recursos																								
ACT.1.2: Control de calidad y monitorización de los planes de trabajo																								
ACT.1.3: Coordinación técnica																								
WP2: Planificación y diseño																								
ACT.2.1: Análisis de requisitos																								
ACT.2.2: Diseño de la interfaz de usuario																								
WP3: Algoritmo de generación de duplas																								
ACT.3.1: Interpretación de los datos de los foros universitarios																								
ACT.3.2: Clasificación de alumnos																								
ACT.3.3: Detección de duplas																								
ACT.3.4: Herramientas de generación de reportes																								
WP4: Visualización e interacción de datos																								
ACT.4.1: Visualización de datos																								
ACT.4.2: Diseño de la interacción y de la interfaz gráfica del usuario																								
WP5: Pruebas y depuración																								
ACT.5.1: Pruebas funcionales																								
ACT.5.2: Pruebas de compatibilidad																								
WP6: Documentación y despliegue																								
ACT.6.1: Documentación																								
ACT.6.2: Despliegue																								

4.3 Plan de explotación

En una primera aproximación, tras el lanzamiento de nuestro servicio, se focaliza en el ámbito del mercado nacional puesto que determinar inicialmente las posibilidades de expansión internacional se considera muy precipitado y poco fiable por la cantidad de factores que intervienen en el desarrollo de este.

Análisis del mercado y demanda

Al estar enfocado en entornos educativos, podemos decir que este será el mercado objetivo para nuestro producto. Sin embargo, esto no solo se limitará a los modelos de universidad a distancia (aún habiéndonos centrado en estos en nuestra memoria debido a la alta tasa de abandono que presenta), sino que podrá extrapolarse a otros como la universidad presencial o, incluso, modelos no universitarios como las FPs. El objetivo final de nuestro producto es que sea capaz de generar duplas mentor-estudiante a partir de las conversaciones en un foro sin importar el nivel o el tipo de educación.

Como parte del análisis de mercado a realizar para este proyecto se establecerán contactos con centros educativos públicos y privados para fijar una serie de ellas como casos de estudio iniciales para conocer las necesidades concretas y obtener unos resultados iniciales con nuestro servicio. Se les ofrecerá el uso de nuestro servicio de manera promocional durante un plazo a determinar.

También sería interesante considerar la posibilidad de liberar cierta partes del código del proyecto para permitir la creación de una comunidad de desarrollo que pueda crear complementos al propio sistema, lo que podría crear un ecosistema en el que el cliente final pudiera personalizar aún más el sistema para su uso particular.

Competencia

En el estudio realizado en la fase de estado de la cuestión, se compara otras propuestas de mentoría con la presentada en este documento. Algo que diferencia a los distintos sistemas de mentoría es el empleo de una gran diversidad de datos estáticos, adaptándose a las características de cada entorno educativo, para la toma de decisiones a la hora de generar las duplas; pero, fuera de eso, los procesos suelen ser bastante similares. Comparados con nuestra aplicación, existen dos características que las diferencia de cualquier otra:

- **Automatización del proceso de generación de duplas:** Todos los procesos de mentoría observados se caracterizan por ser procesos manuales, ejecutados por personas. Nuestra propuesta propone un proceso automático que genera las duplas a partir de los datos aportados y un algoritmo.
- **Empleo de datos dinámicos:** Aunque se ha comentado que existe una gran variedad de datos según los diferentes modelos, estos tienen en común que son datos estáticos, frente a los datos dinámicos que se emplean en el modelo compuesto.

Teniendo en cuenta estos dos puntos distintivos, se ha realizado una búsqueda de productos que proporcionen este servicio haciendo uso de estas características. Y aunque, tal y como observamos en la última sección del capítulo 2, existen numerosos ejemplos de sistemas de mentoría por todo el mundo, no se ha encontrado ninguno que emplee datos dinámicos o utilice procedimientos de automatización para la generación de duplas.

