UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Análisis y predicción del absentismo y éxito de los estudiantes en plataformas de educación online

Álvaro Andújar Amorós
Tutor: Ruth Cobos Pérez

Junio 2016

Análisis y predicción del absentismo y éxito de los estudiante en plataformas de educación online

AUTOR: Álvaro Andújar Amorós

TUTOR: Ruth Cobos Pérez

Dpto. de Ingeniería Informática

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Junio 2016

Resumen

El uso de las nuevas Tecnologías de la Información y de la Comunicación (TIC) hace ya tiempo que no quedaron restringidas a grandes empresas o entidades. Su evolución y desarrollo las han convertido en una de las herramientas más versátiles y con un sin fin de posibilidades, por lo que no es de extrañar que las entidades educativas las hayan incorporado para mejorar sus sistemas educativos y alcanzar a un mayor número de estudiantes.

El recurso elegido por varias universidades de la Comunidad de Madrid para promover estas nuevas TICs ha sido la plataforma de cursos online masivos edX/Open edX [28]. Existen dos formatos que se han implementado en el entorno académico local, los MOOC (Massive Open Online Courses) y los SPOCs (Small Private Online Courses) [43].

El objetivo de este Trabajo Fin de Grado consiste en la creación de un sistema que permita el análisis de la actividad de los estudiantes en los cursos on-line a partir de la generación de unos indicadores que ayuden a detectar el posible absentismo de los participantes en los mismos y predecir quienes podrían ser los alumnos con mayor probabilidad de abandono.

La herramienta analizará la información a partir de datos anonimizados para proteger la privacidad de los estudiantes. Estos datos servirán para generar una serie de indicadores basados en la interacción de los estudiantes en los foros, actividad con el contenido del curso y en la realización de los problemas, contrastados con sus calificaciones del curso. Estos indicadores serán los usados para predecir el abandono estudiantil.

Palabras clave: analizar y predecir, anonimizar, absentismo, indicadores, MOOC y SPOC.

Abstract

The use of new Information Technology and Communication (ICT) is no longer restricted to the use of large companies or organizations. Its evolution and development have made ICTs one of the most versatile tools with endless possibilities. Therefore, it is logical to think that educational institutions would want to use them in order to improve their education systems and reach a greater number of students around the world.

The solution chosen by some many universities in the Comunidad de Madrid to promote these new ICTs has been through the platform of massive online courses, edX [28]. There are two types of courses that could be deployed; the MOOC (Massive Open Online Courses) and the SPOCs (Small Private Online Courses) [43].

The aim of this project is to create an application that allows the analysis of the activity of students in the courses created by the Autonomous University of Madrid and generate indicators in order to analyze the causes of student absenteeism and predict the students most likely to drop with the objective of trying to alleviate it.

The application will be used by the members of the UAMx Office [44] responsible for receiving data from edX and the only team with permission to handle files containing student information. To share this information with teachers and keep private information of students safe, it has created a module in the application that will anonymize the most sensitive information so that they can share data from students to professors and university researchers ensuring the privacy of students.

Keywords: analyze and predict, anonymize, absenteeism, indicators, MOOC and SPOCs.

Agradecimientos

Este trabajo no habría sido posible sin la oportunidad que me brindó Ruth Cobos de realizar una beca en la oficina donde nacieron los cursos MOOC. Gracias a ello, empecé a trabajar con Iván Claros y juntos nos embarcamos en los primeros análisis que me conduciría hasta aquí, gracias por ayudarme a comprender toda la información que recibíamos y por orientarme en el potencial que contenían los paquetes de datos.

Le estoy muy agradecido Pedro García Martín por prestar los datos de su curso MOOC, "La España de El Quijote", sin su consentimiento no se habrían obtenido resultados tan claros sobre mi trabajo.

Un recuerdo muy especial a todos los compañeros de la carrera con los que he compartido asignaturas y viajes, con vosotros las asignaturas eran mucho más divertidas y entretenidas: Rober, Dani, Diego, Isa, Mónica, Pablo y Euler.

Gracias a mis padres por preocuparse todos los días de mi vida como solo los padres lo hacen. Muchas gracias a la Enana, mi hermana, siempre cuidando de mi y ayudándome y chichándome para que de más de mi.

Y finalmente a mi mujer, Sandy, gracias y perdón por los últimos 5 años de carrera que te he hecho pasar, todas las tardes y noches detrás del ordenador, gracias por todo tu apoyo incondicional y por siempre tirar de mi sacando lo mejor de mí.

Este trabajo no estaría aquí sin vosotros y yo tampoco.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

En este capítulo se expone los motivos por los que se ha elegido este tema como Trabajo de Fin de Grado (TFG). A continuación, se explicará en qué marco se ha englobado el proyecto, así como el alcance del mismo especificando sus objetivos. Por último, se expondrá la estructura que sigue el actual documento.

1.1 Motivación

Las nuevas tecnologías de la información no se han visto relegadas a actividades técnicas, ya que la informática ha permitido optimizar la comunicación, conectividad global, automatización de procesos redundantes y mejorar el avance científico.

Es lógico pensar que el sector de la educación centrará su atención en estas nuevas tecnologías y las incorporará en sus procesos habituales de educación o reinventará nuevas maneras de llegar a un mayor número de estudiantes. Esta iniciativa generó nuevas líneas de estudio e investigación sobre su implantación en la Universidad Autónoma de Madrid [4].

Este proyecto tenía como objetivo el de poder participar en la plataforma edX/Open edX [28] y crear contenido bajo el marco de los MOOC. La plataforma edX facilita datos sobre los estudiantes para que las distintas entidades académicas analicen los progresos obtenidos en los diferentes cursos, véase Anexo A para más detalles del proceso de recepción de los paquetes de datos desde edX.

El paquete de datos que se recibe desde edX contiene toda la información de los cursos activos, por ello no es posible analizarlos en referencia a un solo curso directamente sobre los ficheros recibidos. La modalidad de distribución de los paquetes de datos tiene dos modalidades diaria y semanal.

Esto produce que se deba de extraer y combinar información de múltiples ficheros en distintos formatos para obtener una mayor visibilidad sobre la información real de cada curso. Esto define la necesidad de crear un sistema para analizar grandes cantidades de datos conjuntamente pudiendo separar los datos de cada curso y sin necesidad de esperar a tener la recepción semanal del paquete de datos de edX.

Con esa necesidad se determinó el objetivo de crear un sistema donde se pudiera volcar todos los datos que se recibe desde edX, centralizar el acceso a los mismos garantizando su integridad de los datos y que todos tengan un origen homogéneo.

Una observación hecha en los datos de participación de los cursos edX/Open edX, es que su número de matriculados se cuenta en el rango de miles en muchos de sus cursos, pero ha quedado patente que a pesar de ese alto interés son un reducido número de estudiantes los que consiguen finalizar el curso o incluso que no lo abandonen durante el mismo [14]. Esa situación concluyo en que se debía de buscar las causas y tratar de predecir el abandono o desinterés de los estudiantes de un curso MOOC.

1.2 Objetivos

El objetivo de este proyecto es el de desarrollar un sistema, al que denominamos "E-Learning Analytics for Student's Absenteeism" (ELASA), con el que sea posible generar indicadores para predecir el absentismo estudiantil en los cursos MOOC.

Para poder completar la implantación del sistema se deberá de cumplir las siguientes metas:

- Realizar un análisis de las estructuras de los cursos MOOC junto con la información que se facilitan desde la plataforma edX
- Realizar un estudio sobre las tecnologías del aprendizaje electrónico, los cursos online masivos y las analíticas del aprendizaje [11]
- Investigar otras aplicaciones y herramientas ya existentes
- Realizar un análisis de requisitos, funcionales y no funcionales, con el consecuente diseño que ayude a llevar a término el sistema propuesto
- Desarrollar el sistema para que cubra las necesidades básicas planteadas en la definición del Trabajo de Fin de Grado:
    - Almacenar en una base de datos todos los datos de los estudiantes una vez estén anonimizados
    - Generación de informes e indicadores a partir de los datos anónimos almacenados
    - Realizar predicciones y análisis a partir de los indicadores sobre el absentismo estudiantil en los MOOC
- Efectuar pruebas de integración y validación con un conjunto de datos reales
- Concluir con un estudio sobre las posibles mejoras a efectuar en el sistema

1.3 Organización de la memoria

Este documento contiene 6 capítulos que abarcan todos los aspectos del proyecto, desde la creación del TFG, los sistemas implementados, hasta la información técnica y adicional incluida en los Anexos.

En el capítulo actual, Introducción, se detalla la motivación por la cual se ha ideado este proyecto, los objetivos que se quieren alcanzar con el TFG y la organización de la estructura de la memoria.

El resto de la memoria detalla las distintas fases:

- En el capítulo 2 se presenta la investigación realizada para el proyecto y el análisis y comparativa de aplicaciones similares para plataformas de cursos online masivos. Factores como el contexto y los conceptos de las plataformas educativas online, analíticas de aprendizaje y algoritmos de predicción.
- En el capítulo 3 se detallan los requisitos funcionales y no funcionales para cada módulo. Se expone el diseño del sistema, la arquitectura lógica y el modelo de base de datos planteado.
- En el capítulo 4 se expone la metodología de trabajo elegida, la estructura lógica del código desarrollado y su funcionamiento.
- En el capítulo 5 se explican las pruebas realizadas y el resultado de las mismas.

- El capítulo 6 muestra las conclusiones obtenidas y se indican que posibles trabajos futuros se podrían desarrollar a partir de aquí.

Al final, toda la memoria cuenta con distintos Anexos que aclaran información y conceptos usados recurrentemente en este documento:

- Anexo A - Estructura de los paquetes de datos de edX, donde se detalla los procedimientos a seguir para poder acceder y extraer los paquetes de datos de edX
- Anexo B - Se detallan los ficheros que son necesarios para el sistema, así como una explicación detallada de su contenido y estructura. Se indicará el contenido a anonimizar o a excluir antes de incluir los registros en la base de datos.
- Anexo C - Se muestra la estructura de ficheros que el sistema generará automáticamente para extraer información de las ediciones de los cursos MOOC. Además, se explica el contenido que se puede encontrar en cada uno de los ficheros y el significado de las relaciones entre los valores que se extraen en el fichero de mensajes del foro.
- Anexo D – Se facilitado información sobre la planificación ideada con el tiempo a invertir en cada fase. Además, se ha incluido un manual de usuario que muestra ejemplo de todas las funcionalidades y su impacto en la base de datos.

2 Estado del arte

2.1 Introducción

Uno de los puntos que se tuvo en cuenta al comienzo de este proyecto fue la necesidad de encontrar información detallada sobre los datos que se iban a recibir desde la plataforma edX [28]. Continuando con la búsqueda sobre otras aplicaciones o trabajos existentes en los entornos profesionales y/o universitarios relacionados con los entornos de aprendizaje online y cuáles son las tecnologías óptimas para el desarrollo de un proyecto de estas características, de manera que queden cubiertas todas las necesidades que se planteen.

2.2 Investigación de entornos de aprendizaje

2.2.1 Aprendizaje Electrónico

Encontrar una única definición para el Aprendizaje Electrónico [11] [17] [54] se ha vuelto una tarea compleja, ya que cada autor lo expresa según sus propias vivencias. Aun así todos los autores concuerdan en algunos aspectos que se pueden expresar como:

El Aprendizaje Electrónico, más conocido por su término inglés e-learning, es una nueva modalidad de formación a distancia que aplica los procesos educativos y de enseñanza mediante medios electrónicos, como Internet, permitiendo al estudiante recibir y completar sus estudios y cursos con el soporte de aplicaciones o herramientas informáticas.

Las ventajas que motivan la elección del Aprendizaje Electrónico son:

- Optimización del tiempo que se dedica a la formación a distancia.
- Nuevas formas y procesos de aprendizaje.
- El uso de los recursos en línea según conveniencia, permite sacar el máximo partido a los resultados usando todos los dispositivos electrónicos disponibles para el estudiante.
- Los cursos permiten conocer a otros estudiantes con los mismos intereses y romper con la costumbre de una educación local y focalizada.

2.2.2 MOOC

Los MOOC (Massive Open Online Courses)[41][58] son una iniciativa para acercar la educación a cualquier parte del mundo. Ofrecen conocimiento a cambio de que el alumno esté dispuesto a esforzarse. No obliga a estudiar algo que no le interesa y su abanico de posibilidades es tan amplio como la oferta que existe ahora mismo en la red. La educación ha dejado de ser algo de unos pocos y delimitado por fronteras o instituciones, para convertirse en algo global, liberando conocimiento a un público ilimitado.

Es importante conocer que está incluido dentro de la definición de MOOC, pues la enseñanza a distancia que quiere verse amparada por este término deben de cumplir algunos requisitos:

- Estructura organizada de un curso, exponer contenido por temáticas o similitudes, incluir actividades y problemas para que pongan a prueba lo que han aprendido,

además de evaluaciones finales que acrediten el grado de interiorización del curso para cada estudiante.

- Los MOOC son masivos y sin restricciones, los posibles matriculados son, salvo restricciones del hardware de la plataforma, ilimitados.
- En cuanto a su carácter online, no existen restricciones de fronteras, todo el contenido esta en Internet, siendo este el principal medio de comunicación y unión entre todos los participantes.
- No existen limitaciones o restricciones, su contenido es abierto y de libre acceso.

Algunas de las mejoras de los cursos MOOC con respecto a la educación tradicional son:

- La cantidad de estudiantes por curso puede superar, en una sola edición, el total de alumnos que un profesor puede tener en toda su carrera profesional.
- La duración de los cursos no tiene por qué ser igual, sino que se amolda al contenido impartido.
- El coste inicial de producción de un curso es elevado, pero puede ser impartido ilimitadas veces sin coste adicional o por un coste mínimo actualizando los contenidos.
- No existe un horario lectivo, es el estudiante el que decide en qué momento puede realizar el curso en línea.

Actualmente, la Universidad Autónoma de Madrid está llevando a cabo también una iniciativa con otro tipo de cursos online: los llamados SPOCs *(Small Private Online Course)*, que se utilizan a nivel local con los estudiantes en el ámbito académico en el contexto de proyectos de innovación docente.

2.2.3 Analíticas de Aprendizaje (Learning Analytics)

La nueva modalidad de Aprendizaje Electrónico está teniendo mucho éxito. Por lo que numerosas instituciones quieren incorporarlo en sus procesos educativos y de formación profesional. Como se ha visto anteriormente, esta tendencia tiene un gran número de ventajas que hacen que este ganando importancia.

Como consecuencia de la implantación de los cursos MOOC, se están generando una gran cantidad de datos que son registrados a partir de las actividades de los estudiantes. Las analíticas de aprendizaje tienen como objetivo el recopilar, analizar y medir los datos que producen los estudiantes en los cursos en línea. [4]

Los métodos de análisis más comunes se apoyan en el contenido del curso y en las interacciones entre los propios alumnos. Recordando la frase de Peter Drucker, "*Todo lo que se puede medir se puede mejorar*", por ello si queremos mejorar necesitamos analizar los datos.

El análisis de los datos procedentes de los cursos MOOC es de gran valor para profesores e instituciones. Su estudio permitirá a estas entidades asignar los recursos de manera más eficiente, personalizar o adaptar los contenidos a medida para los estudiantes, monitorizar el proceso de los estudiantes durante los cursos y, el objetivo de este proyecto, identificar a los estudiantes de un curso que puedan estar en riesgo de abandono que conlleva el posible fracaso o éxito de una formación online masiva. [50]

2.3 Estudio de tecnologías

En esta sección se estudian los dos modelos de base de datos que podrían ser usados en el Trabajo de Fin de Grado y se comparan los sistemas de gestión de bases de datos más utilizados. Además de analizar los lenguajes que ofrecerían una mayor versatibilidad para el desarrollo del sistema.

2.3.1 Modelos de BBDD

Los sistemas de base de datos se engloban en dos categorías: relacionales y no relacionales. Sin embargo, los sistemas más usados en cualquier proyecto de ingeniería de software son los motores de bases de datos con una arquitectura relacional y, todos ellos, utilizan un lenguaje de consulta basado en SQL, con pequeñas variaciones según la empresa.

2.3.1.1 Base de datos relacionales

Las bases de datos que adoptan el modelo relacional permiten establecer interconexiones o relaciones entre los datos almacenados y, a través de dichas conexiones, normalizar la información almacenada y escalarla según necesidad. [51]

Los sistemas de gestión de bases de datos más comunes y que copan la cuota de la mayoría de los proyectos de ingeniería de software son: MySQL [22], SQLite [33], PostgreSQL [29] y Oracle 9.11 [27].

2.3.1.2 Base de datos no relacionales

Las bases de datos no relacionales, más conocidos como NoSQL [17] [53] acrónimo de los términos ingleses "Not Only SQL", tienen como objetivo proponer una estructura de almacenamiento más versátil, esta ventaja implica perder ciertas funcionalidades como realizar operaciones en varias colecciones de datos teniendo que recurrir a la desnormalización de los datos almacenados.

Algunas de las implementaciones NoSQL más conocidas son: CouchDB [36], MongoDB [21] y Cassandra [35].

2.3.1.1 Comparativa de modelos de base de datos

A continuación se explican las ventajas y desventajas de ambos modelos:

Tipo de base de datos	Ventajas	Desventajas
Relacional	Evita duplicidad de datos. Garantiza la integridad referencial. Favorece la normalización para ser más comprensible y aplicable.	Presentan deficiencias con datos gráficos y multimedia. Escalabilidad. Grandes cantidades de registros produce un mal rendimiento.
No Relacional	Permite estructuras distribuidas. Mayor adaptabilidad a cambios relacionados con el proyecto. Optimizado para grandes cantidades de datos.	No se puede garantizar la unicidad de los datos, datos duplicados. No son compatibles al 100% con bases de datos SQL.

Tabla 2-1: ELASA – Comparativa Bases de datos relacionales y no relacionales

2.3.2 Lenguajes de Desarrollo

Dentro de las tecnologías de software que se pueden aplicar en cada modelo, estas son las que han sido sometidas a estudio.

Python [30], es un lenguaje de scripting independiente de plataforma y orientado a objetos. Es un lenguaje interpretado, lo que significa que no se necesita compilar el código fuente para poder ejecutarlo. Esto ofrece ventajas, como la rapidez de desarrollo, e inconvenientes, como una menor velocidad de ejecución.

Características:

- Propósito general, aunque al comienzo Python no fue orientado a la creación de contenido web, si existen páginas web desarrolladas íntegramente en Python, como por ejemplo Zope [72].
- Multiplataforma, existen numerosas versionas que se pueden ejecutar en muchos sistemas informáticos.
- Interpretado, no es necesario compilar su código antes de ejecutarlo.
- Interactivo, dispone de un intérprete por línea de comandos.
- Orientado a objetos, permite la creación de contenido reutilizable y escalable.
- Sintaxis clara, el lenguaje tiene como requisito obligatorio la indentación del código.

Java [26] se trata de un lenguaje de programación orientado a objetos, cuyo objetivo es poder ser ejecutado en cualquier plataforma sin importar su arquitectura ni la necesidad de reescribir el código usando una Máquina Virtual de Java [24]. El lenguaje Java se creó bajo una serie de requisitos básicos:

- Debía de estar bajo el paradigma de la programación orientada a objetos.
- Permitir la ejecución de cualquier programa en múltiples arquitecturas y sistemas operativos.
- Por defecto debía de incluir soporte para desarrollos en red.
- Permitir un diseño estable para ejecutar aplicaciones de forma segura en sistemas remotos.
- Fácil de usar, mantener y escalar.

Estas dos tecnologías también pueden ser usadas en desarrollos web haciendo uso de frameworks [70] y otras librerías que permiten combinar estos lenguajes con los desarrollos web.

2.4 Métodos de predicción

La minería de datos es un tema cada vez más habitual en los entornos académicos y empresariales. Este concepto hace referencia al proceso de selección y recolección de datos con el fin de analizarlos y encontrar patrones de comportamiento que no se reconocen a simple vista para así poder sacar algún beneficio y mejorar los procesos.

Los métodos predictivos [20] en minería de datos se basan en entrenar a un modelo o algoritmo con un conjunto aleatorio de datos reales. Esto nos permite obtener una variable o indicador que se aplicará conjuntamente al modelo predictivo sobre los datos restantes.

Estos métodos predictivos, que necesitan de variables entrenadas para poder obtener resultados, son denominados métodos asimétricos, supervisados o directos. Su funcionamiento más común es el de identificar patrones o indicadores para clasificar los datos de los grupos de entrenamiento para, posteriormente, aplicar el indicador o múltiples de ellos sobre los conjuntos de datos a clasificar.

Las variables que sirven para entrenar el modelo predictivo se denominan **predictores**, mientras que el resto de las variables son **datos a predecir**. Los algoritmos más comunes aplicados a los métodos predictivos son las redes neuronales, árboles de decisión o clasificación según una función de similitud.

