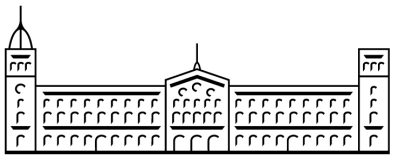

Trabajo final de grado

GRADO DE INGENIERÍA INFORMÁTICA

Facultad de Matemáticas
Universidad de Barcelona

Ciencia de los datos para el análisis de los resultados académicos: estudio estadístico de asignaturas

Autor: Daniel Gabriel Urdas

Director:	Dr. Laura Igual Muñoz
Realitzado en:	Departamento de Matemática Aplicada y Análisis

Barcelona, 27 de junio de 2015

Abstract

This final degree paper is part of an innovative teaching project from the Faculty of Mathematics which aims to create an intelligent support system that both studying tutor and head teacher can use in order to take better decisions in their tasks of supervision and support to students and teachers. This paper focuses mainly on the analysis of academics data in order to make a statistical study of the university subjects. It allows to track the evolution of the subjects and the relationship that may exist between different factors and academic performance of students. Therefore, the study will help identify the different problems or weaknesses of the teaching process and the trajectory of the students in different subjects over the years.

Resumen

Este trabajo final de grado forma parte de un proyecto de innovación docente de la Facultad de Matemáticas, que tiene como objetivo crear un sistema de soporte inteligente que tanto el tutor de estudios como el jefe de estudios puedan utilizar para tomar mejores decisiones en su labor de supervisión y soporte a los alumnos y profesores. Este trabajo se centra principalmente en el análisis de datos académicos para el estudio estadístico de las asignaturas. Permite realizar un seguimiento de la evolución de las asignaturas y de la relación que puede haber entre diferentes factores y los resultados académicos de los alumnos. Por lo tanto, el estudio permitirá observar los diferentes problemas o puntos débiles de la actividad docente y de la trayectoria de los alumnos en las diferentes asignaturas a lo largo de los años.

Resum

Aquest treball final de grau forma part d'un projecte d'innovació docent de la Facultat de Matemàtiques, que té com a objectiu crear un sistema de suport intel·ligent que tant el tutor d'estudis com el cap d'estudis puguin utilitzar per prendre millors decisions en la seva tasca de supervisió i suport als alumnes i professors. Aquest treball es centra principalment en l'anàlisi de dades acadèmiques per a l'estudi estadístic de les assignatures. Permet fer un seguiment de l'evolució de les assignatures i de la relació que hi pot haver entre diferents factors i els resultats acadèmics dels alumnes. Per tant, l'estudi permetrà observar els diferents problemes o punts febles de l'activitat docent i de la trajectòria dels alumnes en les diferents assignatures al llarg dels anys.

Índice

1. Introducción

En la Facultad de Matemáticas, para los grados de Ingeniería Informática y Matemáticas, se pone en marcha en el año 2009 el *Plan de Acción Tutorial (PAT)*. El plan supone la asignación de un tutor a cada alumno por parte de la Facultad. Éste tutor tiene como función guiar y aconsejar al estudiante en su proceso de aprendizaje, de modo que el alumno pueda tomar las mejores decisiones en cuanto a sus estudios universitarios y a su futura vida profesional. Así mismo, un tutor podría hacer un seguimiento académico para cada uno de sus alumnos asignados para poder optimizar los modelos de aprendizaje de cada estudiante.

Uno de los problemas detectados en este plan resulta ser que el tutor no dispone de todos los datos académicos y curriculares necesarios para poder hacer este seguimiento y así ofrecer una tutoría adaptada a las necesidades de cada alumno. Disponer de estos datos ayudaría a identificar y resolver problemas como los que se han observado hasta ahora. Por ejemplo, los estudiantes que tienen asignaturas convalidadas, especialmente de programación, tienen una mayor dificultad en el seguimiento del plan de estudios de las asignaturas relacionadas. Se ha detectado también que una mayoría de los alumnos que poseen malas calificaciones en las asignaturas de formación básica del primer curso, requieren una mayor dedicación para aprobar algunas de las asignaturas de segundo o tercer curso.

Teniendo en cuenta todo lo expuesto con anterioridad, surge la idea de iniciar un proyecto de innovación docente llamado *Sistema inteligente de soporte al tutor de estudios* [1]. El proyecto consiste en crear un sistema informático capaz de analizar todos los datos académicos y curriculares de los alumnos de la Facultad de Matemáticas y así ofrecer al tutor de estudios una herramienta que le ayude en el seguimiento y la trayectoria de cada alumno. También le ayudará a identificar los puntos fuertes y débiles de cada asignatura y los diferentes problemas de los planes de estudios de estas asignaturas.

Éste trabajo final de grado forma parte del proyecto de innovación y se centra principalmente en el estudio estadístico de las asignaturas de los grados de la Facultad de Matemáticas. A partir de los datos de registro y de los resultados académicos de los alumnos de la Facultad de Matemáticas, se hará un estudio exhaustivo para identificar y observar los diferentes problemas o puntos débiles de la actividad docente y de la trayectoria de los alumnos en las diferentes asignaturas a lo largo de los años. Con la identificación de estos problemas se ayudará tanto al tutor de estudios como al jefe de estudios a tomar las mejores decisiones con el fin de perfeccionar y

mejorar la docencia y los diferentes planes de estudios de las asignaturas.