En este último caso, el ejemplo más aproximado que se ha encontrado es el presentado en (Gómez-Flechoso et al., 2019), que emplea la automatización para evaluar el funcionamiento del programa y el efecto que este puede tener en el proceso de aprendizaje. Sin embargo, el uso del aprendizaje automático se limita a eso, sin explorar su uso para el procedimiento de generación de duplas. Quizás este producto, en futuras revisiones, pueda aproximarse a lo que proponemos aquí.

Actividades de promoción y comercialización previstas

Nuestro producto debe ser promocionado desde el inicio del desarrollo y de manera extraordinaria una vez que se hayan realizado las pruebas de conformidad del servicio y se hayan obtenido los primeros resultados satisfactorios en forma de una reducción de las tasas de abandono en las instituciones educativas que se han seleccionado como iniciales.

Esta promoción y difusión se deberá realizar dentro de entornos en los que se mueva nuestra comunidad objetivo, la educativa. Así, por ejemplo, existen ferias como SIMO EDUCACIÓN, evento de referencia tecnológica para los profesionales de la actividad docente en el que se reúne anualmente a las marcas de tecnología y contenidos digitales para la enseñanza. Este tipo de ambientes son perfectos para la promoción de nuestro servicio permitiéndonos estar en contacto directo con miembros del sector educativo que puedan estar interesados en este.

Además de esto, sería interesante contar con un sitio web del proyecto y redes sociales. Ambos deberán estar actualizados con información importante sobre el proyecto para que este destaque.

Para la comercialización, el proyecto se presentaría como un servicio. Se habilitarán una serie de paquetes de servicio en función de diversas características tales como la personalización de los parámetros de entrada o el número máximo de instancias permitidas. Habrá igualmente que determinar los períodos de contrato (anual, bianual…) y los precios de los distintos paquetes de servicio. Estas tarifas se establecerán una vez se conozcan los resultados del estudio de mercado para conseguir el mayor beneficio económico, eso sí siempre siendo realistas y valorando la continuidad de la empresa para ajustar esa relación de beneficio y continuidad.

Como campañas de marketing se pueden promover bonos de servicio en los que el compromiso del cliente tenga recompensa económica. Como ejemplo, una suscripción a un paquete de servicio por un período de cinco años podría ser recompensado con un porcentaje de descuento en el pago de la cuota.

Capítulo 5

Conclusiones y trabajo futuro

5.1 Conclusiones del trabajo

Como se expresa en el primer capítulo de este TFM, este proyecto presentaba dos objetivos generales. Por un lado, plantear un algoritmo que permitiera generar duplas mentor-estudiante y, por el otro, desarrollar una propuesta de solución a partir de las conclusiones obtenidas en el objetivo anterior. Estos presentaban, por su parte, otros subojetivos los cuales se listan a continuación e indicaremos, para cada uno de ellos, si se han cumplido o no y en qué medida o qué faltaría para cumplirlos:

OBJ.1 - Algoritmo de generación de duplas

- OBJ. 1.1 - Deberá ser capaz de clasificar los participantes de un foro según diversas medidas de la red: Basándonos en el estudio realizado por (Marcos-García et al., 2015), hemos empleado determinadas medidas para dividir, según su nivel de participación y de importancia dentro de la red, a los alumnos en dos perfiles: "*líder*" y "*periférico*". Esta agrupación se ha realizado con un algoritmo de aprendizaje no supervisado y, para tomar la decisión de cual emplear, se han realizado pruebas en dos algoritmos de este tipo (EM y *k-means*) comparando tanto los resultados obtenidos al ejecutarlos como sus necesidades computacionales.

- OBJ. 1.2 - Deberá ser capaz de detectar como de compatibles son dos participantes en base a la actividad y las emociones proyectadas en esta: Utilizando las medidas de modularidad de una red para generar comunidades y análisis de sentimientos entre los mensajes intercambiados entre alumnos de diferentes perfiles, hemos sido capaces de detectar la compatibilidad entre usuarios para generar duplas mentor-estudiante. Sin embargo, no estamos seguros de si estos dos métodos usados son los mejores para evaluar la compatibilidad de dos alumnos dentro de un entorno educativo y, quizás, sería conveniente realizar más pruebas con otros acercamientos.