Las **redes neuronales** son métodos de proceso numérico en el que las variables interactúan mediante transformaciones con funciones lineales o no lineales, hasta obtener una salida en forma de coeficiente. Estas salidas se contrastan con los que tenían que haberse obtenido, basándose en unos datos de prueba, dando lugar a un proceso de retroalimentación mediante el cual la red se reconfigura y aprende, hasta obtener un modelo adecuado.

Un **árbol de decisión** es un diagrama de flujo, con estructura de árbol. Se divide en nodos y hojas. Los nodos son las comparaciones entre los indicadores y/o variables y el dato a clasificar. Las hojas son las clases en las que se puede clasificar un dato. La inducción de árboles de decisión suele ser muy habitual para solucionar problemas de clasificación de datos.

Las técnicas de **agrupación en clúster por similitud** son usadas cuando la variable de destino o respuesta no sea importante o no esté disponible. Como su nombre indica, las técnicas de agrupación en clúster son capaces de agrupar los datos de entrada dependiendo de su similitud con datos de un conjunto de entrenamiento, previamente seleccionado, creando una clasificación basada en datos reales [16]. Los algoritmos de similitud que se han sopesado para su implementación en el sistema has sido KNN [57], clasificador Rocchio [59] y Naïve Bayes [61]. En el caso de caso de la implementación de KNN se haría uso de la función que se muestra en la *Figura 2-1*.

$$p(w|c) \sim \frac{\sum_{d \in c} frec(w,d) + 1}{\sum_{d \in c} |d| + |\mathcal{V}|}$$

Figura 2-1: ELASA - NN Función matemática

También se implementará el algoritmo de minería de datos Tf-Idf [56] que genera un coeficiente en función de la frecuencia de un término por cada documento. Realizando una implementación que equipare a los estudiantes como documentos y a los eventos como términos se pretende obtener qué eventos son más relevantes para los estudiantes con mejores notas.

Véase la sección 4.3.3 para más detalles sobre la implementación de estos algoritmos.

Los resultados de las predicciones con indicadores o variables múltiples nos otorgarán información sobre que variables de entrada tienen mayor importancia y cuáles de ellos hacen que el sistema sobre-aprenda. [15]

2.5 Otras aplicaciones relacionadas

En la primera fase del proyecto, se realizó un trabajo previo de investigación sobre las herramientas ya desarrolladas en esta área y que pueden servir de motivación a la hora de desarrollar el nuevo sistema.

Gran parte del software que se utiliza actualmente para el aprendizaje de análisis duplica la funcionalidad del software de análisis web, aplicándolo a la interacción del alumno con el contenido. Las herramientas de análisis de redes sociales se utilizan comúnmente para mapear las conexiones sociales y discusiones.

Algunos ejemplos de herramientas de cuadros de mando y análisis de datos relacionados con los cursos MOOC son:

La herramienta **Analyse** [46] es una aplicación desarrollada por el laboratorio Gradient [47] por la Universidad Carlos III, que tiene como funcionalidad la de realizar análisis y visualizar los datos mediante cuadros de mando (Dashboards) sobre el aprendizaje en línea de los estudiantes que usan la plataforma Open edX [28].

Una aplicación de la propia Universidad Autónoma de Madrid es **Open DLAs** [31], se trata de una herramienta que muestra mediante un cuadro de mando educativo, basado en analíticas de aprendizaje, presente la información generada en los cursos MOOC tanto a instructores como a administradores, con el objetivo de poder identificar qué formato de curso recibe mejor captación en este entorno online masivo y realizar un seguimiento del progreso académico de los estudiantes durante la duración del curso.

Student Success System [5] es una herramienta de análisis predictivo del aprendizaje, que predice la implicación de los estudiantes y los categoriza en cuadrantes en función del riesgo de compromiso con el curso y el rendimiento. La herramienta genera indicadores que ayudan a visualizar si el alumno tiene problemas en el apartado social interactuación con otros estudiantes, el rendimiento de las evaluaciones y problemas con los contenidos del curso.

SNAPP [32], se trata de una herramienta de análisis de aprendizaje que visualiza el grafo resultante de las interacciones de los estudiantes con sus mensajes en los foros de discusión y como se responden entre ellos.

LOCO-Analyst [18] se trata de una herramienta de aprendizaje sensible al contexto para el análisis de los procesos que tienen lugar en un ambiente de aprendizaje basado en la web. Actualmente está siendo usada sobre los cursos realizado con los Sistemas de Gestión de contenido para el Aprendizaje, siglas en ingles LCMS [7] [52], desarrolladas por la Universidad de Saskatchewan en Saskaton, Canadá [48].

Estas aplicaciones pueden categorizarse según su funcionalidad. Aquellas que solo muestren datos agrupados al usuario se denominan cuadros de control o dashboards, otras están más involucradas en el análisis de los cursos MOOC y las relaciones sociales con los estudiantes y finalmente se encuentran aplicaciones que pretenden analizar sacar indicadores y predecir el comportamiento de los estudiantes en los cursos. Se puede ver la clasificación de cada una de las herramientas expuestas en esta sección en la *Figura 2-2*.

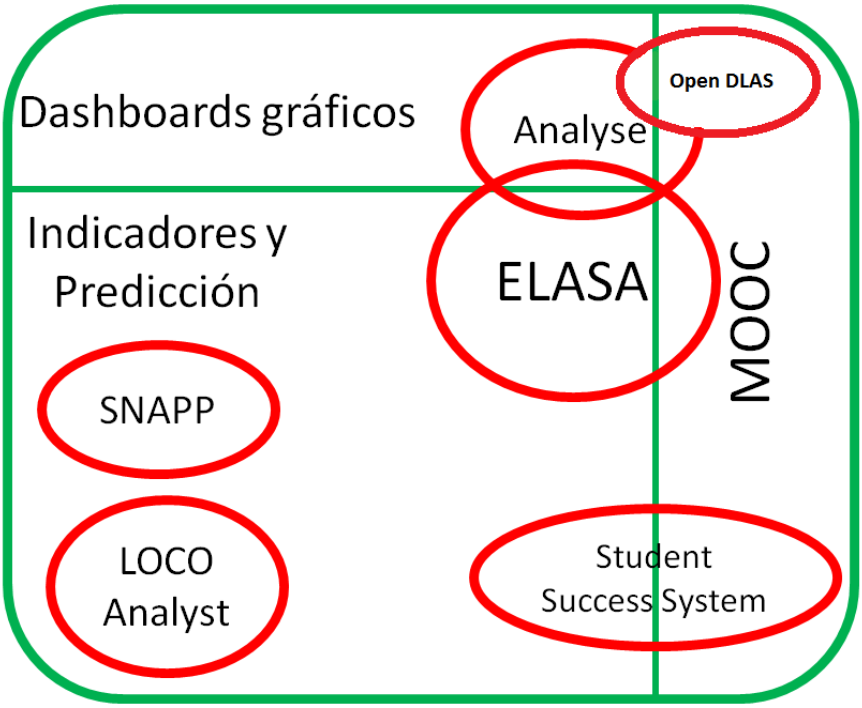

Figura 2-2: ELASA - Herramientas y su clasificación según funcionalidad

3 Análisis de Requisitos y Diseño

3.1 Análisis de Requisitos

En esta sección se presentan los requisitos funcionales y no funcionales para el sistema propuesto en este documento. Es importante que se definan lo mejor posible los requisitos y los casos de uso para conseguir que el sistema sea robusto y se pueda escalar en caso de continuar con su desarrollo.

3.1.1 Requisitos Funcionales:

Los requisitos que aquí se detallan son el resultado de un análisis previo sobre qué necesidades se debían de cubrir con el sistema, dejando la posibilidad de incorporar nuevas necesidades que se puedan detectar al realizar las pruebas de validación con datos reales.

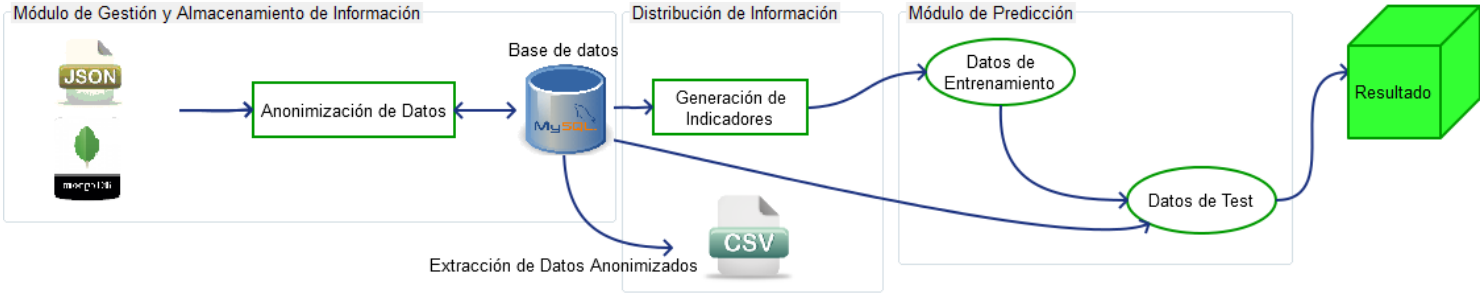

Figura 3-1: ELASA - Propuesta de diseño de los módulos

3.1.1.1 Módulo de Gestión y Almacenamiento de Información

3.1.1.1.1 Sub-modulo Gestor de Base de Datos

RF 1 Introducir un nuevo usuario de control a la herramienta.

RF 1.1 Para introducir un nuevo usuario en la base de datos será necesario introducir un nombre de usuario que no se encuentre ya registrado. Cada usuario dado de alta en el sistema deberá de ir acompañado por una contraseña

RF 2 Añadir el identificador de un nuevo curso MOOC

RF 2.1 El sistema tiene que permitir introducir tantos identificadores para cursos MOOC como se quiera

RF 2.2 Cada curso ha de tener al menos un identificador que esté asociada a la edición de un curso MOOC

RF 3 Añadir un nuevo identificador de una edición para un curso ya existente

RF 3.1 La base de datos ha de estar preparada para recibir tantos identificadores de ediciones para un curso como se quiera, pero nunca podrá existir una combinación de identificadores curso-edición duplicada

RF 4 Introducir la relación entre el nombre de usuario del estudiante y un identificador único

RF 4.1 Este identificador único para los estudiantes será el que se reciba de los ficheros de dato de la plataforma edX

RF 5 Actualizar datos de estudiantes para un curso-edición

RF 5.1 En caso de que un estudiante se cambie de nombre de usuario, será necesario actualizar esa información para su correcta identificación en los ficheros que se reciban de eventos, foros y certificados

RF 5.2 No será necesario actualizar los registros anteriores del estudiante ya que estos se guardan en la base de datos según el identificador único del estudiante creado previamente y que no se puede cambiar

RF 5.3 En caso de que el identificador de estudiante no se encuentre en la base de datos, se dará de alta un nuevo registro con el nombre de usuario del estudiante y el identificador del fichero de datos, *RF4*

RF 6 Añadir datos demográficos de los estudiantes

RF 6.1 Para introducir los datos demográficos, el estudiante ya debe de haber sido incluido en la base de datos

RF 7 Actualizar los datos demográficos de los estudiantes

RF 7.1 Para actualizar los datos demográficos, el estudiante ya debe de haber sido incluido en la base de datos

RF 7.2 En caso de que los datos demográficos del estudiante no hayan sido introducidos estos no se actualizarán, sino que se crearán un nuevo registro asociado al identificador del estudiante, *RF6*

RF 7.3 En caso de que el identificador de estudiante no se encuentre en la base de datos, se dará de alta un nuevo registro con el nombre de usuario del estudiante y el identificador del fichero de datos, *RF4*

RF 8 Añadir los datos de matriculación del estudiante al curso MOOC

RF 8.1 Para introducir los datos de matriculación del estudiante, ya debe de existir un registro previo con el identificador de estudiante

RF 9 Actualizar los datos de registro a la edición del curso de los estudiantes

RF 9.1 Para introducir los datos de matriculación de los estudiantes, ya deben de existir un registro previo con el identificador de cada uno de ellos

RF 9.2 En caso de que el identificador de estudiante no se encuentre en la base de datos, se dará de alta un nuevo registro con el nombre de usuario del estudiante y el identificador del fichero de datos, *RF4*

RF 10 Introducir los certificados obtenidos por los estudiantes al finalizar una edición de un curso MOOC

RF 10.1 Solo se introducirán los certificados de los estudiantes que cuenten con un registro activo en la base de datos

RF 10.2 Durante la fase de creación de registros para los certificados se usará el identificador de cada estudiante para mantener la privacidad en cuanto a la nota obtenida por el mismo

RF 10.3 En caso de que el identificador del estudiante no se encuentre en la base de datos, se dará de alta un nuevo registro con el nombre de usuario del estudiante y el identificador del fichero de datos, *RF4*

RF 11 Introducir los datos relacionados con la actividad social de los foros de cada edición de un curso MOOC

RF 11.1 Solo los identificadores de los mensajes del foro serán introducidos en la base de datos. Como requisito previo, debe de existir un estudiante registrado en la base de datos

RF 11.2 El contenido de los mensajes será descartado, pero se mantendrá la relación de las respuestas entre estudiantes para analizar la interacción entre ellos

RF 11.3 En caso de que el identificador del estudiante no se encuentre en la base de datos, se dará de alta un nuevo registro con el nombre de usuario del estudiante y el identificador del fichero de datos, *RF4*

RF 12 Introducir los datos relacionados con los eventos diarios de una edición de un curso MOOC

RF 12.1 Solo los eventos de estudiantes previamente registrados en el sistema serán incluidos en la base de datos

RF 12.2 La información de los eventos nunca contendrá información privada del estudiante o información alguna que pueda violar su privacidad

RF 12.3 En caso de que el identificador de estudiante no se encuentre en la base de datos, se dará de alta un nuevo registro con el nombre de usuario del estudiante y el identificador del fichero de datos, *RF4*

Para obtener más información acerca de los datos gestionados por este sub-modulo véase Anexo B.

3.1.1.1.2 Sub-módulo de Anonimización de Datos

El módulo de anonimización tiene como objetivo eliminar o ignorar la información privada de los estudiantes durante la creación de los registros en la base de datos, de manera que se puedan realizar investigaciones y reportes sin que exista ninguna violación a la privacidad de los participantes en cursos MOOC. La anonimización se ejecuta en el momento que se introducen los datos provenientes de edX en la base de datos.

Como se explica detalladamente en el Anexo B, todos los ficheros recibidos desde la plataforma edX relacionan la información del estudiante con el nombre de usuario elegido al darse de alta en la página web de edX. De manera que, para anonimizar la información de cada estudiante, se sustituirá el nombre del usuario por un identificador único que edX facilita en uno de los ficheros, esto creará un mapeo entre nombre de usuario e identificador único para cada estudiante.

RF 13 Mapeo del nombre de usuario e identificador único

RF 14 Anonimizar el *RF 6* y *RF 7*

RF 14.1 Usando el mapeo del nombre del usuario con el identificador único, *RF 13*, se anonimizarán los datos demográficos de los estudiantes

RF 15 Anonimizar el RF 8 y RF 9

RF 15.1 Usando el mapeo del nombre del usuario con el identificador único, *RF 13*, se anonimizarán los datos de matriculación de los estudiantes

RF 16 Anonimizar el RF 10

RF 16.1 Usando el mapeo del nombre del usuario con el identificador único, *RF 13*, se anonimizarán los datos de las notas obtenidas por los estudiantes y el tipo de certificado expedido para el estudiante

RF 17 Anonimizar el RF 11

RF 17.1 Usando el mapeo del nombre del usuario con el identificador único, *RF 13*, se anonimizarán los mensajes creados en los foros de los cursos MOOC y únicamente se mantendrá la relación resultante de las conversaciones entre estudiantes

RF 18 Anonimizar el RF 12

RF 18.1 Usando el mapeo del nombre del usuario con el identificador único, *RF 13*, se anonimizarán los eventos, manteniendo cuando sea posible, la información del servidor de edX por encima de cualquier información generada por el estudiante que pueda contener datos privados

3.1.1.2 Módulo de Distribución de Información

3.1.1.2.1 Sub-módulo de Extracción de Datos Anonimizados

Este módulo tiene como objetivo la extracción en bruto de los datos anonimizados para la entrega a profesores, centros de investigación u otras entidades.

RF 19 Extraer información relacionada con los eventos de los cursos MOOC

RF 19.1 Comprobar que existe registro del curso MOOC en la base de datos

RF 19.2 Comprobar que han sido cargados datos sobre eventos para el curso MOOC seleccionado

RF 20 Extraer información relacionada con las relaciones entre estudiantes en los foros de los cursos MOOC

RF 20.1 Comprobar que existe registro del curso MOOC en la base de datos

RF 20.2 Comprobar que han sido cargados datos sobre foros para el curso MOOC seleccionado

RF 21 Extraer información relacionada con los certificados de los cursos MOOC

RF 21.1 Comprobar que existe registro del curso MOOC en la base de datos

RF 21.2 Comprobar que han sido cargados datos sobre los certificados para el curso MOOC seleccionado y que el curso ha finalizado

RF 21.3 En caso de que no haya finalizado el curso MOOC la extracción asignará una puntuación de 0 a todos los estudiantes

3.1.1.2.2 Sub-módulo de Generación de Indicadores

La idea general es proponer indicadores o métricas para clasificar a los estudiantes, principalmente que describan la interacción de los mismos con la plataforma. Los indicadores están diseñados de manera que puedan aplicarse a intervalos de tiempo, es decir, se podrán obtener según apliquemos una fecha inicial o final.

Indicadores creados:

RF 22 Actividad social: Medida de la cantidad de post y/o contestaciones puestas en los foros por parte de los estudiantes

RF 23 Constancia: Número de días diferentes que el estudiante ha realizado algún acceso a la plataforma edX. Sería una forma alternativa de medir la constancia del estudiante

RF 24 Número de sesiones realizadas por el estudiante

RF 24.1 Se tomará como sesión toda actividad relacionada con los eventos o actividades en los foros entando dentro de un intervalo de tiempo

RF 25 Número de accesos a recursos diferentes del curso

RF 26 Número de eventos lanzados por el usuario

Funcionalidades comunes a todos los indicadores:

RF 27 Los indicadores deben de permitir que se les aplique filtros por fechas

RF 27.1 Se podrán obtener los indicadores según apliquemos una fecha inicial

RF 27.2 Se podrán obtener los indicadores según apliquemos una fecha final

RF 27.3 Se podrán obtener los indicadores según apliquemos una fecha inicial y final al mismo tiempo

3.1.1.3 Modulo Modelos de Predicción

El sistema implementará algoritmos para obtener información acerca el absentismo estudiantil en los cursos MOOC, estos algoritmos tienen como objetivo ayudar a los equipos docentes a detectar problemas en los cursos y, así tomar medidas para paliar este abandono.

Los modelos de predicción que se han creado para este módulo son:

RF 28 Modelo Vectorial por Tf-Idf [56]

RF 28.1 Este modelo hace uso del indicador de cantidad de eventos, *RF 26*

RF 29 Modelos de similitud: algoritmo de KNN [57]

RF 29.1 Este modelo hace uso del indicador de constancia, *RF 23*

RF 29.2 Este modelo hace uso del indicador de número de eventos por estudiantes, *RF 26*

RF 29.3 Este modelo hace uso del indicador de número sesiones por cada estudiante, *RF 24*

RF 30 Modelos de Asortatividad [55]

RF 30.1 Este modelo hace uso del indicador de actividad social por estudiante, *RF 22*

3.1.2 Requisitos no Funcionales

Hacer referencia a:

RNF 1 Interfaz

RNF 1.1 La interfaz ha de mostrar las instrucciones claras e intuitivas

RNF 1.2 El sistema deberá de ser capaz de completar toda funcionalidad implementada sin asistencia externa

RNF 1.3 El sistema debe de ser fácil de aprender y que no requiera conocimientos informáticos avanzados para su uso

RNF 1.4 El sistema debe tener la portabilidad para ser ejecutado en cualquier sistema operativo que acepte la instalación de una Máquina Virtual de Java.

RNF 2 Rendimiento en tiempo real [42]

RNF 2.1 Obtener un buen rendimiento para sentencias de Definición de Datos (**DDL**). Este requerimiento no es prioritario, debido a que la estructura de la base de

datos ya ha sido definida. En caso de una modificación en una tabla, índice, etc. el rendimiento será bueno pero no optimizado, ya que no es una actividad que se vaya a producir de manera recurrente.

RNF 2.2 Obtener un buen rendimiento para sentencias de Manipulación de Datos (**DML**). Estas sentencias son muy comunes en el módulo de la base de datos, las sentencias denominadas INSERT y UPDATE, son acciones que se ejecutarán miles de veces al día, además de que el tamaño de la base de datos no dejara de crecer y la obtención de información mediante sentencia SELECT será más costoso cada día. Por ello, se ha escogido el esquema InnoDB [23] para las tablas de la base de datos de manera que se optimice el rendimiento para estas instrucciones.