OBJ.2 - Desarrollo de propuesta de solución

- OBJ. 2.1 - Desarrollar un plan de trabajo: Se ha especificado un plan de trabajo de 24 meses para el desarrollo de una aplicación funcional que haga uso del algoritmo de generación de duplas especificado en OBJ.1. Este plan de trabajo se ha dividido en 7 paquetes de trabajo que quedan representados temporalmente en un cronograma del proyecto.

- OBJ. 2.2 - Desarrollar un presupuesto: Este subobjetivo, finalmente, no se ha podido llevar a cabo. El principal motivo de ello es que, desde mi punto de vista, sería interesante estudiar si se podría adscribir a alguna ayuda de desarrollo del proyecto como las ayudas CDTI. Dado que habría que realizar un estudio sobre las diferentes ayudas y si, finalmente, estás se pueden adaptar a la naturaleza de nuestro proyecto (con el tiempo que esto puede llevar), se ha decidido proponer este punto como un trabajo a futuro.

- OBJ. 2.3 - Desarrollar un plan de explotación: Este plan de explotación se ha expuesto realizando un análisis del mercado y de la demanda de nuestro producto. Además, se ha hecho un estudio de la competencia (con ejemplos de otros proyectos de la misma índole indicando qué los diferencia del nuestro) y se han desarrollado las actividades de promoción y de comercialización previstas.

Una vez analizados los objetivos propuestos, procedemos, para cerrar esta sección, a comentar las oportunidades que este trabajo puede proporcionar, desde un punto de vista personal.

Como ya comentamos en este TFM, la educación a distancia está experimentando un tremendo auge durante estos últimos años (potenciado especialmente por la pandemia global del COVID-19). Esto no resulta sorprendente, pues este modelo es cómodo y permite la suficiente flexibilidad como para poder compaginarlo con la vida laboral o con determinadas situaciones domésticas.

Sin embargo, las tasas de abandono en este modelo universitario siguen siendo una asignatura pendiente (si no la más importante). Estos valores son extremadamente altos (especialmente si los comparamos con el modelo presencial) y es por ello por lo que es imprescindible tomar medidas que permitan reducir esos datos.

La mayor diferencia entre el modelo presencial y el no presencial reside en el carácter personal. La naturaleza del primero, en el que debemos compartir un espacio físico con otros alumnos, hace que sea más sencillo que surjan relaciones y grupos entre ellos, a partir de intereses comunes. Esto es muy importante pues, en momentos difíciles y estresantes, es más fácil encontrar apoyo y sentirse arropado por otros compañeros. El docente también, al ser las

clases presenciales, puede notar que alumnos están más perdidos, preocuparse por ellos y asesorarlos si es necesario.

Estas relaciones, en modelos a distancia, resultan más complicadas de construir y, en la mayoría de los casos, se reducen a mensajes en el foro o a correos electrónicos. Esto puede provocar que ciertos alumnos se sientan aislados y sin apoyo para continuar con sus estudios.

Considero que una aplicación con estas características, capaz de detectar patrones de casos que puedan estar cercanos al abandono y que pueda sugerir posibles duplas para un sistema de mentoría en el que un alumno (compatible con él) pueda ayudarlo durante su vida académica, es de bastante utilidad.

Sin embargo, personalmente, reducir el abandono y la perdida de talento consecuente debería ser un objetivo secundario. El objetivo principal de esta aplicación debe tener un carácter más humano y estar centrado en el estado anímico de los alumnos buscando que estos no solo no piensen en abandonar los estudios, sino que al realizar estos encuentren una mayor satisfacción personal y, con ello, una mayor felicidad, evitando así posibles depresiones o ataques de ansiedad. Y este objetivo, en una época como la que nos encontramos y con los casos de suicidio en máximos en España (Mediavilla, 2023), es más que suficiente.