RNF 3 Escalabilidad

RNF 3.1 La estructura de la base de datos se ha definido según el esquema InnoDB. Este esquema permite restricciones como por ejemplo, FOREIGN_KEYS, que facilita la definición de nuevas tablas y la creación de relaciones con otras ya existentes.

RNF 3.2 En caso de cambio en la estructura de ficheros explicada en el Anexo A, los módulos de la aplicación deben de estar lo suficientemente aislados para que los cambios en el código fuente tengan un impacto mínimo.

RNF 4 Seguridad

RNF 4.1 La aplicación deberá de incluir un sistema de autentificación para acceder a cualquier funcionalidad de la herramienta.

RNF 4.2 La base de datos deberá de tener un usuario y contraseña conocidos solo por los perfiles administradores.

RNF 5 Privacidad de datos

RNF 5.1 Los datos que se tienen que eliminar u obviar durante el proceso de anonimización, son todos aquellos que contengan información personal de los estudiantes así como las respuestas a ejercicios y el contenido de los mensajes de los foros.

3.2 Diseño

Como se ha explicado en el apartado anterior, los requisitos funcionales se han clasificado en módulos, por lo que se ha seguido la misma tendencia para el diseño del sistema. En este apartado se dará una descripción más detallada de la estructura modular ideada:

- Módulo de Gestión y Almacenamiento de Información
  - Sub-módulo de Gestión de Base de Datos
  - Sub-módulo de Anonimización de Datos
- Módulo de Distribución de Información
  - Sub-módulo de Extracción de Datos Anonimizados
  - Sub-módulo de Generación de Indicadores
- Módulo Modelos de Predicción

3.2.1 Entorno de desarrollo

Para el entorno de desarrollo se ha utilizado XAMPP [1], un servidor de plataforma libre que integra en una sola aplicación:

- ✓ Servidor Web Apache[37]

- ✓ Servidor FTP FileZilla [12]
- ✓ Servidor de base de datos MySQL [22]
- ✓ Intérprete de lenguaje de scripts PHP [40]

Para la implementación del código se ha usado el entorno de programación NetBeans IDE 8.0.2 [25], con el lenguaje de programación Java ya que permite usar sistemas de repositorios como Github [13], Subversion [34] y Mercurial [19].

Adicionalmente, se va a explicar la arquitectura lógica resultante del diseño del sistema además de profundizar en el diseño de la base de datos, explicando todas sus características y restricciones. Los ficheros que se han identificado como posible entrada de datos tendrían los formatos CSV [62], JSON [66] y MongoDB.

En la *Figura 3-2* se muestra el esquema de la arquitectura lógica:

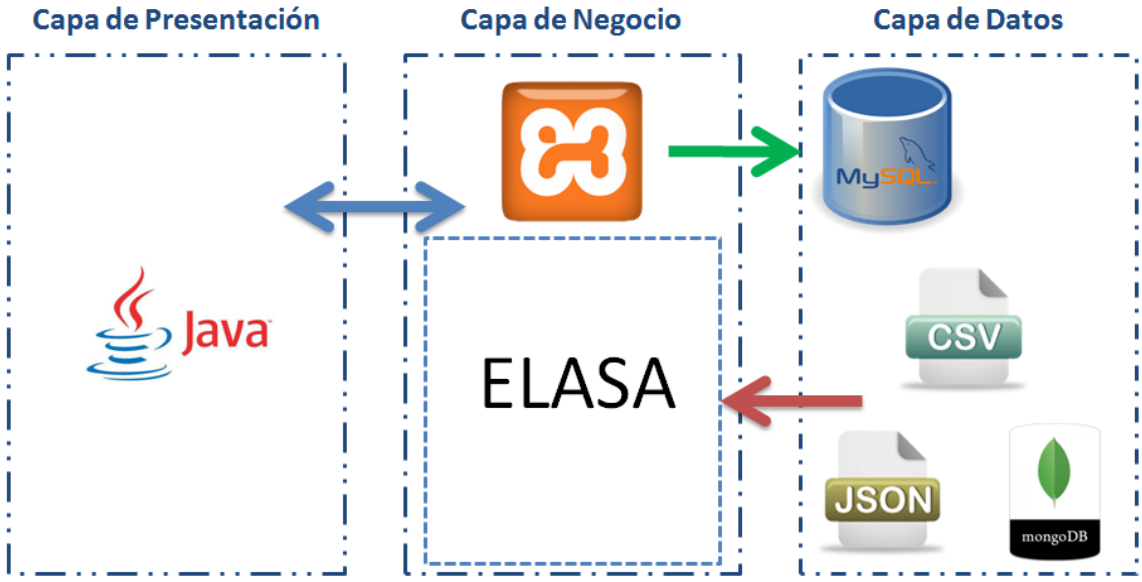

Figura 3-2: ELASA - Arquitectura lógica del sistema

3.2.2 Arquitectura lógica

3.2.2.1 Capa de presentación

La capa de presentación es aquella con la que interactúa el usuario que hace uso del sistema. Se encarga de presentar la información al usuario y de capturar la interacción que éste realiza a través de la interfaz creada. Toda acción por parte del usuario se comunica a la capa de negocio, generando un proceso interno acorde con la información recibida desde la capa de presentación.

3.2.2.2 Capa de negocio

La capa de negocio es la sección que contiene y gestiona la lógica del sistema. Se comunica con la capa de presentación mostrando la interfaz con todas las opciones que podrán ser usadas por los usuarios y con la capa de datos para recibir ficheros de datos en bruto, almacenar en la base de datos información y extraer de la base de datos información.

Esta capa contiene los ficheros Java que gestionan la conexión con la base de datos, el tratamiento de los ficheros que se reciban desde la capa de datos y se aplicarán las peticiones o acciones que el usuario envié desde la capa de presentación. También se considera que el software XAMPP, que gestiona la configuración de predeterminada con MySQL y Apache, está incluido en esta capa.

3.2.2.3 Capa de datos

La capa de datos es la responsable de almacenar los datos de los estudiantes, recibir ficheros y extraer cualquier información que el sistema solicite. Recibe peticiones desde la capa de negocio y envía las respuestas resultantes.

Esta capa también englobará la funcionalidad de recibir los ficheros con los datos de los cursos MOOC en formato SQL, JSON y MongoDB. Además, por petición de la capa de presentación, la capa de negocio puede generar ficheros CSV [62] que contienen extracciones de datos desde la base de datos para los usuarios.

3.3 Modelo de datos

A continuación se da una breve descripción de cada uno de los contenedores que se muestran en la *Figura 3-3*. En la base de datos los contenedores que almacenan los datos se denominan tablas, de ahora en adelante usaremos ese término.

- Tabla elasa_user_ctrl: esta tabla almacenará los registros de los usuarios que podrán acceder a la base de datos del sistema. El nombre de usuario tiene que ser único y la contraseña esta codificada por la función SHA1 [68].

- Tabla elasa_user_map: esta tabla contiene los identificadores únicos obtenidos de los ficheros de edX y la relación con los nombre de usuarios de los estudiantes en los cursos MOOC. Los nombres de usuarios son almacenados después de que se les aplique la funcionalidad de encriptación con la función SHA1.

- Tabla elasa_courses: esta tabla contiene los nombres de los cursos en el sistema de edX y los códigos asignados por el sistema a cada edición. A cada relación curso-edición se le asignará un valor numérico único que servirá para identificar los registros de las distintas tablas con el curso-edición al que pertenecen.

- Tabla elasa_students: en esta tabla se almacenan los datos de matriculación en el sistema por parte de los estudiantes. Para evitar los datos privados se elimina cualquier información que pueda contener datos personales y se usará el identificador privado facilitado por edX.

- Tabla elasa_profile: esta tabla guardará los datos demográficos, de género y de edad para realizar agrupaciones en los registros según grupos sociales y filtrar los datos que se extraigan a petición del investigador. Para evitar los datos privados estos pasan por el proceso de anonimización.

- Tabla elasa_events: esta tabla contendrá la información anonimizada de los eventos. Cada evento estará conectado con un identificador de una edición de un curso por la clave primaria de la tabla y con el identificador del propio estudiante.

- Tabla elasa_social: esta tabla contendrá la información anonimizada de los mensajes que se escriban en el foro. Cada mensaje tiene un identificador único que usaremos a modo de clave primaria y única para mantener la relación de respuestas entre estudiantes. Además, cada mensaje tendrá un valor numérico para indicar a que curso-edición pertenece y el identificador del estudiante para almacenar el creador del mensaje.

- Tabla elasa_certificates: esta tabla solo contendrá información sobre las notas de los estudiantes de los cursos-edición finalizados. En caso de intentar almacenar cursos-ediciones no finalizados, el sistema guardará una calificación de 0 a cada estudiante matriculado. Cada certificado estará relacionado con un curso-edición con un valor numérico para identificar a donde pertenece y con un identificador del estudiante. Además, cada certificado tendrá un dato que indica si el certificado expedido ha sido de la modalidad verificada o básica.

En la *Figura 3-3* se representa el modelo de base de datos propuesto para el proyecto:

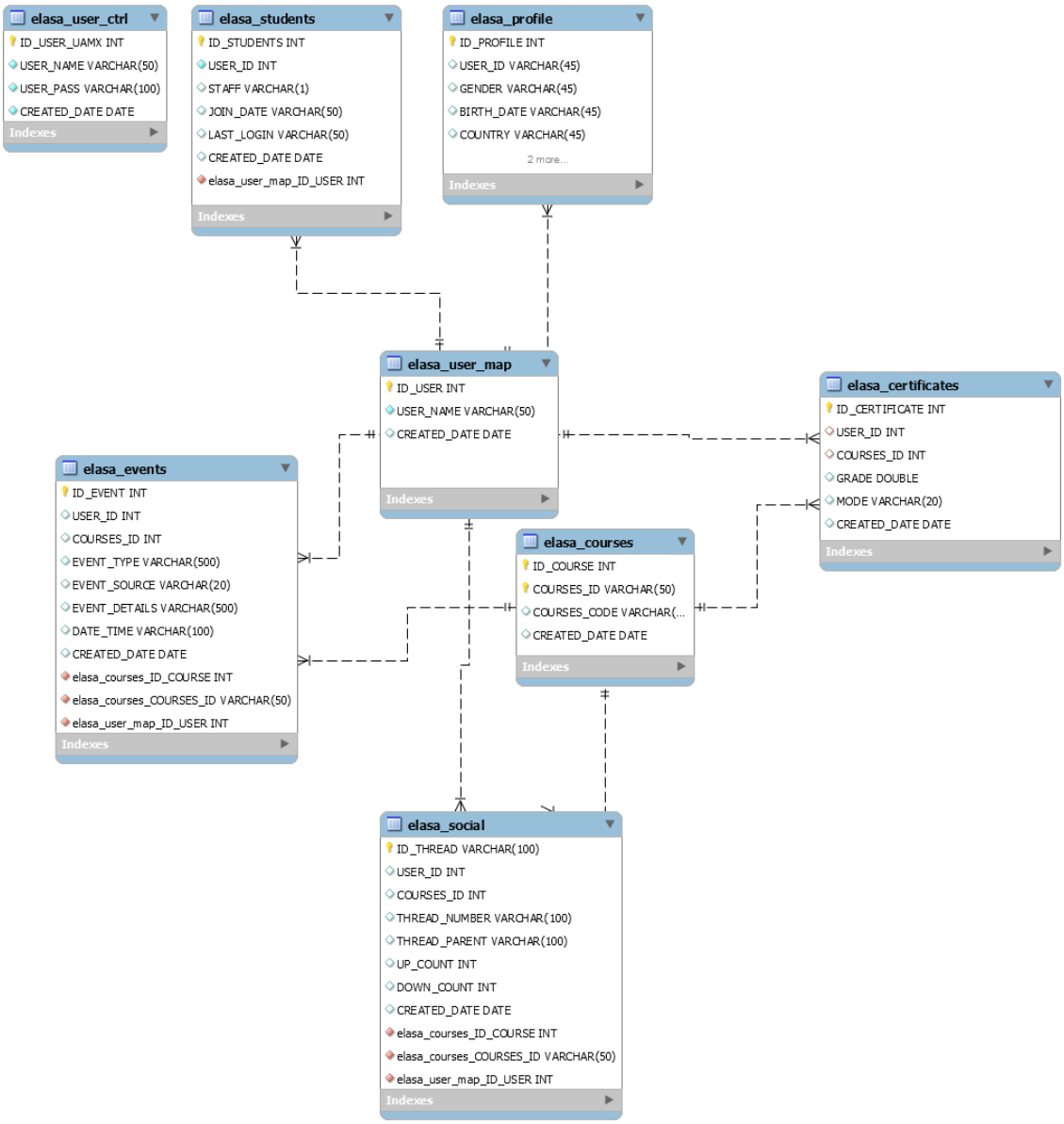

Figura 3-3: ELASA - Modelo de Base de Datos

4 Desarrollo

En este capítulo se detalla la fase de desarrollo del sistema ELASA siguiendo lo anteriormente explicado para los requisitos funcionales y el diseño modular. Se explicará la metodología seguida en la implementación del sistema y que ficheros serían necesarios para cumplir con cada requerimiento expuesto en el capítulo 3. También se detalla en este capítulo el procedimiento de analíticas de aprendizaje y los algoritmos de predicción implementados para el proyecto.

4.1 Ciclo de vida

Para garantizar el éxito de cualquier proyecto de ingeniería de software, es necesario analizar los requerimientos y las fases necesarias para llevarlo a cabo, es decir, buscar un ciclo de vida que se adapte nuestras necesidades y a las distintas fases por las que va a pasar el desarrollo.

El objetivo del TFG es claro, mientras que los requerimientos se podrán ampliar y actualizar según avanza el desarrollo del mismo, es decir, se podría dar el caso de incluir requisitos funcionales nuevos durante la vida del mismo. Por este motivo, lo más conveniente fue seleccionar una metodología de trabajo Ágil [6] con un ciclo de vida iterativo e incremental.

Se ha elegido este ciclo de vida debido a que en cada iteración se obtiene un producto ejecutable que cubre las nuevas funcionalidades añadidas y sobre las que se realizan las pruebas correspondientes. Tras esto se comienza una nueva iteración del proceso analizando los nuevos requerimientos y solucionando los errores que se puedan haber encontrado.

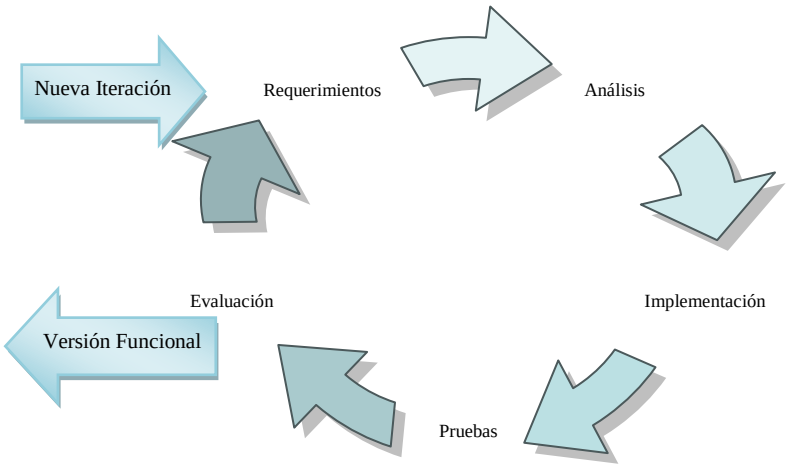

Figura 4-1: Ciclo de vida

4.2 Tecnologías y lenguajes empleados en el desarrollo

Al tratarse de un sistema cerrado sin necesidad de conexión a internet, ya que solo el personal autorizado tendría acceso a los datos, y para evitar accesos indeseados a los datos privados del estudiante, el lenguaje de programación elegido ha sido Java [26]. Se trata de un lenguaje multiplataforma, estable y que permite gran escalabilidad de sus

funcionalidades si el diseño es modular. Por ello y dado que los ficheros que facilita la plataforma edX [28] están sujetos a cambios sin previo aviso a las instituciones, se ha elegido este paradigma de programación para el desarrollo del proyecto.

Además Java cuenta con una gran cantidad de librerías con funcionalidades adicionales y todas las clases que se puedan crear en este proyectos serían completamente adaptables a un sistema online si fuese necesario, ya sea haciendo uso de J2EE [22] o implementando esta funcionalidad con otros frameworks.

Todo el sistema se ha ideado para que la aplicación sea ejecutada en un ordenador local donde está alojada la base de datos. La base de datos está gestionada por un software que integra de MySQL [22] y Apache [37], además de otras funcionalidades, llamado XAMPP [41].

4.3 Descripción de módulos

4.3.1 Gestión y Almacenamiento de Información

4.3.1.1 Gestión de Base de Datos

El sub-módulo de Gestión de Base de Datos es el encargado de toda la funcionalidad relacionada con las tablas y la interacción con los registros. Estos ficheros de dividen entre los paquetes *es.uam.tfg.elasa.structs* y *es.uam.tfg.elasa.bbdd*.

El primer paquete contiene las clases donde se almacenan los datos en estructuras que imitan las columnas de las tablas de la base de datos, de esta manera se cumple con el objetivo del RNF3, si fuera necesario incluir una nueva columna en una de las tablas, el impacto sería mínimo pues solo haría falta incluir un nuevo atributo en la estructura correspondiente en esos ficheros Java. Se puede ver las estructuras de las clases incluidas en el paquete en la *Figura 4-2*.

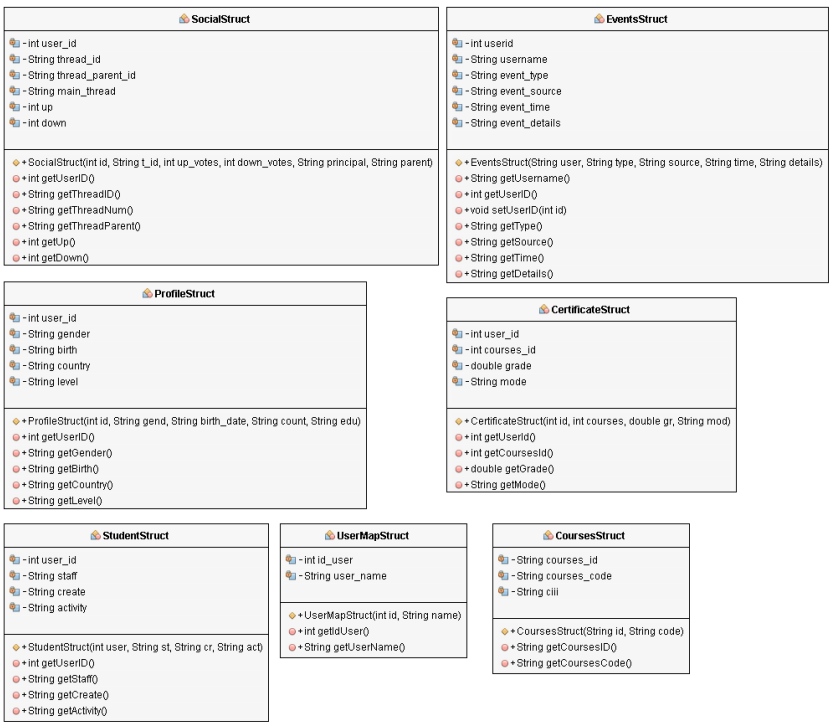

Figura 4-2: ELASA - Paquete *es.uam.tfg.elasa.structs*

El segundo paquete del sistema contiene los ficheros encargados de realizar todas las instrucciones ya sean DDL o DML [51] de cada tabla, *Figura 3-3*. Además de los ficheros para gestionar cada tabla de la base de datos, también se ha creado una clase Java que se encargará de establecer la conexión con la base de datos y autenticarse contra la tabla **elasa_user_ctrl** para permitir el acceso a las funcionalidades del sistema. Se puede ver las clases que gestionan cada tabla definida en la base de datos en la *Figura 4-3*.

Esta clase Java tratará de conectar con la base de datos según lo configurado en las preferencias del servidor Apache. Si la información de la configuración es correcta pero no encuentra la base de datos, el sistema supone una primera conexión en un entorno de trabajo y creará toda la estructura de la base de datos automáticamente. Se puede ver la estructura de las clases en la *Figura 4-2*.

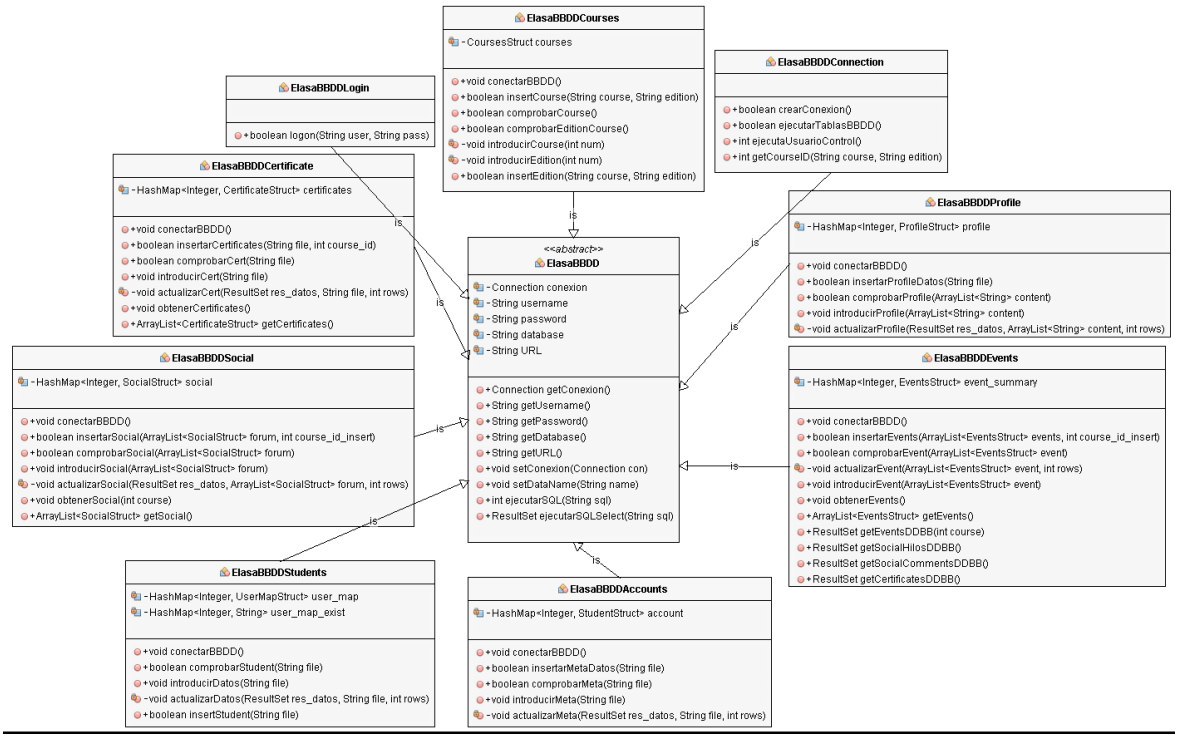

Figura 4-3: ELASA - Paquete *es.uam.tfg.elasa.bbdd*

4.3.1.2 Anonimización de Datos

La funcionalidad de anonimización de la información de los estudiantes no está ligada a ninguna acción por parte del usuario del sistema, sino que son clases java que están vinculadas a los métodos de lectura de los ficheros de datos. Se puede ver la estructura de las clases en la *Figura 4-4*.

Como se explica en el Anexo A, los paquetes de datos recibidos desde la plataforma edX son muy grandes y tienen mucha información que es necesario obviar y eliminar manteniendo una actitud ética y correcta hacía los estudiantes.

La funcionalidad de anonimización de datos se compone de dos paquetes *es.uam.tfg.elasa.parse* y *es.uam.tfg.elasa.security*, estos paquetes tienen un objetivo común, aunque funcionan de manera independiente, sobre los datos.

El paquete *es.uam.tfg.elasa.security* se compone de una clase que codifica, según la función SHA1 [68], la información que se pase como parámetro. La información más

básica que codifica es la contraseña de los usuarios que van a acceder al sistema ELASA. También es usado para codificar el nombre de usuario de los estudiantes de los cursos MOOC, de manera que no sea posible relacionar las acciones de un estudiante a través de su nombre de usuario.

Las funcionalidades del paquete *es.uam.tfg.elasa.parse* están asociadas a la lectura de los ficheros más complejos que se reciben desde edX. Estas clases java interactúan y limpian los datos facilitando su posterior agrupación antes de incluirlos, insertándolos o actualizando registros, en la base de datos.

El fichero que contiene la información de matriculación de los estudiantes es tratado con la clase **ParseProfile**. Este fichero contiene la información facilitada por el estudiante al registrarse en la plataforma edX. Se necesita un tratamiento especial debido a que los campos necesarios para registrar un nuevo usuario son de texto libre, es decir, se pueden introducir saltos de página y tabulaciones sin que edX lo controle. Por este motivo se ha generado una regla lógica simulando una expresión regular para asegurar la correcta lectura de cada registro.

Para anonimizar los datos privados de los estudiantes en relación a sus actividades durante el curso, se ha creado el fichero **ParseEvent**. Esta clase obtiene del fichero diario los eventos de cada estudiante. Toda información privada contenida en el fichero queda ignorada, manteniendo únicamente los campos explicados en el Anexo B.

Se tiene que tener en cuenta que el fichero diario contiene los registro de todos los eventos de todos los cursos activos en la plataforma, por ello es necesario comparar el origen de los eventos con las ediciones de cada curso. Esto se realiza mediante el método *checkSource(...)*, de esta manera se obtiene el identificador único que la base de datos ha asignado a esa combinación de curso-edición, manteniendo la persistencia en los datos y facilitando la identificación de los eventos entre todos las ediciones.

La clase **ParseMongoSocial** no solo anonimiza los datos de los estudiantes, sino que además crea las conexiones con los ficheros MongoDB [21] para poder acceder a los datos con sentencias sencillas y haciendo uso de la funcionalidad de las bases de datos NoSQL [17].

Los datos referentes a los mensajes de los foros vienen en ficheros semanales y separados por cada edición, en estos ficheros se incluye toda la información existente en los foros además de los mensajes nuevos generados durante la semana previa a su liberación por parte de edX. De manera, se ha creado un método, *isMainThead (...)*, que comprueba cual es el último mensaje registrado en el foro para una edición elegida y se comienza la inserción en la base de datos desde esa nueva entrada.

Para más información acerca de los datos anonimizados véase el Anexo B.

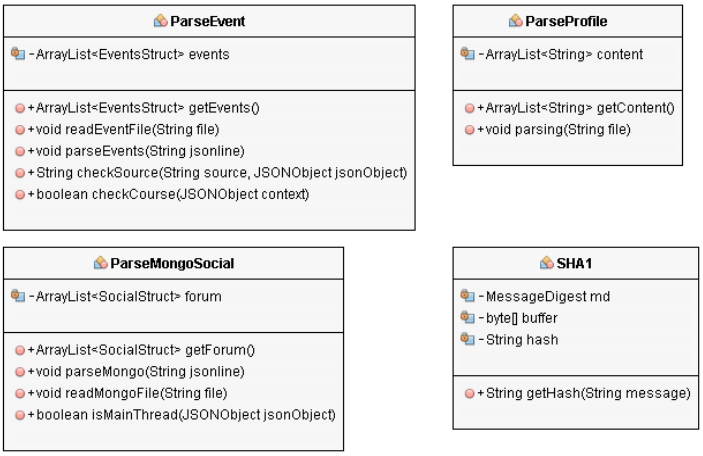

Figura 4-4: ELASA - Paquete *es.uam.tfg.elasa.parse* y ***es.uam.tfg.elasa.security***

4.3.2 Distribución de Información

4.3.2.1 Extracción de Datos Anonimizados

Las clases java encargadas de extraer la información almacenada en la base de datos están diseñadas e implementadas según la estructura de las tablas de la base de datos. La información de la matriculación de los estudiantes y la información demográfica no se pueden extraer a través de la funcionalidad del sistema. Esta información, aunque anonimizada, no deberá de estar disponible para usuarios que no sean responsables del análisis de los paquetes de datos recibidos desde la plataforma edX. Se puede ver la estructura de las clases en la *Figura 4-5*.

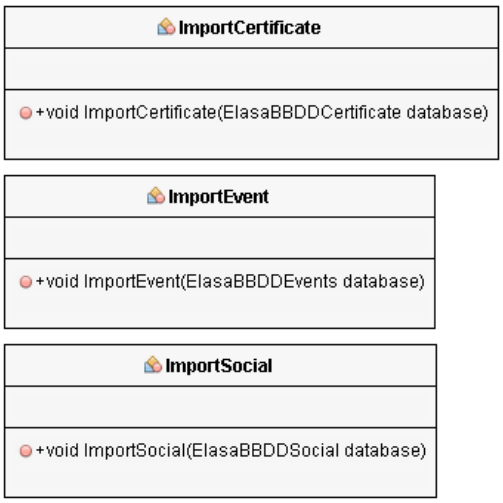

Figura 4-5: ELASA - Paquete *es.uam.tfg.elasa.importing*

La información de los certificados, eventos y los mensajes de los foros relacionados con cada estudiante se extrae en 3 ficheros CSV [62] distintos. Estos ficheros CSV contienen columnas comunes entre ellos como son el identificador que identifica la edición de cada curso y el identificador de cada estudiante. Para más detalles sobre la estructura de los ficheros generados por el sistema véase Anexo C.

4.3.2.2 Generación de Indicadores

La generación de indicadores es una parte vital del sistema ya que sin ellos no sería posible aplicar los métodos de predicción que se van a explicar en el siguiente apartado. Los indicadores que se han desarrollados han sido los que se detallaron en el capítulo 3 - Análisis de Requisitos, véase sección 3.1.1.2.2. Se puede ver la estructura de las clases en la *Figura 4-6*.

A todos los indicadores se les ha implementado la funcionalidad de obtener los datos según una serie de parámetros que reciben por el usuario: la edición de un curso y las fechas entre las que se quiere los datos. En caso de no recibir una fecha de inicio o de final, se toma como fecha inicial 01/01/1900 y como fecha final 31/12/2100. Todos los indicadores tienen método en común para volcar la información generada en ficheros CSV, haciendo uso del método *toFile (...)*.

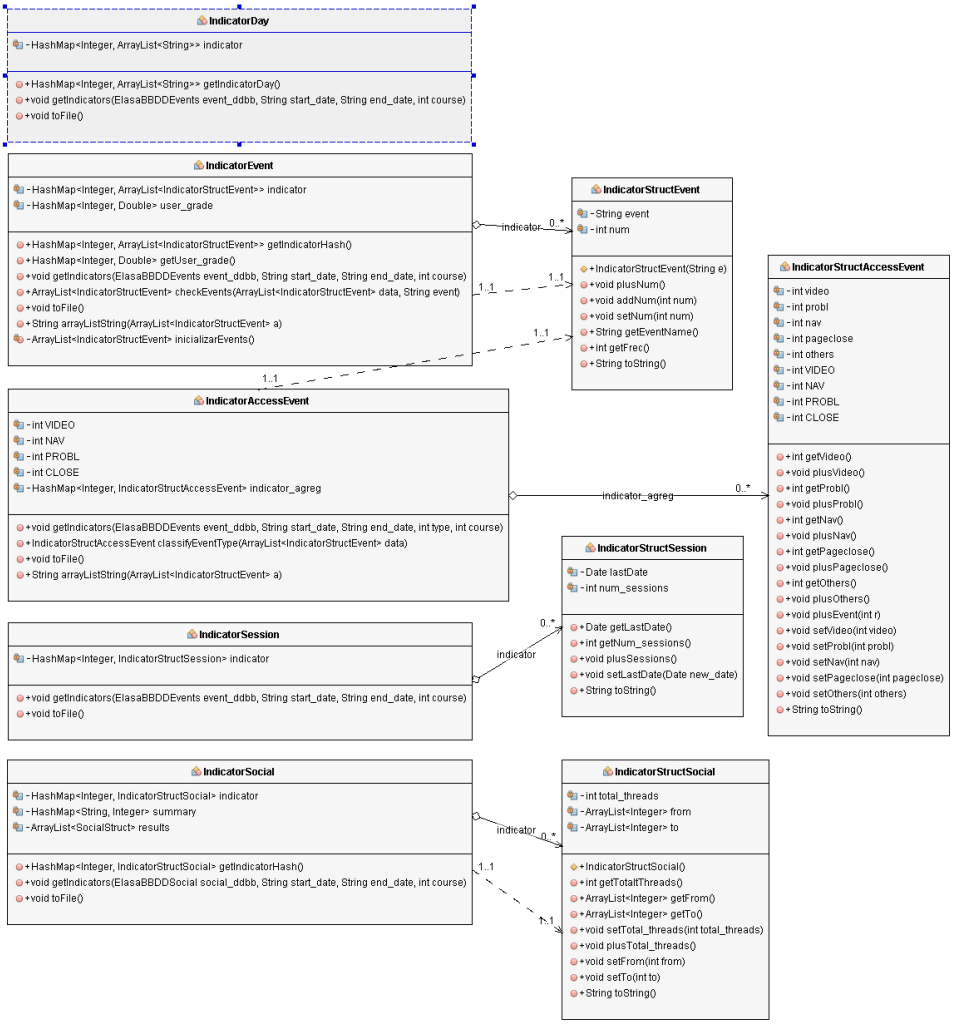

Figura 4-6: ELASA - Paquete *es.uam.tfg.elasa.indicator*

Los indicadores son generados de manera dinámica y conectándose siempre contra la base de datos, por lo que su información siempre estará tan actualizada como lo estén los registros de la edición de los cursos MOOC.

Actividad social (*RF22*): este indicador será usado para ser aplicado al algoritmo de Asortatividad que se explicará en la sección 4.3.3. Es generado por la clase

IndicatorSocial, la información de este indicador es agrupada en una estructura que emula un grafo dirigido [64] manteniendo así la información de que estudiante contesta a que estudiante.

Número de días (Constancia) (*RF 23*): este indicador es generado por la clase **IndicatorDay**, recoge el número de días que cada estudiante se ha conectado al curso. Una vez generado el indicador, este puede aplicarse a los modelos de predicción, que explicaremos a continuación.

Número de sesiones (*RF 24*): este indicador se ha planteado después de realizar una investigación sobre el tipo de contenidos que componen los cursos MOOC. Los eventos que se reciben en los paquetes de datos desde edX no tienen información relacionada con las sesiones de los navegadores de los estudiantes. Solo la fecha y hora en la que se generan los datos por cada evento.

De manera, para obtener cuantas sesiones han realizado los estudiantes durante el curso, se ha analizado la duración media del contenido que puede encontrarse en un curso para dar por finalizada una sesión por inactividad.

Se planteó que cada fin de sesión viniera fijado por la aparición de un evento denominado "*page_close*", para más información véase el Anexo B. Según la plataforma edX este evento se registra cuando un estudiante cierra el navegador o una pestaña del navegador pero no tiene por qué ser un indicador de final de actividad.

Después de analizar el contenido de los cursos MOOC se llegó a la conclusión de que los vídeos se pueden considerar como los componentes que consumen mayor tiempo a los estudiantes con una inactividad resultante de su visualización.

Teniendo en cuenta el análisis aquí detallado se asumió que el límite temporal de inactividad se debía de fijar según la duración del video más largo de cada edición de cada curso MOOC. Por lo que cualquier inactividad más allá de ese margen temporal podría considerarse un final de sesión.

Una vez realizado este análisis se creó la clase **IndicatorSession**, esta clase genera los indicadores de manera dinámica y conectándose siempre contra la base de datos. Teniendo en cuenta el parámetro que indica ese valor máximo de inactividad que marcará el final de sesión. Una vez generado el indicador, este se podrá aplicar a los modelos de predicción, que explicaremos a continuación.

Número de eventos (*RF 26*): el objetivo de este indicador es el de agrupar el número de eventos realizado por cada estudiante entre dos fechas indicadas por dos parámetros de entrada. Este indicador es generado por la clase **IndicatorEvent**, una vez generado el indicador, este puede aplicarse a los modelos de predicción, que explicaremos a continuación.

Cantidad de eventos por acceso al día (*RF 25*): el objetivo de este indicador es el de agrupar el número de eventos de un mismo tipo generado por cada estudiante en cada día, este indicador estará acotado por dos fechas estipuladas en los parámetros de entrada. Este indicador se genera mediante la clase **IndicatorAccessEvent** su implementación es muy similar a la descrita en el *RF26.*

4.3.3 Modelos de Predicción

Para el modelo de predicción sobre el absentismo estudiantil se han implementado varios algoritmos con el objetivo de clasificar a los estudiantes según sus eventos. Los algoritmos que hacen uso de los eventos registrados nos permiten analizar cuáles de ellos son más relevantes en el éxito del curso para un estudiante. El algoritmo de Asortatividad implementado genera un coeficiente de correlación entre nodos, siendo los nodos los estudiantes y las aristas los mensajes entre ellos; el coeficiente obtenido podría ser un indicador valioso sobre el absentismo de los estudiantes.

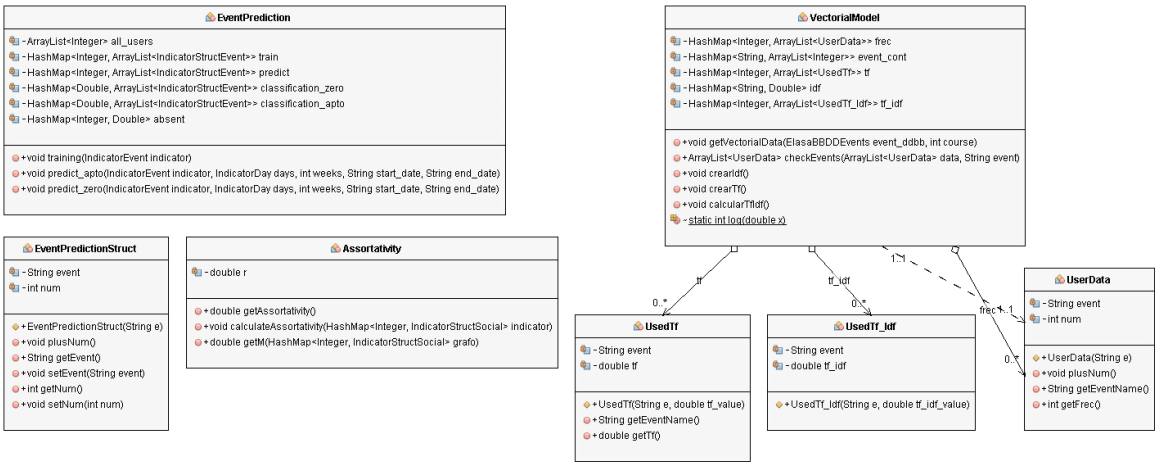

Figura 4-7: ELASA - Paquete *es.uam.tfg.elasa.prediction y es.uam.tfg.elasa.social_network*

Los modelos predictivos elegidos son:

- Modelo Vectorial Tf-Idf [56]: Es un algoritmo muy usado en la minería de datos para analizar la relevancia de un documento según las palabras del alfabeto. Se ha implementado el mismo modelo adaptándolo para que los documentos sean los estudiantes y los términos los eventos generados.

  Posteriormente se relaciona las notas de los estudiantes con el coeficiente de relevancia obtenido para cada estudiante y se crea un ranking para poder clasificar a los mejores estudiantes y cuáles son los eventos más influyentes a la hora de generar el coeficiente. El objetivo es analizar si el coeficiente de los estudiantes puede relacionarse con el posible abandono del curso o si por el contrario nos da información de estudiante con altas probabilidades de aprobarlo.

- Modelo de similitud KNN [57]: El algoritmo ha seguido la fórmula matemática que se ve en la *Figura 2-1*. Tiene el objetivo de obtener una probabilidad por cada estudiante para ver si el algoritmo lo clasifica en el grupo de alta probabilidad de abandono del curso. El algoritmo hace uso de todos los estudiantes que hayan realizado algún curso MOOC y que el curso haya finalizado para generar el grupo

de entrenamiento, se obtienen todos los indicadores anteriormente explicados para cada estudiante y se aplican al algoritmo para obtener el coeficiente de similitud.

Para evitar los valores 0 se suma 1 a los numeradores la división se realiza contra el sumatorio de todos los estudiantes del conjunto de datos y se suaviza para evitar estimaciones demasiado estrictas, sobreajuste [69] y probabilidades cero como se ve en la *Figura2-1*;

Una vez que tengamos los indicadores de cada estudiante generados se entrenan con una muestra representativa para crear un coeficiente que será usado para separar aquellos con probabilidad de abandono durante el curso de los que no. El entrenamiento tiene en cuenta los indicadores previamente explicados y la calificación final obtenida, por lo que el grupo de entrenamiento solo pueda estar compuesto por estudiantes participes en ediciones ya finalizadas.

Los indicadores de sesiones y de accesos a cada evento por día son usados para generar un peso en relación a la importancia del evento a partir los dos coeficientes de los indicadores, estos se agrupan por las notas de los estudiantes. Esto pesos se usarán para aplicarlos al conjunto de datos a predecir.

Quijote501x - 1T2015	<5	5-7	7-9	>9
Session	9,15	132,94	172,24	245,19
Navigation	11,5	200,69	258,91	390,48
Problems	0,43	17,73	20,04	33,07
Vídeos	7,44	119,16	142,9	212,52
Problems Checks	0,49	26,19	27,39	31,93

Tabla 4-1: ELASA – Valor medio de acceso a cada grupo de eventos por estudiante en la primera edición del curso "La España de El Quijote"

La Tabla4-1, contiene datos reales que se han usado en el capítulo 5 a modo de prueba. El grupo de entrenamiento generará valores para cada rango de notas, como se ve en la *Tabla 4-1*, usando estos valores para asignar pesos según la diferencia entre los valores de los alumnos de 5-7 y los alumnos de <5, esta comparativa será la primera barrera para fijar el límite de estudiantes que han aprobado y separarlos de los estudiantes que no han aprobado o han abandonado el curso.