5.2 Trabajo futuro

En este último apartado comentaremos propuestas para ampliar los contenidos de este trabajo o, incluso, para evaluar otras vías para ciertas decisiones que se han tomado aquí:

- **Desarrollo de la aplicación:** El primero de estos trabajos debería ser el desarrollo del proyecto que se ha sugerido en este TFM, haciendo uso del estudio realizado en el capítulo 3 y de la propuesta plasmada en el capítulo 4.
- **Evaluar el uso del análisis de sentimientos para la detección de compatibilidad:** Como comentábamos en las conclusiones del capítulo 3, el uso del análisis de sentimientos es de gran utilidad para diversos escenarios, pero puede no ser igual de útil para todos, pues exige que los mensajes tengan cierta carga sentimental. Ejemplos como las redes sociales o una sección de opiniones sobre un producto son buenos ejemplos en los que este tipo de análisis pueda ser útil; pero, en otros como un foro universitario (en el que la mayoría de mensajes son breves y neutrales) quizás este no sea el mejor acercamiento. Sería conveniente evaluar en profundidad si el uso del análisis de sentimientos en entornos educativos puede resultar beneficioso o, por el contrario, no proporciona la suficiente

información como para considerarlo un método funcional para evaluar las emociones y el estado anímico de los alumnos.

- **Emplear datos estáticos en conjunto con los datos dinámicos:** En este trabajo nos hemos centrado en los datos dinámicos para la generación de las duplas mentor-estudiante. Sin embargo, no deberíamos olvidar la importancia del uso de los datos estáticos, especialmente para conocer la compatibilidad entre dos alumnos. Datos como la edad, la situación doméstica o la localización geográfica pueden resultar de gran importancia y no tenerlos en cuenta puede provocar que duplas que, teóricamente, deberían funcionar acaben fracasando. El principal problema con este punto es como plantear y desarrollar esto respetando la normativa dictada por el GDPR; pero, de conseguirse, las duplas generadas tendrán una base mucho más robusta y, por lo tanto, mayores posibilidades de resultar exitosas.

- **Desarrollo de un presupuesto acorde con la naturaleza del proyecto:** Aunque se ha realizado la propuesta de proyecto, uno de los puntos que faltan es el presupuesto de este. Se propone como trabajo desarrollar un presupuesto que analice diversos métodos de financiación para el desarrollo del proyecto aquí expuesto. Podrían barajarse desde ayudas a nivel nacional o europeo que puedan adaptarse a la naturaleza de este y que nos permitan colaborar con universidades (y otras entidades educativas) de cualquier tipo (como es el caso de las ayudas CDTI o el programa Horizonte Europa); u otra forma de financiación por parte de algún programa universitario.

Bibliografía

(Abyaa, 2019) - Abir Abyaa, Mohammed Khalidi Idrissi, Samir Bennani (2019). Learner modelling: systematic review of the literature from the last 5 years. Education Tech Research Dev (2019) 67:1105–1143. https://doi.org/10.1007/s11423-018-09644-1

(Anaya, 2019a) - Anaya, A. R., 2019. *Sistemas de aprendizaje colaborativo*. Apuntes del Tema 1 de la asignatura Métodos de desarrollo y análisis de entornos colaborativos y redes sociales del Máster de Informática, UNED.

(Anaya, 2019b) - Anaya, A. R., 2019. *Teoría sociales de la minería social*. Apuntes del Tema 2 de la asignatura Métodos de desarrollo y análisis de entornos colaborativos y redes sociales del Máster de Informática, UNED.

(Anaya, 2019c) - Anaya, A. R., 2019. *Introducción a la minería de datos en redes sociales*. Apuntes del Tema 4 de la asignatura Métodos de desarrollo y análisis de entornos colaborativos y redes sociales del Máster de Informática, UNED.