Además a cada grupo de la *Tabla 4-1* se le asigna el valor medio de la constancia de todos los estudiantes de manera que se obtiene un peso por cada evento y un valor de constancia por cada grupo de notas. Este valor será diferente a cada ejecución ya que los estudiantes que componen el grupo de entrenamiento varían entre ejecuciones.

Al conjunto de datos restantes se le aplica el algoritmo de predicción junto con los indicadores y los pesos. El primer indicador que se aplica es el de constancia, de manera que si el estudiante tiene una constancia similar a los valores de los grupos de notas del conjunto de entrenamiento, este será seleccionado para buscar valores similares entre sus eventos similares a los valores del grupo de entrenamiento.

En caso de que la constancia del estudiante sea baja o similar a los valores obtenidos del grupo de entrenamiento de estudiantes que han abandonado o

suspendido el curso, se considera que el estudiante pueda ser muy eficiente en cada una de sus conexiones por lo que se aplica el indicador. Por ello se aplica el algoritmo KNN para obtener un coeficiente relacionado con los eventos que se asemeje a los estudiantes en probabilidad de abandono.

A los estudiantes que tengan una constancia similar a los estudiantes de entrenamiento que acabaron el curso se les aplica un algoritmo de KNN buscando la similitud de los eventos del estudiante a predecir con los eventos de los estudiantes del grupo de entrenamiento aplicando el peso definido previamente.

Aquellos estudiantes que obtengan un coeficiente de similitud en el rango de estudiantes de entrenamiento clasificados con abandonos o suspensos, son clasificados como estudiantes, que a pesar de realizar conexiones diarias e interactuar con el contenido del curso MOOC, tienen una alta probabilidad de abandono.

Los eventos que más peso tienen son los eventos de navegación a lo largo del curso, los eventos que registren accesos e interacción con contenido del curso MOOC, visualización e interacción con los vídeos y, por supuesto, aquellos eventos de corrección de los problemas. En caso de que uno de estos indicadores tenga un valor próximo a 0, se le clasifica al estudiante en posibilidad de abandono.

- Asortatividad: El algoritmo de Asortatividad pretende obtener un coeficiente de correlación de los grados entre dos estudiantes conectados.

  En caso de obtener un valor positivo, el coeficiente indica que existe una correlación entre nodos con grado similar, se dice que la red es totalmente asortativa, es decir, el foro se ha agrupado en comunidades que mantienen la mayor parte del contacto entre ellos.

  Mientras que un valor negativo indica correlaciones entre nodos de diferente grado, se dice que la red es totalmente disortativa, es decir, los estudiantes no tienden a responder siempre a los mismos estudiantes sino que mantienen una comunicación muy variada.

5 Pruebas y resultados e Integración

Con el objetivo de realizar las pruebas sobre los métodos de anonimización y los algoritmos de predicción implementados en el sistema, se han usado dos conjuntos de datos distintos. El primero conjunto de datos ha sido construido desde la plataforma edX, ya que en su Wikipedia para ayudar al análisis de los datos facilita gran cantidad de datos que pueden ser usados para orientar a los equipos de análisis haciendo uso de la documentación sobre sus ficheros [9].

Con el objetivo de realizar las pruebas iniciales y la investigación para la anonimización de los ficheros y su contenido, se crean ficheros de prueba con estos ejemplos de datos ficticios. Posteriormente se facilitaron otro conjunto de datos para realizar pruebas con datos reales, este paquete de datos contiene las dos primeras ediciones del curso de "*La España de El Quijote*" [10], curso con el código Quijote501x y cuyas ediciones se denominan 1T2015 como referencia a la primera y 3T2015 a la segunda.

5.1 Pruebas

Las pruebas realizadas sobre las funcionalidades básicas de la base de datos como la autenticación, la inserción de estudiantes, etc., y los métodos de anonimización de la información de los estudiantes, se realizaron con el conjunto de datos ficticios creados desde los ejemplos de la documentación. Posteriormente se validaron las pruebas con los datos reales de las ediciones del curso del Quijote501x.

5.1.1 Creación base de datos

Las pruebas sobre la base de datos consistieron en realizar varias pruebas de creación y borrado de la base de datos de manera iterativa, de manera que la estructura definida en el modelo de datos de la *Figura 3-3* se respetase y se crease sin errores.

5.1.2 Anonimización

Una vez los datos reales fueron incluidos en la base de datos del sistema se realizaron pruebas unitarias sobre la misma, con el objetivo de detectar errores en la anonimización de la información privada de los estudiantes o errores en la integridad y persistencia de los datos.

Las pruebas de integridad de los estudiantes se realizaron siguiendo las acciones a continuación explicadas:

- Creación de una base de datos nueva.
- Inserción de los identificadores de la edición del curso que se va a probar
- Insertar los datos de los estudiantes
- Realizar una actualización del mismo fichero y controlar en base de datos que no aumenta el número de estudiantes
- Comprobar que el número de registros de la base de datos son los mismos que el número de filas existen en el fichero facilitado por la base de datos

- Introducir el total de los mensajes generados en el foro del curso por los estudiantes, la base de datos deberá de ser capaz de mapear todos los mensajes con los identificadores anónimos de los estudiantes previamente insertados
- Comprobar que el total de registros en la base de datos se corresponde con el total de filas que está dentro del fichero facilitado por edX
- Insertar un fichero aleatorio que contenga eventos generados por los estudiantes, comprobar que todos los eventos son mapeados correctamente con cada identificador anónimo de los estudiantes previamente insertados
- Comprobar que todos los registros han sido introducidos y no han quedado eventos sin mapear
- Insertar los certificados generados para los estudiantes, comprobar que todos los estudiantes tienen asignados una nota a un identificador anónimo
- Comprobar que la base de datos tiene tantos registros en su tabla que alojará la información con filas tienen los ficheros recibidos

Estas pruebas fueron realizadas por cada edición de los cursos del Quijote en dos bases de datos distintas. Una vez las pruebas fueron un éxito, se realizaron las mismas pruebas en una base de datos nueva que contendría la información de ambas ediciones del curso facilitado.

Los datos de las ediciones con las que se hicieron las pruebas son:

Ediciones	Número de estudiantes	Número total de eventos	Número total de mensajes en foro
1T2015	3.530	442.562	862
3T2015	2.152	359.449	485

Tabla 5-1: ELASA - Datos de las ediciones del curso Quijote501x

5.2 Resultados

5.2.1 Asortatividad

El algoritmo de Asortatividad pretende obtener información sobre la interacción de los estudiantes durante el curso y obtener un indicador sobre si la actividad social es importante para la predicción del abandono de los estudiantes.

Los coeficientes obtenidos fueron:

Ediciones	Coeficientes Asortatividad
Quijote501x - 1T2015	-0.1247
Quijote501x - 3T2015	-0.1714

Tabla 5-2: ELASA - Coeficiente de Asortatividad del curso Quijote501x

Los coeficientes obtenidos nos indican que la relación entre los estudiantes no genera comunidades sino que los mensajes y la interacción de los participantes se mantiene dentro de las conversaciones. Este tipo de coeficientes son muy comunes en páginas web y foros, ya que los usuarios tienden a interactuar sobre hilos o conversaciones creadas, las temáticas no suelen traspasar esas barreras y quedan acotadas dentro de las contestaciones que se dan los estudiantes en referencia a un tema o pregunta.

A simple vista estos coeficientes no generan información sobre que usuarios han podido o no abandonar el curso, por ello se relacionó el número de comentarios y contestaciones de cada estudiante con su nota final obtenida. Se crearon dos rankings por cada edición para comparar las notas finales con el cumulo de interacciones en los foros de los 10 estudiantes con mayores valores en el número de mensajes creados y el número de respuestas generadas en cada MOOC. Todas las notas finales de los estudiantes están comprendidas en un rango de 0-1.

Para la primera edición del curso el sistema ha obtenido los siguientes resultados:

Student_ID	Course	Nota Final	Núm. de Mensajes
6560220	Quijote501x - 1T2015	0.27	5
6438284	Quijote501x - 1T2015	0.31	5
6549972	Quijote501x - 1T2015	0.0	4
5527913	Quijote501x - 1T2015	0.0	4
6580336	Quijote501x - 1T2015	0.64	5
6392866	Quijote501x - 1T2015	0.4	3
6067471	Quijote501x - 1T2015	0.12	2
6573000	Quijote501x - 1T2015	0.16	2
4117936	Quijote501x - 1T2015	0.14	2
5765986	Quijote501x - 1T2015	0.0	2

Tabla 5-3: ELASA - Relación Nota Final Numero de Mensajes, Quijote501x - 1T2015

Student_ID	Course	Nota Final	Núm. de Respuestas
6549972	Quijote501x - 1T2015	0.0	44
6575074	Quijote501x - 1T2015	0.99	17
5527913	Quijote501x - 1T2015	0.0	16
6469856	Quijote501x - 1T2015	0.93	16
6579592	Quijote501x - 1T2015	0.94	11
1546929	Quijote501x - 1T2015	0.91	11
5052141	Quijote501x - 1T2015	0.95	9
5656463	Quijote501x - 1T2015	0.87	8
6582814	Quijote501x - 1T2015	0.86	8
5801500	Quijote501x - 1T2015	0.74	7

Tabla 5-4: ELASA - Relación Nota Final Numero de Respuestas, Quijote501x - 1T2015

Para la segunda edición del curso el sistema ha obtenido los siguientes resultados:

Student_ID	Course	Nota Final	Núm. de Mensajes
8163776	Quijote501x - 3T2015	0.93	11
8409612	Quijote501x - 3T2015	0.83	9
8548130	Quijote501x - 3T2015	0.8	8
7731638	Quijote501x - 3T2015	0.55	7
5790600	Quijote501x - 3T2015	0.0	7
8546014	Quijote501x - 3T2015	0.95	7
8512791	Quijote501x - 3T2015	0.2	5

5546757	Quijote501x - 3T2015	0.95	5
8534546	Quijote501x - 3T2015	0.86	5
6515198	Quijote501x - 3T2015	0.95	5

Tabla 5-5: ELASA - Relación Nota Final Numero de Mensajes, Quijote501x - 3T2015

Student_ID	Course	Nota Final	Núm. de Respuestas
8512791	Quijote501x - 3T2015	0.2	15
8409612	Quijote501x - 3T2015	0.83	14
8546014	Quijote501x - 3T2015	0.95	11
8163776	Quijote501x - 3T2015	0.93	6
5546757	Quijote501x - 3T2015	0.95	6
8176946	Quijote501x - 3T2015	0.84	6
3488348	Quijote501x - 3T2015	0.95	5
2333823	Quijote501x - 3T2015	0.22	5
4769867	Quijote501x - 3T2015	0.19	4
6266725	Quijote501x - 3T2015	0.89	4

Tabla 5-6: ELASA - Relación Nota Final Numero de Respuestas, Quijote501x - 3T2015

Los valores obtenidos nos indican que los estudiantes son más propensos a ayudarse y a comunicarse entre ellos sin formar comunidades, simplemente respondiendo a las conversaciones o dudas creadas en el foro sin importar que estudiante sea autor. Las tablas nos indican que los estudiantes que acaban el curso y obtienen una alta calificación final suelen ser muy activos en lo que se corresponde a responder a compañeros no tanto a crear nuevos hilos y conversaciones.

5.2.1 Modelos de Predicción

5.2.1.1 Modelo vectorial Tf-Idf

Con el modelo vectorial Tf-Idf [56] implementado se ha llegado a la conclusión de no ser un modelo muy acertado haciendo uso de los indicadores por eventos para cursos online. El problema radica en que el número de eventos totales que se dispone en los ficheros recibidos de la plataforma edX no es muy grande en ningún caso llega al centenar, por lo que los estudiantes con alta probabilidad de abandonar los cursos tienden a tener un valor igual o que tiende a 0 y los estudiantes que acaban los cursos tienden a obtener un valor que tiene a 1, pero el algoritmo no genera coeficientes entre esos dos rangos por lo que no son resultados de gran utilidad.

Por ello, estos coeficientes no nos dan información sobre si los estudiantes pueden abandonar el curso o no, debido a que si un estudiante no realiza acciones en un curso durante un día, el algoritmo le clasificaría con un 0 y por lo tanto abandono del curso. Pero al día siguiente podría generar mucha actividad y por lo tanto obtendría un valor 1 y el sistema no lo clasificaría como abandono del curso.

Se trata de un algoritmo que hace uso de una variable muy sencilla para ser aplicado a una situación en la que hay que tener en cuenta varios niveles de datos, como son actividad, sesiones, constancia, tipos de eventos, etc.

5.2.1.1 Modelo de predicción por similitud

Las pruebas sobre cada edición del curso se han realizado siguiendo la lógica de un curso que avanza en el calendario académico, es decir, cada predicción incluye todos los ficheros de datos desde que comenzó la fecha de comienzo del curso hasta la fecha de fin de semana lectiva.

Para evitar un sobreajuste negativo en los datos se ha tenido en cuenta el hecho que no todos los estudiantes matriculados comienzan el curso o participan en él. Por lo que para ambas ediciones, las pruebas no han tenido en cuenta en el entrenamiento ni en la predicción aquellos estudiantes que no han generado ningún evento entre la fecha de inicio y la fecha de fin marcada para la predicción. Con esta restricción obtendríamos el siguiente juego de datos para cada edición:

Ediciones	Número de matriculados	Núm. estudiantes participes	Núm. de estudiantes en la últ. semana	Número aprobados
1T2015	3.530	1.719	453	199
3T2015	2.152	1.110	320	169

Tabla 5-7: ELASA – Estudiantes matriculados frente a estudiantes activos totales

Como se puede observar en los datos obtenido a continuación, se realizaron pruebas unitarias para cada edición del curso del Quijote. El conjunto de datos de entrenamiento se crea de manera dinámica y aleatoria con cada ejecución, por lo que se generó un proceso iterativo con realizará 100 predicciones, con el objetivo de obtener unos datos que minimizarán el impacto de obtener un conjunto de entrenamiento con las mejores y las peores combinaciones de estudiantes.

Los valores presentados en las *Tablas 5-8 y 5-9* son las medias de esas iteraciones:

Quijote501x - 1T2015	Predicción de abandonos	Porcentaje de Acierto
Semana 1	203	30.07
Semana 2	389	46.58
Semana 3	519	58.71
Semana 4	595	61.91
Semana 5	618	64.3
Semana 6	711	70.95
Semana 7	748	72.27
Semana 8	775	73.46

Tabla 5-8: ELASA – Predicción y Aciertos Quijote501x - 1T2015

Quijote501x - 3T2015	Predicción de abandonos	Porcentaje de Acierto
Semana 1	170	36.4
Semana 2	232	44.96
Semana 3	289	49.65
Semana 4	340	54.05
Semana 5	321	50.87
Semana 6	375	56.22
Semana 7	362	55.52
Semana 8	403	60.23

Tabla 5-9: ELASA – Predicción y Aciertos Quijote501x - 3T2015

De los resultados se puede deducir que existe una tendencia a mejorar en los resultados mientras avanza el curso, esta mejora es más importante desde la semana 1 a la semana 3 y luego tiende a estabilizarse, pero queda patente que cuanta más información se tenga más fina será la predicción, la *Figura 5-1* no da una mejor visión de esa tendencia.

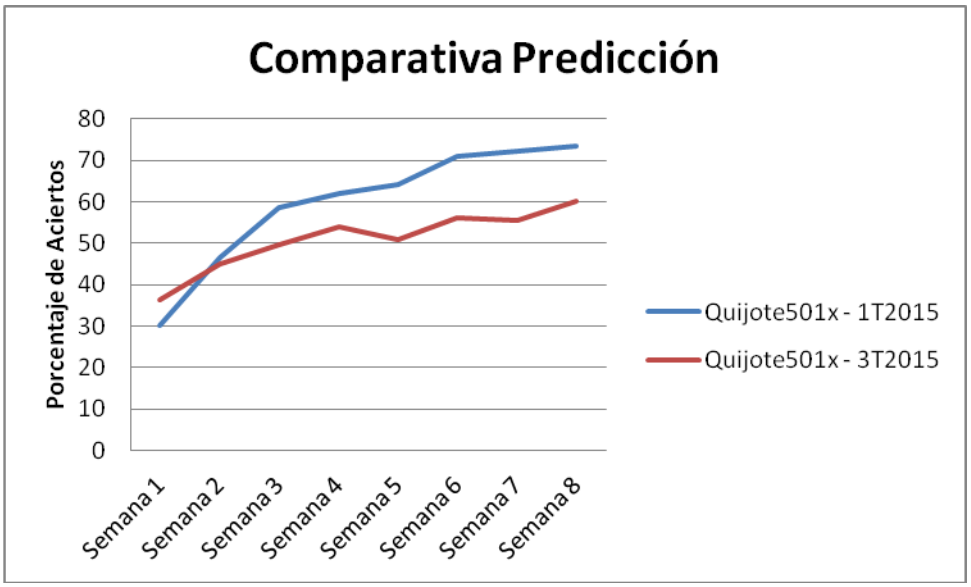

Figura 5-1: ELASA - Comparativa Predicción Quijote

Si comparamos estos datos con el número de estudiantes que se conectan cada día con la primera edición del Quijote por ejemplo, véase la *Figura 5-2*, vemos que existe una relación entre la mejora de la predicción y el número de estudiantes concurrentes cada día. La actividad en los cursos tiende a tener picos los martes, días en los que se libera el contenido de la siguiente semana de los cursos MOOC, pero cada semana existe un abandono de estudiantes que es patente en los datos y que estaría relacionado con la mejora en el acierto del algoritmo de predicción.

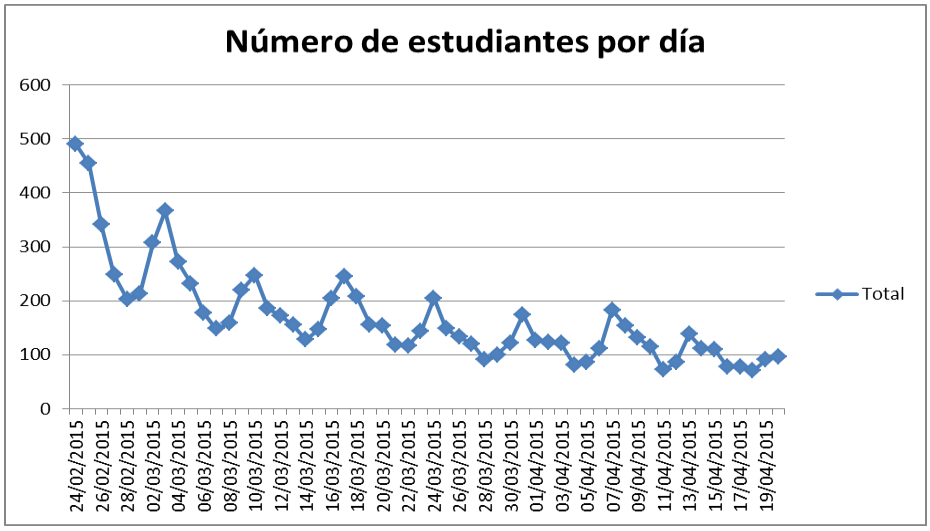

Figura 5-2: ELASA - Número de estudiantes activos por día en Quijote501x - 1T2015

Dado que los datos con los que se han realizado las pruebas son todos de un mismo modelo de un curso y de una misma rama de estudio, Humanidades, los resultados obtenidos no

deberían de ser generalizados a todos los cursos de todas las modalidades de los cursos MOOC.

5.3 Integración en la Oficina UAMx

Como test final de integridad y validación, se facilitó una versión inicial del sistema al equipo de análisis de datos de la Oficina UAMx [44]. Esta versión contenía las funcionalidades de los módulos de *Gestión y Almacenamiento de Información* y el de *Distribución de Información*. Se realizó una reunión con el responsable de explotación de datos de los MOOC para la implantación del sistema, realizar un tutorial y explicación de la versión a entregar, así como aclarar los requisitos de hardware que podían ser necesarios y el modelo de la base de datos.

En la Oficina UAMx pusieron la aplicación a trabajar con distintas ediciones de sus distintos cursos MOOC. Se generaron informes con datos anonimizados para revisar la correcta funcionalidad y detectar errores del sistema en una situación real.

Al poco tiempo se entregó una nueva versión que solventaba pequeños errores de integridad entre ediciones de un mismo curso, dado que edX cambio la forma de identificar las ediciones de los cursos dentro de los ficheros de eventos, certificados y foros.