(Armando, 2017) - *Los 10 tipos de Redes Sociales y sus características*. https://psicologiaymente.com/social/tipos-de-redes-sociales

(Asbee et al., 1999) - Asbee, S., Simpson, O., & Woodall, S. (1999). *Student–student mentoring in distance education*. Journal of Access and Credit Studies, 2(2), 220–232.

(Azure, 2023) - *¿Qué es el aprendizaje automático?*. https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-machine-learning-platform/

(Bleidorn y Hopwood, 2019) - Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. Personality and Social Psychology Review, 23(2), 190– 203. https://doi.org/10.1177/1088868318772990

(Boxtel et al., 2000) - Boxtel, C.V., der Linden, J.V., Kanselaar, G., 2000. Collaborative learning tasks and the elaboration of conceptual knowledge. Learning and Instruction 10, 311–330.

(Boyle et al., 2010) - Boyle, F. A., Kwon, J., Ross, C. E., & Simpson, O. (2010). Student–student mentoring for retention and engagement in distance education. Open Learning: The Journal of Open and Distance Learning, 25(2), 115–130. https://doi.org/10.1080/02680511003787370

(Conati et al., 2018) - Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in education needs interpretable machine learning: Lessons from open learner modelling. arXiv preprint arXiv:1807.00154.

(Cornell University, 2023) - *Collaborative Learning*. https://teaching.cornell.edu/teaching-resources/active-collaborative-learning/collaborative-learning

(Crosling et al., 2008) - Crosling, G., Thomas, L., & Heagney, M. (2008). *Introduction: Student success and retention*. In G. Crosling, L. Thomas, & M. Heagney (Eds.). *Improving student retention in higher education* (pp. 1–13). Oxford: Routledge.

(Đambić et al., 2016) - Đambić, G., Krajcar, M., & Bele, D. (2016). Machine learning model for early detection of higher education students that need additional attention in introductory programming courses. International Journal of Digital Technology & Economy, 1(1), 1– 11.

(Dans, 2010) - Dans, E. (2010). Todo va a cambiar. Tecnología y evolución: adaptarse o desaparecer. Deusto, 287.

(DCR, 2022) - *Understanding static and dynamic data.* https://datacentrereview.com/2022/12/understanding-static-and-dynamic-data/

(Earle, 2007) - Earle, D. (2007). *Te wahi i ng ā taumata atakura* (Supporting Māori achievement in bachelors degrees). Wellington, NZ: Ministry of Education.

(ETSII, 2023) - *Programa de Mentoría ETSII Orienta*. https://www.informatica.us.es/index.php/programa-de-mentoria

(FC, 2021) - Flare Compare (2021). https://flarecompare.com/Machine%20Learning%20(ML)/EM%20vs.%20K-Means%20Comparing%20Two%20Clustering%20Algorithms/

(Fernández, 2022) - Fernández de Buján, Federico (2022). *La enseñanza a distancia, Su concreción en la UNED. Mi definición. Perfiles de alumno, del profesor y del tutor*. Revista Derechos Humanos y Educación, nº Extraordinario (2022), 65-85.

(Fundación BBVA, 2019) - Informe U-Ranking, 7ª edición. Fundación BBVA (2019).

(García et al., 2016) - García, S., Ramírez-Gallego, S., Luengo, J., Herrera, F. (2016). *Big Data:* Preprocesamiento y calidad de datos. Big Data monografía, 17-23.

(Gibbs et al., 2007) - Gibbs, G., Regan, P., & Simpson, O. (2007). *Improving student retention through evidence based proactive systems at the Open University*. Journal of Student Retention, 8(3), 359–376.

(Gómez-Flechoso et al., 2019) - Gómez-Flechoso, M. & Alonso García, Miguel & Sánchez-Ruiz, A.A. & Díaz, D. & Benitez, P.. (2019). Gestión Automatizada del Programa de Mentoría en la Universidad Complutense de Madrid. 408-413. 10.26754/CINAIC.2019.0085.