6 Conclusiones y trabajo futuro

6.1 Conclusiones

A continuación se presentan las conclusiones más relevantes que se han alcanzado tras la ejecución del sistema, con lo que se logran satisfacer todos los objetivos que se recogieron en la introducción.

Se ha realizado un exhaustivo estudio de los datos facilitados por la plataforma edX con las interacciones de los estudiantes con sus cursos MOOC para poder crear un modelo de base de datos capaz de almacenar la información recibida en los paquetes de datos desde edX.

Se ha desarrollado un sistema que permite analizar grandes cantidades de datos, no solo de ediciones individuales sino de cursos completos y de todos los alumnos que, en algún momento puedan ser partícipes de los cursos MOOC de edX y Open edX.

Se han implementado algoritmos de predicción que pretenden identificar los estudiantes que tienen una alta probabilidad de abandono, así como indicadores que puedan servir como ejemplo de cuáles son los aspectos más destacados para la actividad de los estudiantes.

Se ha construido un prototipo del sistema que ha sido probado en un entorno real, con conjuntos de datos reales de miles de estudiantes y centenares de miles de eventos.

Los resultados obtenidos en las predicciones y en el análisis de los indicadores muestran que el abandono de un estudiante puede tener una relación directa con los eventos de video, navegación y problemas. La constancia es un indicador que permite mejorar los resultados hasta la semana 4 aproximadamente. Posteriormente un análisis más detallado de los eventos y de las relaciones en los foros ayuda a perfilar el algoritmo que predice el absentismo estudiantil.

6.2 Trabajo futuro

Toda la funcionalidad que se diseñó ha sido implementada con éxito y los algoritmos de predicción devuelve resultados con una clara correlación con los análisis en bruto de los datos. Se han detectado que los eventos relacionados con la visualización de vídeos, la navegación y la actividad con los problemas que califican a los estudiantes tienen gran importancia en la detección de casos de abandono en los cursos MOOC.

Por ello se debería de seguir esta línea de desarrollo y profundizar en los siguientes puntos:

- Los eventos relacionados con los vídeos se deberían de analizar individualmente. Ver cuáles de ellos son los más importantes y cuáles no. Esta mejora afinaría los indicadores y la predicción.
- Realizar una investigación sobre si la longitud de los vídeos y el idioma puede afectar al interés de los estudiantes. Analizar cuanto del contenido multimedia y audiovisual se consume en su totalidad, cuanto parcialmente y cuanto tiene una cantidad de accesos inusualmente alto (esto nos podría indicar contenido especialmente difícil o poco claro para los estudiantes).

- Se podría analizar la navegación de los estudiante, identificando la sección con poca actividad y que sirva para afinar los indicadores y, por lo tanto, la predicción.
- Analizar todos los eventos relacionados con los problemas de los cursos. Investigar si los estudiantes con mejores notas consumen todos los intentos de realizar los ejercicios con el objetivo de conseguir la máxima calificación y realizar una comparativa de si existe una mejora significativa sobre la calificación final al realizar las actividades repetidamente hasta el límite permitido por el sistema.
- Analizar el comportamiento de los foros sociales durante el curso. Realizar un análisis semántico del contenido de los posts para complementar las encuestas que realizan los estudiantes al final del curso, para saber si están alineadas las sensaciones en el foro con las respuestas de las encuestas.

Referencias

[1] Apache Friends, «XAMPP,» [En línea]. Disponible: https://www.apachefriends.org/es/index.html. [Último acceso: Junio 2016]

[2] AWS, Cloud Computing «Amazon Web Service (AWS)» [En línea]. Disponible: https://aws.amazon.com/es/. [Último acceso: Junio 2016]

[3] C. Paramio, «El concepto NoSQL, o cómo almacenar tus datos en una base de datos no relacional,» [En línea]. Disponible: http://www.genbetadev.com/bases-de-datos/el-concepto-nosql-o-como-almacenar-tus-datos-en-una-base-de-datos-no-relacional/. [Último acceso: Abril 2011].

[4] D. Zielke. M. Bültmann. M. A. Chati. U. Schroeder. A. L. Dyckhoff, Design and Implementation of a Learning Analytics Toolkit for Teachers, RWTH Aachen University, Germany: Educational Technology & Society, 2012.

[5] D2L Corporation, «Student Success System» [En línea]. Disponible: http://www.d2l.com/products/student-success-system/. [Último acceso: Junio 2016]

[6] Desarrollo Iterativo Incremental - Proyectos Agiles [En Línea]. Disponible: https://proyectosagiles.org/desarrollo-iterativo-incremental/ [Último acceso: Mayo 2016]

[7] Education Technology Trends - Part II - LMS and LCMS, Education Marketplaces, «LCMS,» [En línea]. Disponible: http://www.programmableweb.com/news/education-technology-trends-part-ii-lms-and-lcms-education-marketplaces/2013/08/23/. [Último acceso: Junio 2016]

[8] edX Inc., Data Czar Selection and Responsibilities «Data Czar and Data Teams» [En línea]. Disponible: http://edx.readthedocs.io/projects/devdata/en/stable/internal_data_formats/data_czar.html/. [Último acceso: Junio 2016]

[9] edX Inc., EdX Research Guide [En línea]. Disponible: http://edx.readthedocs.io/projects/devdata/en/latest/index.html/. [Último acceso: Junio 2016]

[10] edX, La España del Quijote «La España del Quijote» [En línea]. Disponible:https://www.edx.org/course/la-espana-de-el-quijote-uamx-quijote501x-0/. [Último acceso: Junio 2016]

[11] El E-Learning como medio educativo y de desarrollo profesional para las organizaciones [En Línea] Disponible: http://exa.unne.edu.ar/informatica/SO/Gisemono.pdf [Último acceso: Junio 2016]

[12] Filezilla, «FTP Filezilla,» [En línea]. Disponible: https://filezilla-project.org/. [Último acceso: Junio 2016]

[13] Github, «Github,» [En línea]. Disponible: https://github.com/. [Último acceso: Junio 2016]

[14] I. Claros, R. Cobos, G. Sandoval, M. Villanueva, "Creating MOOCs by UAMx: experiences and expectations," Proceedings of the EMOOCs 2015 European MOOC Stakeholders Summit,Mons, BE, pp. 18-20, 2015

[15] J.A. Ruipérez-Valiente, P.J. Muñoz-Merino, C. Delgado Kloos, "A Predictive Model of Learning Gains for a Video and Exercise Intensive Learning Environment," Proceedings of the Artificial Intelligence in Education, pp. 760-763, 2015

[16] J.A. Ruipérez-Valiente, P.J. Muñoz-Merino, C. Delgado Kloos, "Detecting and Clustering Students by their Gamification Behavior with Badges: A Case Study in Engineering Education," International Journal of Engineering Education, Aceptado para publicación, 2016

[17] Justificaturespuesta, «Aprendizaje Electrónico» [En línea]. Disponible: http://justificaturespuesta.com/que-es-el-e-learning-20-ventajas-del-aprendizaje-electronico/. [Último acceso: Mayo 2016].

[18] LOCO-Analyst, «LOCO-Analyst» [En línea]. Disponible: http://jelenajovanovic.net/LOCO-Analyst/index.html. [Último acceso: Junio 2016]

[19] Mercurial, «Mercurial,» [En línea]. Disponible: https://www.mercurial-scm.org/. [Último acceso: Junio 2016]
[20] Métodos predictivos y descriptivos, «predictivos,» [En línea]. Disponible: http://www.slideshare.net/lalopg/mtodos-predictivos-y-descriptivos-minera-de-datos/. [Último acceso: Junio 2016]
[21] MongoDB Inc., «MongoDB,» [En línea]. Available: https://www.mongodb.org/. Disponible: [Último acceso: Junio 2016].
[22] MySQL, «MySQL,» [En línea]. Disponible: https://www.mysql.com/. [Último acceso: Junio 2016]
[23] MySQL, InnoDB Storage, «esquema InnoDB» [En línea]. Disponible: http://dev.mysql.com/doc/refman/5.7/en/innodb-storage-engine.html/. [Último acceso: Junio 2016]
[24] Oracle Cloud, «Máquina Virtual de Java» [En línea]. Disponible: https://docs.oracle.com/javase/specs/jvms/se7/html/jvms-1.html#jvms-1.2. [Último acceso: Junio 2016]
[25] Oracle Netbeans, «NetBeans IDE 8.02,» [En línea]. Disponible: https://netbeans.org/. [Último acceso: Junio 2016]
[26] Oracle, «Java» [En línea]. Disponible: https://java.com/es/. [Último acceso: Junio 2016]
[27] Oracle, «Oracle 9.11,» [En línea]. Disponible: http://www.oracle.com/index.html/. [Último acceso: Junio 2016]
[28] Plataforma edX, «edX» [En línea]. Disponible: https://www.edx.org/. [Último acceso: Mayo 2016].
[29] PostgreSQL, «PostgreSQL,» [En línea]. Disponible: https://www.postgresql.org/. Disponible: [Último acceso: Junio 2016]
[30] Python Software Foundation «Python» [En línea]. Disponible: https://www.python.org/. [Último acceso: Junio 2016]
[31] R. Cobos. S. Gil., A. Lareo, F.A. Vargas, "Open-DLAs: An Open Dashboard for Learning Analytics," L@S 2016
[32] Sakai Collaborative Learning Environment by Marist College, «SNAPP» [En línea]. Disponible: https://confluence.sakaiproject.org/pages/viewpage.action?pageId=84902193/. [Último acceso: Junio 2016]
[33] SQLite, «SQLite,» [En línea]. Disponible: https://www.sqlite.org/. [Último acceso: Junio 2016]
[34] The Apache Software Foundation - Subversion, «Subversion,» [En línea]. Disponible: https://subversion.apache.org/. [Último acceso: Junio 2016]
[35] The Apache Software Foundation, «Apache Cassandra,» [En línea]. Disponible: http://cassandra.apache.org/. [Último acceso: Junio 2016].
[36] The Apache Software Foundation, «Apache CouchDB,» [En línea]. Disponible: http://couchdb.apache.org/. [Último acceso: Junio 2016].
[37] The Apache Software Foundation, «Servidor Apache,» [En línea]. Disponible: https://httpd.apache.org/. [Último acceso: Junio 2016]
[38] The GnuPG Project «GPG» [En línea]. Disponible: https://www.gnupg.org/. [Último acceso: Junio 2016]
[39] The GnuPG Project «TAR» [En línea]. Disponible: https://www.gnu.org/software/tar/. [Último acceso: Junio 2016]
[40] The PHP Group «PHP» [En línea]. Disponible: http://php.net/. [Último acceso: Junio 2016]
[41] Thecompleteuniversityguide, «MOOCs» [En línea]. Disponible: http://www.thecompleteuniversityguide.co.uk/distance-learning/moocs-(massive-open-online-courses)/. [Último acceso: Mayo 2016].
[42] Tipos de sentencias SQL y sus componentes sintácticos, «Rendimiento en tiempo real» [En línea]. Disponible: http://www.desarrolloweb.com/articulos/tipos-de-sentencias-sql.html/. [Último acceso: Junio 2016]

[43] UAM - MOOCs y SPOCs, «MOOCs y SPOCs,» [En línea]. Disponible: http://www.emadridnet.org/es/hacia-la-innovacion-educativa-los-moocs-y-spocs-de-la-universidad-autonoma-de-madrid. [Último acceso: Mayo 2016].
[44] UAMx, «Oficina UAMx,» [En línea]. Disponible: www.uam.es/uamx. [Último acceso: Mayo 2016].
[45] Universidad Autónoma de Madrid, «Universidad Autónoma de Madrid,» [En línea]. Disponible: http://www.uam.es/. [Último acceso: Mayo 2016].
[46] Universidad Carlos III, «Analyse» [En línea]. Disponible: http://www.it.uc3m.es/pedmume/ANALYSE/. [Último acceso: Junio 2016]
[47] Universidad Carlos III, «Gradient» [En línea]. Disponible: http://gradient.it.uc3m.es/. [Último acceso: Junio 2016]
[48] University of Saskatchewan Canada, «Universidad de Saskatchewan,» [En línea]. Disponible: http://www.usask.ca/. [Último acceso: Junio 2016]
[49] W3schools «HTML» [En línea]. Disponible: http://www.w3schools.com/html/. [Último acceso: Junio 2016]
[50] Wikipedia «Learning Analytics» [En línea]. Disponible: https://es.wikipedia.org/wiki/Learning_analytics [Último acceso: Junio 2016]
[51] Wikipedia Base de Datos relacional [En línea]. Disponible: https://es.wikipedia.org/wiki/Base_de_datos_relacional [Último acceso: Junio 2016]
[52] Wikipedia LCMS, «LCMS,» [En línea]. Disponible: https://en.wikipedia.org/wiki/Learning_management_system/. [Último acceso: Junio 2016]
[53] Wikipedia NoSQL «NoSQL» [En línea]. Disponible: https://es.wikipedia.org/wiki/NoSQL [Último acceso: Junio 2016]
[54] Wikipedia, «Aprendizaje Electrónico» [En línea]. Disponible: https://es.wikipedia.org/wiki/Aprendizaje_electr%C3%B3nico. [Último acceso: Mayo 2016].
[55] Wikipedia, «Asortatividad» [En línea]. Disponible: https://es.wikipedia.org/wiki/Asortatividad/. [Último acceso: Junio 2016]
[56] Wikipedia, «Modelo Vectorial por Tf-Idf» [En línea]. Disponible: https://en.wikipedia.org/wiki/Tf%E2%80%93idf/. [Último acceso: Junio 2016]
[57] Wikipedia, «Modelos de similitud: algoritmo de KNN» [En línea]. Disponible: http://arantxa.ii.uam.es/~castells/docencia/ir/apuntes/7-recomendacion.pdf/. [Último acceso: Junio 2016]
[58] Wikipedia, «MOOCs» [En línea]. Disponible: https://en.wikipedia.org/wiki/Massive_open_online_course. [Último acceso: Mayo 2016].
[59] Wikipedia, Algoritmo Rocchio «Rocchio» [En línea]. Disponible: https://en.wikipedia.org/wiki/Rocchio_algorithm/. [Último acceso: Junio 2016]
[60] Wikipedia, Archivo Batch «Ficheros por lotes (batch)» [En línea]. Disponible: https://es.wikipedia.org/wiki/Archivo_batch/. [Último acceso: Junio 2016]
[61] Wikipedia, Clasificador Naïve Bayes «Naïve Bayes» [En línea]. Disponible: https://en.wikipedia.org/wiki/Naive_Bayes_classifier/. [Último acceso: Junio 2016]
[62] Wikipedia, Comma-separated values «CSV» [En línea]. Disponible: https://en.wikipedia.org/wiki/Comma-separated_values/. [Último acceso: Junio 2016]
[63] Wikipedia, Extensible Markup Language «XML» [En línea]. Disponible: https://es.wikipedia.org/wiki/Extensible_Markup_Language/. [Último acceso: Junio 2016]
[64] Wikipedia, Grafo dirigido «grafo dirigido» [En línea]. Disponible: https://es.wikipedia.org/wiki/Grafo_dirigido/. [Último acceso: Junio 2016]
[65] Wikipedia, GZIP «GZ» [En línea]. Disponible: https://en.wikipedia.org/wiki/Gzip/. [Último acceso: Junio 2016]
[66] Wikipedia, JSON «JSON» [En línea]. Disponible: https://es.wikipedia.org/wiki/JSON/. [Último acceso: Junio 2016]
[67] Wikipedia, Pretty Good Privacy «PGP» [En línea]. Disponible: https://en.wikipedia.org/wiki/Pretty_Good_Privacy/. [Último acceso: Junio 2016]

[68] Wikipedia, SHA1 «SHA1» [En línea]. Disponible: https://es.wikipedia.org/wiki/Secure_Hash_Algorithm#SHA-1/. [Último acceso: Junio 2016]
[69] Wikipedia, Sobreajuste «Sobreajuste» [En línea]. Disponible: https://es.wikipedia.org/wiki/Sobreajuste/. [Último acceso: Junio 2016]
[70] Wikipedia, Sotfware Framework «framework» [En línea]. Disponible: https://en.wikipedia.org/wiki/Software_framework/. [Último acceso: Junio 2016]
[71] Wikipedia, ZIP (File Format) «ZIP» [En línea]. Disponible: https://en.wikipedia.org/wiki/Zip_(file_format)/. [Último acceso: Junio 2016]
[72] Zope Foundation «Zope» [En línea]. Disponible: http://www.zope.org/. [Último acceso: Junio 2016]

Glosario

MOOC	Massive Open Online Course
SPOC	Small Private Online Course
TFG	Trabajo de Fin de Grado
TIC	Tecnologías de la Información y de la Comunicación
UAM	Universidad Autónoma de Madrid
ELASA	E-Learning Analyser for Student's Absenteeism
BBDD	Base de datos
AWS	Amazon Web Service

Anexos

A Estructura de paquete de datos de edX

La plataforma edX [28] facilita a las instituciones que trabajen con su plataforma información acerca de sus estudiante de cómo interactúan con los cursos y sus resultados finales. Además dan visibilidad de información relacionada con el registro del usuario en la plataforma, esta información suele ser requerida por la plataforma cuando el usuario se registra por primera vez.

Las instituciones que tengan convenio con edX para recibir estos datos, no solo para usar su plataforma para crear cursos MOOC [41] [58], reciben los datos en ficheros a través de un procedimiento que involucra al servicio Amazon Web Service (AWS) [2].

Los grupos de investigación y las instituciones que utilicen los servicios de exportación de datos de edX con el objetivo de investigar y analizar sus cursos y a sus estudiantes son denominados Data Czars o Data Teams [8]. Estos Data Teams son pequeños grupos de investigadores que tienen acreditación para descargar y desencriptar los ficheros de edX. La figura de Data Teams en la Universidad Autónoma de Madrid [45] recae en el equipo de análisis de datos de la Oficina UAMx [44], dada la sensibilidad de los datos normalmente la gestión de los paquetes de datos es responsabilidad de personal muy limitado.

Los involucrados en el Data Team deben de tener cono conocimientos de:

- Administración de bases de datos SQL [13] y NoSQL [53]
- Minería de datos y análisis de datos
- Gestión de archivos grandes mediante Amazon Web Service
- Cifrado con técnicas PGP [67] y GPG [38]
- Formatos de compresión TAR [39], GZ [65], y ZIP [71]
- Formatos de datos CSV [62], JSON [66], MongoDB[21], XML [63] y HTML [49]
- Administración y creación de consolas y archivos de procesamiento de ficheros por lotes (batch) [60]

El Data Team ha de estar autorizado para acceder a los datos, esto obliga a obtener un certificado a modo de llave electrónica para el acceso seguro al repositorio de datos desde el AWS. Los datos recibidos están encriptados, se utiliza un proceso de claves pública-privada que se comparten entre la plataforma edX y las instituciones a través de GNU Privacy Guard (GnuPG o GPG).

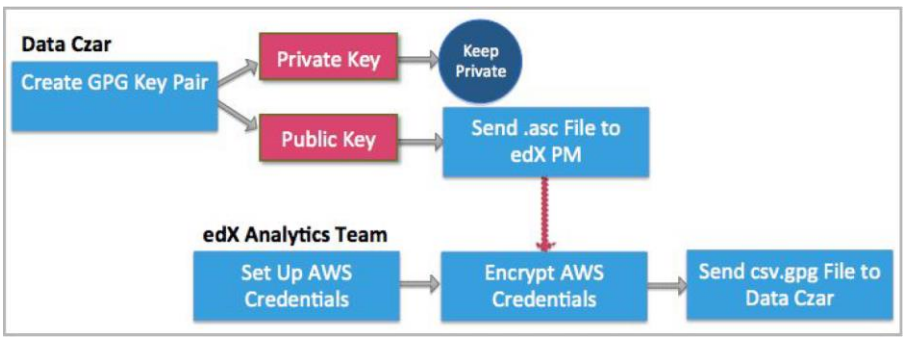

Figura 0-1: edX – Esquema proceso encriptación paquete de datos

Una vez establecidas las claves o llaves electrónicas para el Data Team se podrá descifrar los archivos. Como muestra la Figura 0-2, en ese momento se podrá acceder al servicio AWS para extraer los paquetes de datos.

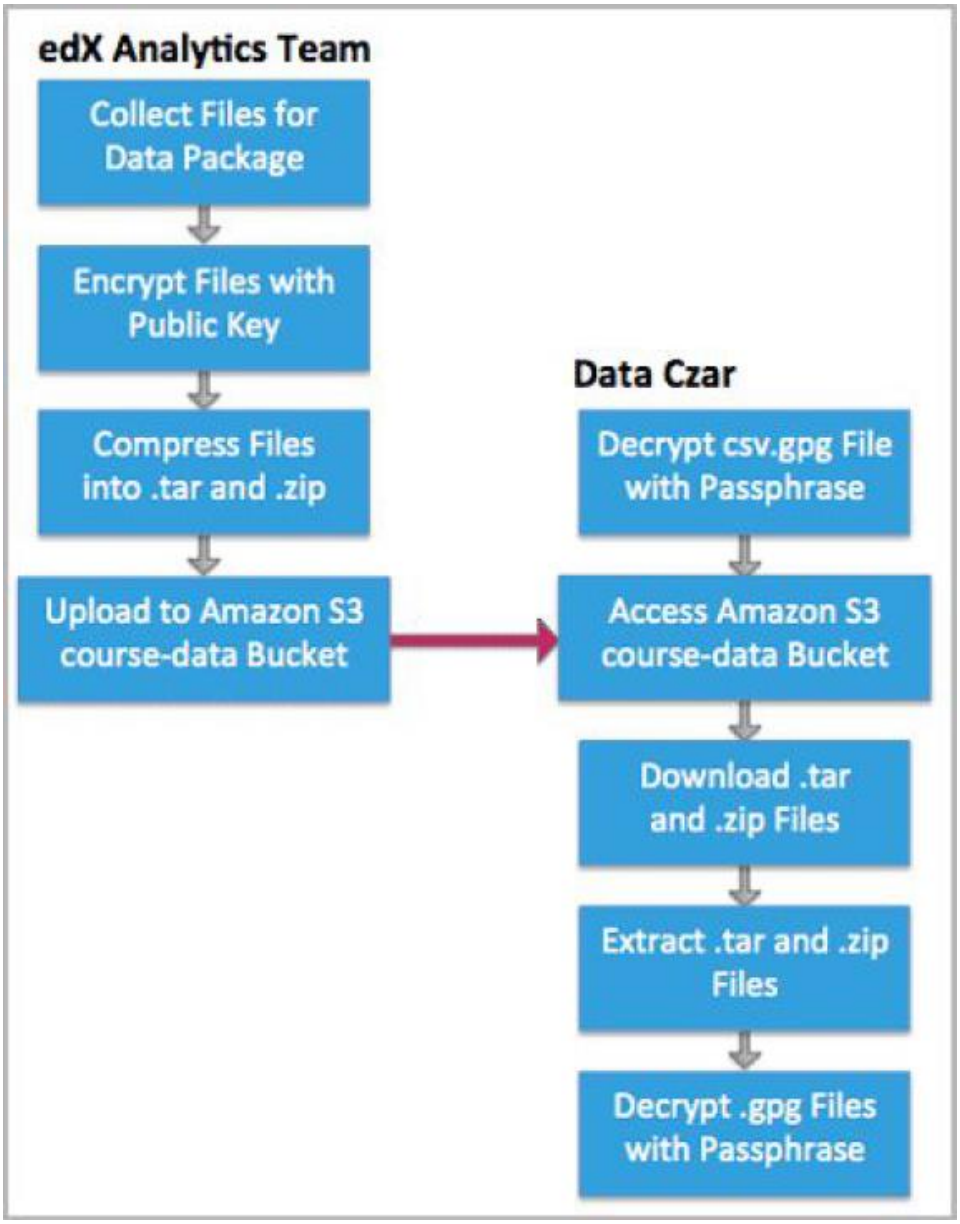

Figura 0-2: edX – Esquema del proceso de acceso a AWS y desencriptación de paquete de datos

Los paquetes de datos contienen conjuntos de ficheros comprimidos con los registros de las bases de datos de edx.org y edge.edx.org, este último es el entorno de desarrollo y pruebas para los cursos MOOC. Se obtienen dos tipos de ficheros las acciones diarios y los datos agregados semanalmente.

En el caso de los ficheros de acciones semanales, los datos están agrupados en ficheros diarios en formato SQL [13] y MongoDB [21]. Todos los ficheros vienen identificados siguiendo el mismo procedimiento:

- {org}{course}{date}auth_user_prod_analytics{site}.sql: información sobre registro y matriculación de los estudiantes en los MOOC.
- {org}{course}{date} certificates_generatedcertificate{site}analytics.sql: tabla de notas final y estado de la certificación que relaciona al estudiante con la nota final del curso.
- {org}{course}{date} courseware_studentmodule{site}analytics.sql: estado del courseware por cada estudiante.
- {org}{course}{date} user_id_map{site}analytics.sql: información que relaciona a los estudiantes entre su nombre de usuario y un identificador numérico anónimo.
- {org}{course}{date}{site}.mongo: información de las discusiones (foros).
- {org}{course}{date} wiki_article{site}analytics.sql: información de los artículos añadidos.
- {org}{course}{date} wiki_articlerevision{site}analytics.sql: información de artículos borrados o modificados.
- {org}{course}{date} auth_userprofile{site}analytics.sql: información sobre los datos del perfil de los usuarios.
- {org}{course}{date} student_courseenrollment{site}analytics.sql: información sobre el tipo de curso que va a realizar el estudiante.
- {org}{course}{date} user_api_usercoursetag{site}analytics.sql: información sobre la estructuración del curso y de sus contenidos.

En el caso de los ficheros de acciones diarias, los datos están agrupados en ficheros diarios en formato JSON [66]. Todos los ficheros son creados con un nombre temporal asignado por los sistemas de la plataforma edX.

B Anonimización de ficheros edX

De todos los ficheros que se han explicado en el Anexo A anteriormente, el sistema solo hará uso de 5 de ellos. A continuación se detalla cual es el objetivo de cada uno de ellos, que datos se anonimizarán y que campos de los ficheros son los que se incluirán en la base de datos.

Los ficheros necesarios para el proceso de anonimización y la inserción de los registros de los estudiantes en la base de datos son:

- {org}{course}{date} user_id_map{site}analytics.sql
- {org}{course}{date} auth_userprofile{site}analytics.sql
- {org}{course}{date} auth_user_prod_analytics {site}.sql
- {org}{course}{date} certificates_generatedcertificate{site}analytics.sql
- {org}{course}{date}{site}.mongo

Además de los ficheros semanales que se acaban de indicar, la anonimización de los eventos se produce a partir de los ficheros logs que registran la actividad de los estudiantes y que son recibidos diariamente.

{org}{course}{date}user_id_map{site}analytics.sql

Este fichero es el más importante para la anonimización de los datos de los estudiantes, ya que sin él no sería posible esta funcionalidad. Su contenido es simple solo tiene 3 campos como se ve en la Figura 0-3.

```
hash_id id  username

e9989f2cca1d699d88e14fd43ccb5b5f  9999999 AAAAAAAA
```

Figura 0-3: edX – user_id_map

El valor **hash_id** es usado por edX para ofuscar la identidad de estudiante. Los manuales de desarrollo y analíticas de aprendizaje de edX no dan más información sobre el campo o como puede relacionarse con otros ficheros.

El valor **id** es el identificador único que usa edX en sus sistemas internos para mantener siempre la integridad de datos en su base de datos sabiendo que un usuario puede realizar un cambio en su nombre de usuario en la plataforma.

El valor **username** es el nombre de usuario elegido por el estudiante al realizar el registro en la plataforma edX.

El motivo de por qué es tan importante este fichero es debido a que todos los ficheros incluidos en el paquete de datos usan el valor **username** para identificar a sus estudiantes, por ello si queremos anonimizar las actividades de los mismos, se debe de incluir en la base de datos de ELASA primero este fichero con el objetivo de poder mapear cada **username** con el **id** que nos facilita edX y así eliminar información privada del estudiante.

El sistema almacena en la base de datos el **id** que se recibe en el fichero y se aplica la función SHA1, explicada en el capítulo de Diseño, sobre el **username** para mantener la privacidad.

{org}{course}{date}auth_userprofile{site}.sql

Este fichero contiene toda la información del registro de cada estudiante en la plataforma edX. Un ejemplo del contenido que se puede encontrar en el fichero se puede ver en la *Figura 0-4*.

```
id  user_id name  language  location  meta  courseware  gender
mailing_address year_of_birth level_of_education  goals allow_certificate
country  city  bio  profile_image_uploaded_at

9999999  AAAAAAAA  AAAAAAAAA English MIT {"old_emails":
[["aaaaa@xxxxx.xxx", "2012-11-16T10:28:10.096489"]], "old_names":
[["BBBBBBBBBBBB", "I wanted to test out the name-change functionality",
"2012-10-22T12:23:10.598444"]]} course.xml  NULL  NULL  NULL  NULL  NULL
1      NULL   Hi! I'm from the US and I've taken 4 edX courses so far. I
want to learn how to confront problems of wealth inequality. 2015-04-19 16:41:27
```

Figura 0-4: edX – auth_userprofile

La anonimización de este fichero pasa por comprobar que todos los valores **user_id** están registrados en la base de datos. Este valor se corresponde con el valor **id** explicado anteriormente del fichero ***{org}{course}{date}user_id_map{site}analytics.sql***. En caso de que el valor **user_id** no esté registrado se notificará al usuario del sistema y los datos no serán introducidos en la base de datos ya que no se podría realizar la anonimización correctamente.

Si el mapeo entre **id** y **user_id** es satisfactorio, se recogerán de cada estudiante los valores de **gender**, **birth_date**, **country** y **level_education**.

{org}{course}{date}auth_user_prod_analytics{site}.sql

Este fichero contiene información de la actividad del estudiante en la plataforma edX y en un curso MOOC en concreto. Un ejemplo del contenido que se puede encontrar en el fichero se puede ver en la *Figura 0-5*.

```
id  username  first_name  last_name  email  password  is_staff  is_active
is_superuser  last_login  date_joined status  email_key  avatar_typ
country  show_country  date_of_birth  interesting_tags  ignored_tags
email_tag_filter_strategy display_tag_filter_strategy
consecutive_days_visit_count

9999999    AAAAAAAAA    AAAAAA  AAAAAA 1 1 0 2014-01-01 17:28:27 2012-03-04
00:57:49   NULL      0 NULL      0 0
```

Figura 0-5: edX – auth_user_prod_analytics

La anonimización de este fichero pasa por comprobar que todos los valores **id** están registrados en la base de datos. Este valor se corresponde con el valor **id** explicado anteriormente del fichero ***{org}{course}{date}user_id_map{site}analytics.sql***. En caso de que el valor no esté registrado se notificará al usuario del sistema y los datos no serán introducidos en la base de datos ya que no se podría realizar la anonimización correctamente.

Si el mapeo entre ambos valores **id** es correcto, se recogerá de cada estudiante los valores de **staff**, **date_joined** y **last_login**. El campo **staff** puede almacenar valores alfabéticos entre '*Y*' y '*N*'. Este valor nos indica que estudiantes son en realidad personales de edX que trabaja en el control y monitorización de los cursos MOOC. Este campo es especialmente útil ya que todo usuario que tenga un valor *Y* en este campo tiene que ser excluido de los indicadores y de la predicción de los cursos.

{org}{course}{date}certificate_generatedcertificates{site}.sql

Este fichero contiene la información relacionada con las calificaciones finales de cada estudiante y su modalidad de certificado para un curso MOOC en concreto. Un ejemplo del contenido que se puede encontrar en el fichero se puede ver en la *Figura 0-6*.

```
id  user_id  download_url  grade  course_id  key  distinction  status  verify_uuid
download_uuid  name  created_date  modified_date error_reason  mode

26  9999999
https://s3.amazonaws.com/verify.edx.org/downloads/9_hash_1/Certificate.pdf
0.84  BerkeleyX/CS169.1x/2012_Fall  f_hash_a  0  downloadable  2_hash_f
9_hash_1  AAAAAA  2012-11-10  00:12:11  2012-11-10  00:12:13  honor

27  9999999  0.0  BerkeleyX/CS169.1x/2012_Fall  0  notpassing  AAAAAA
2012-11-10  00:12:11  2012-11-26  19:06:19  honor
```

Figura 0-6: edX – certificate_generatedcertificates

La anonimización de este fichero pasa por comprobar que todos los valores **user_id** están registrados en la base de datos. Este valor se corresponde con el valor **id** explicado anteriormente del fichero ***{org}{course}{date}user_id_map{site}analytics.sql***. En caso de que el valor no esté registrado se notificará al usuario del sistema y los datos no serán introducidos en la base de datos ya que no se podría realizar la anonimización correctamente.

Si el mapeo entre los valores **id** y **user_id** es correcto, se recogerá de cada estudiante los valores de **grade** y **mode**. Las calificaciones de los estudiantes están comprendidas entre 0 y 1 por defecto en los ficheros de los paquetes de datos de edX.

{org}{course}{date}{site}.mongo

Este fichero contiene la información de todos los mensajes que puedan escribirse en los foros para un curso MOOC en concreto. Un ejemplo del contenido que se puede encontrar en el fichero se puede ver en la *Figura 0-7*.

```
{ "_id" : { "$oid" : "52e54fdd801eb74c33000070" }, "votes" : { "up" : [],
"down" : [], "up_count" : 0, "down_count" : 0, "count" : 0, "point" : 0 },
"visible" : true, "abuse_flaggers" : [], "historical_abuse_flaggers" : [],
"parent_ids" : [], "at_position_list" : [], "body" : "I'm hoping this
Demonstration course will help me figure out how to take the course I enrolled
in. I am just auditing the course, but I want to benefit from it as much as
possible, as I am extremely interested in it.\n", "course_id" :
"edX/DemoX/Demo_Course", "_type" : "Comment", "endorsed" : true, "endorsement"
: { "user_id" : "9", "time" : ISODate("2014-08-29T15:11:49.442Z") },
"anonymous" : false, "anonymous_to_peers" : false, "author_id" : "NNNNNNN",
"comment_thread_id" : { "$oid" : "52e4e880c0df1fa59600004d" },
"author_username" : "AAAAAAAAAA", "sk" : "52e54fdd801eb74c33000070",
updated_at" : { "$date" : 1390759901966 }, "created_at" : { "$date" :
1390759901966 } }
```

Figura 0-7: edX – fichero mongoDB

La anonimización de este fichero pasa por comprobar que el valor **author_id** que este asignado al mensaje este registrado en la base de datos. Este valor se compara con los valores **id** del fichero explicado ***{org}{course}{date}user_id_map{site}analytics.sql***. En caso de que el valor no esté registrado se notificará al usuario del sistema y los mensajes no serán introducidos en la base de datos ya que no se podría realizar la anonimización correctamente.

Si el mapeo entre los valores **id** y **author_id** es correcto, cada identificador de cada mensaje es almacenado, esto valor se extraería de {"_id" : {"$oid" : "**Identificador del mensaje**"}}. También se almacena la información de los campos **up_count**, **down_count** y **comment_thread_id**.

El valor de **comment_thread_id** nos da la información necesaria para saber si el mensaje que almacenamos en la base de datos es un mensaje nuevo, una contestación al hilo principal o es una respuesta a otro comentario previo pero no al hilo principal.

Ficheros de eventos

El fichero de evento es distinto a los ficheros anteriormente explicados. Son ficheros generados de manera diaria, con una gran cantidad de eventos por día y combinan todos los eventos de todos los cursos activos para la institución que recibe el paquete de datos. La Figura0-8 es un ejemplo de darle un evento que muestra las respuestas de un estudiante a ejercicios del curso.


```
{"agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/30.0.1599.101 Safari/537.36", "context": {"course_id": "edx/AN101/2014_T1",
"module": {"display_name": "Multiple Choice Questions"}, "org_id": "edx", "user_id":
9999999}, "event": {"answers": {"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1":
"yellow", "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1": ["choice_0", "choice_2"]},
"attempts": 1, "correct_map": {"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1":
{"correctness": "incorrect", "hint": "", "hintmode": null, "msg": "", "npoints": null,
"queuestate": null}, "i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_4_1":
{"correctness": "correct", "hint": "", "hintmode": null, "msg": "", "npoints": null,
"queuestate": null}}, "grade": 2, "max_grade": 3, "problem_id": "i4x://edx/AN101/problem/
a0effb954cca4759994f1ac9e9434bf4", "state": {"correct_map": {}, "done": null, "input_state":
{"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": {}, "i4x-edx-AN101-problem-
a0effb954cca4759994f1ac9e9434bf4_4_1": {}}, "seed": 1, "student_answers": {}}, "submission":
{"i4x-edx-AN101-problem-a0effb954cca4759994f1ac9e9434bf4_2_1": {"answer": "yellow", "correct":
false, "input_type": "optioninput", "question": "What color is the open ocean on a sunny day?",
"response_type": "optionresponse", "variant": ""}, "i4x-edx-AN101-problem-
a0effb954cca4759994f1ac9e9434bf4_4_1": {"answer": ["a piano", "a guitar"], "correct": true,
"input_type": "checkboxgroup", "question": "Which of the following are musical instruments?",
"response_type": "choiceresponse", "variant": ""}}, "success": "incorrect"}, "event_source":
"server", "event_type": "problem_check", "host": "precise64", "referer": "http:\/\/localhost:8001\/
container\/i4x:\/\/edX\/DemoX\/vertical\/69dedd38233a46fc89e4d7b5e8da1bf4?action=new",
"accept_language": "en-US,en;q=0.8","ip": "NN.N.N.N", "page": "x_module",
"time": 2014-03-03T16:19:05.584523+00:00", "username": "AAAAAAAAAA"}
```


Figura 0-8: edX – ejemplo de evento

Para registrar cada evento se recogen de cada acción un conjunto de campos en concreto:

- event_source
- event_type
- time
- username
- course_id

Para anonimizar las acciones de cada estudiante es necesario haber insertado en la base de datos el fichero ***{org}{course}{date}user_id_map{site}analytics.sql*** de manera que se

pueda comparar el **username** de cada evento con los que se obtienen del fichero y sustituir el **username** por el **id** de los usuarios mapeados previamente.

Dentro de cada campo se puede encontrar distintos valores, según que evento se este anonimizando o si el evento ha sido generado por el estudiante o es una respuesta del servidor. A continuación se va a explicar los posibles valores que tomarían.

Event_source - este campo contiene valores alfanuméricos que nos indican cual ha sido la interacción de origen que ha generado el evento. Los valores pueden ser **server**, **browser**, **mobile** y **task**.

Time - se trata de un campo que contiene la fecha y la hora en UTC en la que se ha registrado el evento. Tiene un formato especifico 'YYYY-MM-DDThh:mm:ss.xxxxxx'.

Course_id - se trata de un identificador asignado a la edición de un curso. Dependiendo del año en el que se haya generado el fichero podrá tener un formato u otro. Este identificador alfanumérico se corresponde con la combinación de los campos curso y edición que se almacena en la base de datos. De manera que si la edición del curso se ha registrado correctamente en la base de datos, todos los eventos podrán ser agrupados correctamente bajo su edición correspondiente de los cursos.

Event_type - se trata del campo fijo con mayor variabilidad del fichero. Se ha mantenido la clasificación de sub-categorías que describe edX para poder explicarlo de manera claro y concisa. El campo contiene valores alfabéticos que se clasifican según su funcionalidad.

- Enrollment Events
- Navigational Events
- Video Interaction Events
- Problem Interaction Events
- Certificate Events

Además de estas categorías existen eventos relacionados con funcionalidades de testeo o ejercicios y problemas básicos de edX. Estos eventos no han sido descartados pero todos ellos han sido catalogados en una clasificación propia de ELASA como *Otros*.

Enrollment Events

Se tratan de eventos generados cuando un estudiante se matricula en el curso, se desmatricula, cambia la modalidad del certificado que se le va a otorgar si aprueba el curso y eventos de comprobación interna del servidor sobre la correcta actualización de datos de estudiantes.

- edx.course.enrollment.activated - es un evento generado por el estudiante.
- edx.course.enrollment.deactivated - es un evento generado por el estudiante.
- edx.course.enrollment.mode_changed - es un evento generado por el estudiante.
- edx.course.enrollment.upgrade.clicked - es un evento generado por el servidor.
- edx.course.enrollment.upgrade.succeeded - es un evento generado por el servidor.

Navigarional Events

Se tratan de eventos generados cuando un estudiante se desplaza por el contenido del curso o por la propia estructura del curso.

- page_close - el estudiante ha cerrado el navegador o la pestaña donde tenía la web de edX abierta.
- seq_goto - el estudiante se desplaza a una unidad o sección sin seguir la secuencia establecida en el curso.
- seq_next - el estudiante salta a la siguiente unidad a la que se encuentra.
- seq_prev - el estudiante salta a la unidad previa a la que se encuentra.

Video Events

Se tratan de eventos generados cuando un estudiante interactúa con el contenido multimedia del curso.

- hide_transcript - el estudiante oculta los subtítulos de un video.
- load_video - el estudiante comienza a cargar un video.
- pause_video - el estudiante pausa un video.
- play_video - el estudiante comienza a visualizar un video o continua visualizandolo si anteriormente ha generado un evento pause_video.
- seek_video - el estudiante se mueve a los largo del video haciendo uso de la funcionalidad del reproductor.
- stop_video - El estudiante ha llegado al final del video. Esto no significa que el video se haya visto en su totalidad sino que la barra de progreso del reproductor ha llegado al final
- show_transcript - el estudiante activa los subtítulos de un video.
- speed_change_video - el estudiante cambia la velocidad de avance del video.

Problems Events

Se tratan de eventos generados cuando un estudiante realiza los problemas del curso.