(Hilbert et al., 2021) - Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. Review of Education, 9, e3310. https://doi.org/10.1002/rev3.3310

(Hollenbeck et al., 2012) - Hollenbeck, J. R., Beersma, B., & Schouten, M. E. (2012). Beyond team types and taxonomies: A dimensional scaling conceptualization for team description. Academy of Management Review 37, 82-106.

(Holster, 2022) - Holster J.D. (2022). Introduction to R for Data Science: A LISA 2020 Guidebook. Chapter 7 Network Analysis. https://bookdown.org/jdholster1/idsr/

(IBM, 2021) - *Supervised vs. Unsupervised Learning: What's the difference?*. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

(Infinita, 2022) - *Mercado potencial: ¿Qué es y cómo se calcula?*. https://www.infinitiaresearch.com/noticias/mercado-potencial-que-es-y-como-se-calcula/

(Jeffares, 2019) - Jeffares, A. (2019). *K-means: A Complete Introduction*. https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

(Johnson y Johnson, 2005) - Johnson, D.W., Johnson, R.T., 2005. *Cooperative Learning, Values, and Culturally Plural Classrooms*.

(Jung et al., 2014) - Jung, Yong & Kang, Min & Heo, Jun. (2014). *Clustering performance comparison using K-means and expectation maximization algorithms*. Biotechnology, biotechnological equipment. 28. S44-S48. 10.1080/13102818.2014.949045.

(Kleinberg, 2002) - Kleinberg, J., 2002. *An impossibility theorem for clustering*. Proceedings of the 15th International Conference on Neural Information Processing Systems, NIPS'02, pp. 463–470, MIT Press.

(Lee et al., 2009) - Lee, H., & Kwon, J. (2009). *An evaluation of the effectiveness of KNOU Student Mentoring Program and its improvement plan*. Internal paper, Institute of Distance Education, Korean National Open University, p. 73.

(López e Ibarra, 2021) - López Torres, B. J., Ibarra Escobedo, R. (2021). Análisis de Redes Sociales: Aplicaciones en las ciencias sociales. Taberna Libraria Editores, 13.

(Marcos-García et al., 2015) - Marcos-García J-A., Martínez-Monés A., Dimitriadis Y. (2015). DESPRO: A method based on roles to provide collaboration analysis support adapted to

the participants in CSCL situations. Computers & Education, 82:335–53. https://doi.org/10.1016/j.compedu.2014.10.027

(Mediavilla, 2023) - Daniel Mediavilla (2023). *Los suicidios crecen en España desde 2018 y la pandemia agravó el problema*. El País. https://elpais.com/salud-y-bienestar/2023-01-26/los-suicidios-crecen-en-espana-desde-2018-y-la-pandemia-agravo-el-problema.html

(Mi Educación Online, 2023) - *¿Cuáles son los retos que enfrenta la educación a distancia?*. https://mieducacionenlinea.com/retos-de-la-educacion-a-distancia/

(Moseley, 2020) - Moseley, C., 2020. *Collaboration vs cooperation: what's the difference?*. https://blog.jostle.me/blog/collaboration-vs-cooperation

(Nathiya et al., 2010) - G, Nathiya & Punitha, S. & Punithavalli, Dr. (2010). *An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm*. International Journal of Computer Science and Information Security. 7.

(Prieto, 2021) - *El mercado de la educación digital crece fuertemente*. https://cincodias.elpais.com/cincodias/2021/04/22/opinion/1619083717_793145.html

(Pulgar et al., 2014) - Pulgar, E. L., Cedeño, F. G., & De Santiago Alba, C. (2013). El abandono y el egreso en la UNED. En Libro De Actas. VI Jornadas De Redes De Investigación En Innovación Docente De La UNED, 211–218. http://e-spacio.uned.es/fez/eserv/bibliuned:501063/Luque_et_al_Abandono_Egreso_VI_Redes_2014.pdf

(RD Station, 2023) - *Redes sociales*. https://www.rdstation.com/es/redes-sociales/

(Sancho, 2020) - *Algoritmos de clustering*. https://www.cs.us.es/~fsancho/?e=230

(Saqr et al., 2018) - Saqr M, Fors U, Tedre M, Nouri J (2018). How social network analysis can be used to monitor online collaborative learning and guide an informed intervention. PLoS ONE 13(3): e0194777. https://doi.org/10.1371/journal.pone.0194777

(Saxena et al. 2019) - Saxena, A., Saxena, P., Reddy, H., Gera, R. (2019) - *A Survey of Studying the Social Networks of Students*.

(Shaik et al., 2023) - Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. Natural Language Processing Journal, Volume 2. https://doi.org/10.1016/j.nlp.2022.100003

(Shivanandhan, 2020) - *What is Sentiment Analysis? A Complete Guide for Begginers*. https://www.freecodecamp.org/news/what-is-sentiment-analysis-a-complete-guide-to-for-beginners/

(SIIU, 2022) - *Principales resultados EEU 2021-22*. Sistema Integrado de Información Universitaria. Ministerio de Universidades (Gobierno de España).

(Simpson, 2005) - Simpson, O. (2005). E-Learning Democracy and Social Exclusion-Issues of Access and Retention in the UK. In SAGE Publications eBooks. http://oro.open.ac.uk/11856/

(Stahl, 2004) - Stahl, G., 2004. Groupware goes to school: adapting BSCW to the classroom. IJCAT 19, 162–174.

(Xiaozhou, 2020) - *Expectation-maximization algorithm, explained*. https://yangxiaozhou.github.io/data/2020/10/20/EM-algorithm-explained.html#general-framework-what-is-em

(Zepke et al., 2003) - Zepke, N., Leach, L., & Prebble, T. (2003). *Student support and its impact on learning outcomes*.

Apéndice A: Recursos

Archivos

- **Memoria_JuanGarciaRuiz:** Presente documento que contiene la explicación y los resultados del proyecto.
- **AA-I-foro-20-21-anon.xlsx:** Conjunto de datos inicial correspondiente al escenario Alfa.
- **Foro General PDDII Anon.xlsx:** Conjunto de datos inicial correspondiente al escenario Beta.
- **datos_tituls-01-2023-30_08_2023_21_52.csv:** Estadísticas anonimizadas de matriculación de la UNED.

Programas

- Pencil Project: Utilizada para la realización del diagrama de flujo expuesto al inicio del capítulo 3.
- Gephi: Empleada durante la fase de preprocesamiento, para la visualización de los grafos y para el calculo de diferentes medidas de la red, y en la fase de procesamiento, para calcular la modularidad.
- Weka: Usada en la fase de procesamiento, concretamente en la clasificación de alumnos por medio de algoritmos de aprendizaje no supervisados.
- Análisis de Foros UNED: Empleada en la fase de procesamiento para el análisis de sentimientos en los mensajes intercambiados entre los distintos usuarios de la red.

Apéndice B: Siglas

En este apéndice se indicará el significado de cada una de las siglas empleadas en este documento:

- **BBVA:** Banco Bilbao Vizcaya Argentaria.
- **CSCL:** *Computer-supported collaborative learning* (Aprendizaje colaborativo asistido por ordenador).
- **EM:** Esperanza-Maximación.
- **ETSII:** Escuela Técnica Superior de Ingeniería Informática.
- **FP:** Formación Profesional.
- **IU:** Interfaz de Usuario.
- **IUED:** Instituto Universitario de Educación a Distancia.
- **ML:** *Machine Learning* (Aprendizaje automático).
- **NLP:** *Natural Language Processing*(Procesamiento del lenguaje natural).
- **SNA:** *Social Network Analysis*(Análisis de redes sociales).
- **TFM:** Trabajo de Fin de Máster.
- **UNED:** Universidad Nacional de Educación a Distancia.
- **WP:** *Working Package* (Paquete de trabajo).