- problem_check - el estudiante quiere corregir su problema. Este evento es lanzado por el estudiante sobre un problema y genera el evento problem_check desde el servidor.
- problem_check - se trata de un evento que solo se generado si precede el evento problem_check originario del estudiante. Este evento da la información pertinente sobre la puntuación de problema y los datos de la corrección.
- problem_save - el estudiante quiere guardar el progreso realizado en un problema pero no que sea corregido.
- problem_show - el estudiante ha activado la visualización de las respuestas a los problemas. Este evento nunca puede ser generado antes de realizar al menos un intento de resolver el problema.

C Estructura de ficheros con datos anonimización

A continuación se detalla la estructura de cada fichero que se puede generar automáticamente desde la herramienta y los datos que contienen.

C.1 Fichero MOOC-{course}-{edition}-Social.csv:

Los datos del fichero que contiene la extracción de los datos de los foros tiene una estructura con la que se podría realizar el grafo dirigido para visualizar las interacciones de los estudiantes entre ellos. A continuación se explica detalladamente cómo interpretar los datos y ejemplos de cada caso.

Columna	Descripción
User_ID	Identificador único del estudiante
CourseID	Identificador de la edición de un curso
ID_Thread	Identificador del mensaje escrito por el estudiante
Thread_Number	Identificador del mensaje escrito por el estudiante (si es igual a **ID_Thread** es el primer mensaje de una nueva conversación, si es distinto significa que es una contestación a un mensaje previo)
Thread_Parent	Identificador del mensaje escrito por el estudiante (si es nulo significa que será el primer mensaje de una nueva conversación)
Up_Count	Sumatorio de los votos positivos hacia el mensaje
Down_Count	Sumatorio de los votos negativos hacia el mensaje

Tabla 0-1: ELASA - Estructura Social.csv

Caso 1 - El estudiante crea un mensaje que al mismo tiempo es un nuevo hilo o conversación en el foro.

El **ID_Thread** es el identificador único del mensaje, el **Thread_Number** es el identificador único del hilo donde se ha alojado el mensaje. Por lo que en este caso ambos campos contendrán el mismo valor. El campo **Thread_Parent** será un campo que estará vacío, ya que el mensaje no es contestación a ningún mensaje. Se puede ver en la *Figura0-3* un ejemplo de este caso.

User_ID	CourseID	ID_Thread	Thread_Number	Thread_Parent	Up_Count	Down_Count
5358146	1	54ebcc382a472dc7ae000a7d	54ebcc382a472dc7ae000a7d	null	0	0
5052141	1	54ebdb7078f73339ab000aff	54ebdb7078f73339ab000aff	null	0	0
5492081	1	54ebe8f32a472d4797000a8a	54ebe8f32a472d4797000a8a	null	0	0
4257642	1	54ebec982a472d56ca000b0e	54ebec982a472d56ca000b0e	null	0	0
2005558	1	54ebf0412a472dc229000afe	54ebf0412a472dc229000afe	null	0	0
6049988	1	54ebff7e78f7331846000b87	54ebff7e78f7331846000b87	null	0	0
893988	1	54ec06e92a472de316000c0d	54ec06e92a472de316000c0d	null	0	0

Figura 0-9: ELASA – Ejemplo caso 1, fichero social.csv

Caso 2 - El estudiante crea un mensaje nuevo en un hilo/conversación ya existente en el foro.

El **ID_Thread** es el identificador único del nuevo mensaje, el **Thread_Number** es el identificador único del hilo donde se ha alojado el mensaje que será distinto del **ID_Thread**. El campo **Thread_Parent** será un campo que estará vacío, ya que el mensaje no es contestación a ningún mensaje sino al hilo/conversación. Se puede ver en la *Figura0-4* un ejemplo de este caso.

User_ID	CourseID	ID_Thread	Thread_Number	Thread_Parent	Up_Count	Down_Count
6522671	1	54ec676778f733320d000b4d	54ec676778f733320d000b4d	null	0	0
6529417	1	54ec77d52a472d2b72000bf3	54ec676778f733320d000b4d	null	0	0
6499074	1	54ec78222a472de316000c6c	54ec78222a472de316000c6c	null	0	0
5848142	1	54ec7c592a472da09c000b0a	54ec7c592a472da09c000b0a	null	0	0
2646561	1	54ec81c278f7332aab000bbd	54ec81c278f7332aab000bbd	null	0	0
6429182	1	54ec82dc78f733320d000b6e	54ec82dc78f733320d000b6e	null	1	0
6511428	1	54ec86f178f7336b5b000aff	54ec676778f733320d000b4d	null	0	0

Figura 0-10: ELASA – Ejemplo caso 2, fichero social.csv

Caso 3 - El estudiante crea un mensaje nuevo en un hilo/conversación ya existente en el foro como respuesta a un mensaje que no es el que abrió el hilo/conversación.

El **ID_Thread** es el identificador único del nuevo mensaje, el **Thread_Number** es el identificador único del hilo donde se ha alojado el mensaje que será distinto del **ID_Thread**. El campo **Thread_Parent** tendrá un identificador que será el valor único del mensaje al que se está respondiendo. Este caso es lo que en los foros de la web se denomina "*Citar*". Se puede ver en la *Figura0-5* un ejemplo de este caso.

User_ID	CourseID	ID_Thread	Thread_Number	Thread_Parent	Up_Count	Down_Count
6560220	1	54edddd22a472d2b72000d40	54ed9a5d78f733210a000cc5	54ed9bac78f733ac22000cdd	0	0
6572791	1	54eddf7e2a472d0b4b000c94	54ed9a5d78f733210a000cc5	null	0	0
6560220	1	54ede1842a472db234000d6a	54ede1842a472db234000d6a	null	0	0
6560220	1	54ede4dd78f7338fb6000cbe	54ed9a5d78f733210a000cc5	null	0	0
6560220	1	54ede5792a472d6515000d1e	54ede5792a472d6515000d1e	null	0	0
6572791	1	54edecba78f733803e000c8c	54edecba78f733803e000c8c	null	0	0
6560220	1	54edfadd2a472d85bb000da8	54ed53302a472db234000d1e	54ed6e2c78f7332af1000d01	0	0
5527913	1	54ee00a32a472d43ee000d79	54ed53302a472db234000d1e	54ed6e2c78f7332af1000d01	0	0

Figura 0-11: ELASA – Ejemplo caso 3, fichero social.csv

C.2 Fichero MOOC-{course}-{edition}-Eventos.csv:

En este fichero se volcará la información referente a una edición de un curso MOOC. Los posibles **Event_Type**, los **Event Source** y **Event Details** han sido explicados detalladamente en el Anexo B. Los valores que se extraerán en la columna **User_ID** serán los identificadores únicos creados por edX para cada alumno.

Columna	Descripción
User_ID	Identificador único del estudiante
CourseID	Identificador de la edición de un curso
Event_Type	Identificador del tipo de evento generado por el estudiante
Event_Source	Evento generado por el estudiante o por el servidor
Event_Details	Detalles sobre el Event_Type, para más información véase el Apartado A
Date_Time	Fecha y hora en la que se ha registrado el Event_Type

Tabla 0-2: ELASA - Estructura Eventos.csv

C.3 Fichero MOOC-{course}-{edition}-Certificados.csv:

El contenido del fichero de certificados incluye los **User_ID**, el identificador único de edX para cada estudiante, el valor asignado en base de datos a cada edición de un curso MOOC, la nota final (**Grade**), que está comprendida entre 0 y 1, y el **Mode,** este campo indica si los alumnos que han finalizado el curso van a obtener un certificado verificado o no. Los certificados verificados tienen mayor peso y es necesario realizar un ingreso en la plataforma edX para obtener esta modalidad.

Columna	Descripción
User_ID	Identificador único del estudiante
CourseID	Identificador de la edición de un curso
Grade	Nota final del alumno, comprendida entre 0 y 1
Mode	Valor alfabético que indica si se ha pagado por obtener un certificado edX verificado

Tabla 0-3: ELASA - Estructura Eventos.csv

D Planificación y manual de usuario

Con el fin de controlar el tiempo que se va a invertir en el Trabajo de Fin de Grado, se creó un diagrama de Gantt que modelizará todas las fases por las que iba a pasar el proyecto. La primera planificación no fue perfecta y se ha ido ajustando a la realidad del proyecto. Al mismo tiempo se han añadido tareas, marcadas como hitos, que resultaron en reuniones con mi tutora para definir aspectos importantes del diseño.

En aras de la claridad, en el diagrama de Gantt se muestra cada tarea que formo parte del desarrollo del TFG, se pueden ver en la *Figura 0-12*.

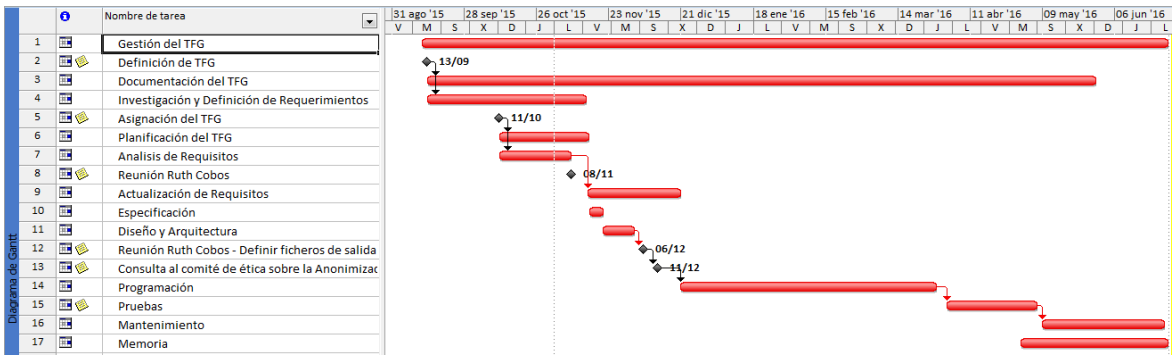

Figura 0-12: ELASA – Planificación para el desarrollo del Trabajo de Fin de Grado

La estructura de paquetes explicados en el capítulo 3.2 de Diseño, son los que se pueden ver en la *Figura 0-13*.

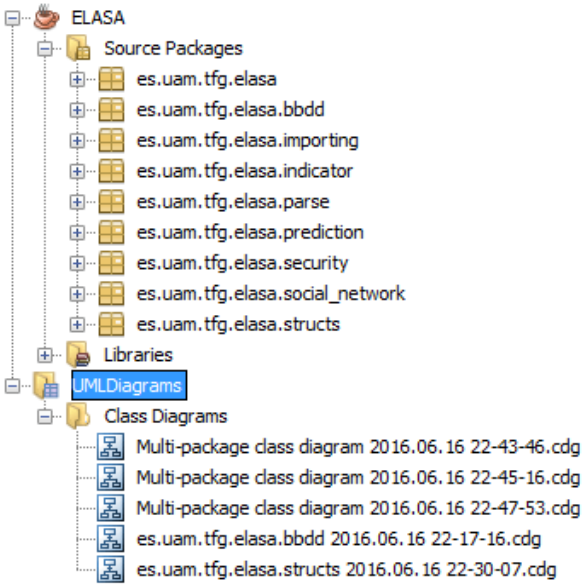

Figura 0-13: ELASA – Estructura de paquetes en el entorno de desarrollo

Cuando se ejecuta el sistema, el usuario de control se autentica contra la base de datos. Si es satisfactorio, el sistema desplegará el menú con todas las opciones como se ve en la *Figura 0-14*.

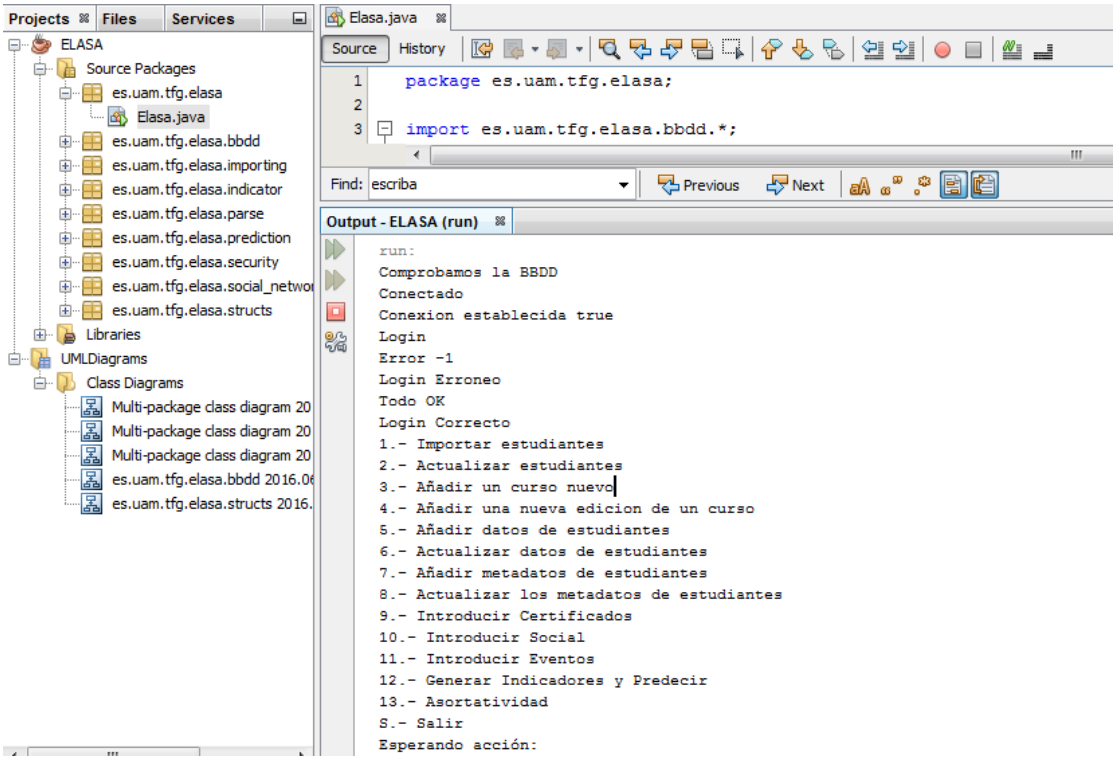

Figura 0-14: ELASA – Ejecución del sistema ELASA con la base de datos ya existente

A continuación, se va a explicar en detalle cada opción expuestas en la Figura previa. Al tratarse de una interfaz sencilla por menú, todo la funcionalidad se vería reflejada en la base de datos. Por ello se van a realizar breves explicaciones con capturas de las diferentes funcionalidades y su resultado en la base de datos. Por último, se verá el resultado de la predicción del curso Quijote501x - 1T2015, tras haber generado los indicadores.

En la Figura 0-15 se puede observar el entorno de gestión de bases de datos de XAMPP, seguidamente crearemos la base de datos "*elasa_bbdd_prueba_memoria*".

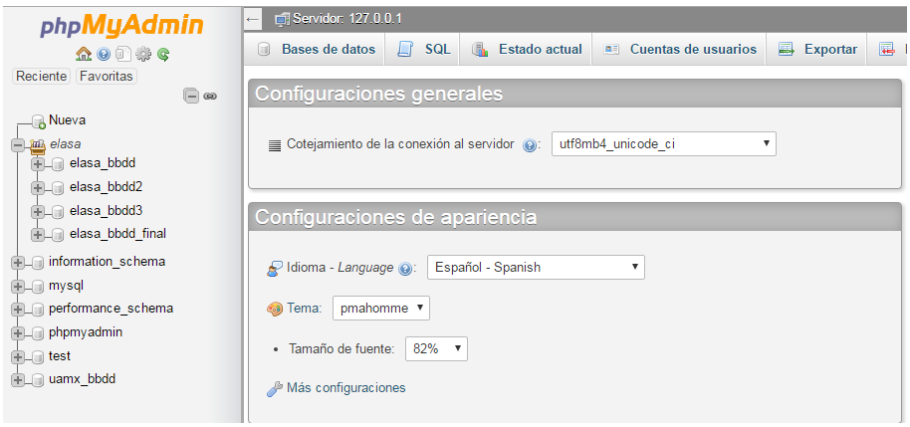

Figura 0-15: ELASA – Bases de datos existentes

Se ejecuta el sistema ELASA, la primera acción siempre será la de comprobar que existe la base de datos a la que se quiere conectar. Como se explica en la sección del desarrollo, punto 4.3.1.1, si no encuentra ninguna base de datos con las características definidas, el sistema automáticamente genera una nueva siguiendo el modelo explicado en la sección 3.3 Modelo de Datos, como se ve en la *Figura 0-16 y en la Figura 0-17*.

```
run:
Comprobamos la BBDD
Base de datos no encontrada
Conexion establecida true
Login
Error -1
Login Erroneo
Todo OK
Login Correcto
```

Figura 0-16: ELASA – Busqueda y creación de una nueva bases de datos

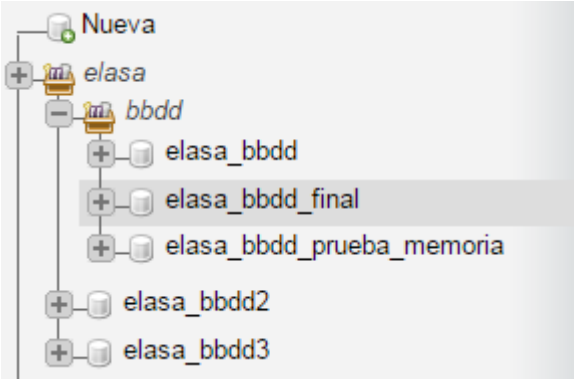

Figura 0-17: ELASA – Nueva base de datos elasa_bbdd_prueba_memoria

Se crea toda la estructura de tablas, claves primarias y externas para cada tabla.

Figura 0-18: ELASA – Nueva estrucutra de elasa_bbdd_prueba_memoria

A continuación, se lanzan las opciones 1 a la 11 para introducir los datos reales de los ficheros de la primera opción del Quijote, sin duda el proceso de carga de los ficheros correspondiente a los eventos son los que mayor carga de trabajo generan sobre la herramienta llevando los tiempo de ejecución a los 10 minutos para anonimizar e insertar

en la base de datos más de 400.000 registros. Una vez terminada la ejecución de cada acción la base de datos de "*elasa_bbdd_prueba_memoria*" queda rellena como se ve en la *Figura 0-19*.

Figura 0-19: ELASA – Estructura de "elasa_bbdd_prueba_memoria" con datos cargados

Se ejecuta la opción que dará el valor referente a la interacción de los estudiantes en el foro del curso de Quijote501x, el algoritmo de Asortatividad, se puede apreciar la salida en la *Figura 0-20*.

```
11.- Introducir Eventos
12.- Generar Indicadores y Predecir
13.- Asortatividad
S.- Salir
Esperando accion: 13
El valor de la Asortatividad para el curso 1T2015 es: -0.12470188025843428
1.- Importar estudiantes
2.- Actualizar estudiantes
3.- Añadir un curso nuevo
4.- Añadir una nueva edicion de un curso
5.- Añadir datos de estudiantes
6.- Actualizar datos de estudiantes
7.- Añadir metadatos de estudiantes
8.- Actualizar los metadatos de estudiantes
9.- Introducir Certificados
10.- Introducir Social
11.- Introducir Eventos
12.- Generar Indicadores y Predecir
13.- Asortatividad
S.- Salir
Esperando acción:
```

Figura 0-20: ELASA – Ejecución del algoritmo Asortatividad para el curso Quijote501x

Se ejecuta la opción que dará la predicción del algoritmo KNN para el curso Quijote501x. Al no incluir parámetros restrictivos de fechas de inicio y de fin la predicción se realizará sobre la totalidad de los estudiantes y contando con todos los eventos cargados en la base de datos, se puede apreciar el resultado en la *Figura 0-21*.

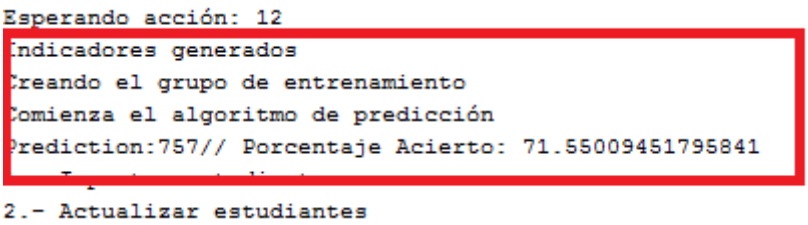

Figura 0-21: ELASA – Ejecución del algoritmo KNN de similitud para el curso Quijote501x

Como se pueden ver en la *Figura 0-20* y en la *Figura0-21*, el algoritmo devuelve el resultado dado en el capítulo 5 - Pruebas. Mientras que el algoritmo KNN no devuelve un valor exacto pero si en el rango que se muestra en la Tabla 5-8, esto es debido a que los conjuntos de entrenamiento son generados dinámicamente en cada ejecución, por lo que es casi imposible que se obtengan los mismos resultados en dos ocasiones.