El trabajo se centra en el seguimiento de las asignaturas y en la relación que puede haber entre los diferentes factores externos (vía de acceso, lugar de los estudios secundarios, nacionalidad, etc.) y los resultados académicos de los alumnos. Por una parte comprobaremos si factores como formación previa de los alumnos, las asignaturas convalidadas, la nota de acceso y otros datos, repercuten en la trayectoria de los alumnos y en la evolución de los resultados en cada asignatura estudiada. Por otra parte se estudiará la evolución de las medias generales en cada asignatura y se hará un estudio sobre la dificultad de estas asignaturas en función de las notas obtenidas por los estudiantes.

Utilizaremos tabmién un tipo de grafo interactivo que permite una mejor compresión de la información exraida de los datos. En este caso se mostrara mediante un grafo, las relaciones que pueden haber entre las asignaturas en función de sus calificaciones.

2. Motivación del trabajo

Éste trabajo final de grado forma parte del proyecto de innovación docente: *Sistema inteligente de soporte al tutor de estudios* [1] del Departamento de Matemática Aplicada y Análisis (MAIA) y del Departamento de Métodos de Investigación y Diagnostico en Educación (MIDE) de la Universidad de Barcelona.

Se trata de un proyecto docente que pretende crear un sistema informático automatizado capaz de procesar los datos académicos de los alumnos de la Facultad de Matemáticas. Éste sistema será una herramienta de soporte al tutor de estudios y al jefe de estudios, con la finalidad de ayudar a entender mejor la evolución académica de los alumnos y la evolución y calidad curricular de las asignaturas. Identificados los diferentes problemas que pueden haber, se podrán tomar decisiones para aumentar la calidad de la docencia y respectivamente las notas y los conocimientos adquiridos por el alumnado.

El proyecto está repartido en 5 fases diferentes:

1. **Fase 1**: Adquisición, ordenación, centralización y anonimización de los datos curriculares disponibles de los alumnos. Ésta fase consta de la obtención de los datos y la preparación de estos para poder utilizarlos en la siguientes fases de análisis.

2. **Fase 2**: Análisis de los datos mediante técnicas de la ciencia de los datos. En ésta fase se analizarán los datos mediante técnicas de estadística y ciencias de los datos que permitan hacer minería de datos.

3. **Fase 3**: Análisis de los datos mediante técnicas de aprendizaje automático. En ésta fase se aplicarán diferentes técnicas de Aprendizaje Artificial en función de cada pregunta planteada y de los datos de los que se disponga para hacer el análisis.

4. **Fase 4**: Desarrollo del sistema inteligente. Ésta fase representa la implementación del sistema inteligente que será una herramienta de soporte con gráficos y datos cuantitativos y cualitativos, tanto para el tutor de estudios como para el jefe de estudios. Además ayudará en la toma de decisiones sobre las acciones de mejora en las asignaturas y en los planos docente y tutorial.

5. **Fase 5**: Evaluación del sistema. En ésta fase se analizarán los resultados del sistema, la eficacia del mismo, y la identificación de los posibles errores producidos.

En éste trabajo final de grado se ha decidido implementar las fases 1, 2 y 3 del proyecto de innovación docente. Las fases 4 y 5 se implementarán en un futuro y no forman parte del trabajo.

3. Descripción del problema

3.1. Explicación de los datos

Los datos de los que disponemos tienen una amplia información sobre los alumnos matriculados en la Facultad de Matemáticas y sobre las asignaturas que están cursando. Se trata de los alumnos que cursan los grados de Ingeniería Informática y Matemáticas. Estos datos nos han sido proporcionados por el departamento de planificación y gestión académica de la Universidad de Barcelona.

Estructura de los datos

A continuación se detallará la estructura y la organización de los datos que utilizamos en el desarrollo del proyecto. Los datos se dividen en tres tablas:

Registo de alumnos

Esta tabla contiene la información de registro de todos los alumnos matriculados en los grados de Matemáticas e Ingeniería Informática desde el año 2009 hasta el año 2015. Esta es la información que proporciona el alumno al iniciar los estudios universitarios y realizar la primera matricula. En particular se dispone de la siguiente información de cada alumno: sexo, año de nacimiento, nacionalidad, información sobre si cursa dos grados a la vez, información sobre si el alumno dispone de la beca de carácter general, información sobre el lugar de los estudios de educación secundaria, información sobre el lugar de los estudios de ciclo formativo de grado superior si es el caso del alumno, la vía del acceso a la universidad, la nota de acceso, el año de obtención y el lugar de las PAU, el año de la primera matriculación, el grado cursado y si proviene de un sistema educativo extranjero.

Descripción asignaturas

Esta tabla dispone de información relacionada con todas las asignaturas de los dos grados de la Facultad de Matemáticas: Matemáticas e Ingeniería Informática. Contiene los siguientes campos: identificador del grado al que pertenece la asignatura, identificador de la asignatura, nombre, número de créditos e información sobre el curso y el semestre en los que se cursa la asignatura.

Calificaciones alumnos

Esta tabla contiene toda la información relacionada con las cualificaciones de los alumnos para cada asignatura cursada. Así mismo cada registro de la tabla contiene: el identificador del alumno, el curso en el que se matriculó la asignatura, el identificador del grado que el alumno está cursando, el identificador de la asignatura, la nota obtenida y el tipo de la cualificación (ordinaria, convalidada o reconocida).

Para disponer de consultas sobre toda la información disponible de las tres tablas haremos lo siguiente:

- Se combina la tabla que representa los registros de los alumnos con la tabla que representa las cualificaciones de los alumnos utilizando el identificador del alumno para poder relacionar la información de las dos tablas.
- Se combina la tabla resultante en el punto anterior con la tabla que representa la información de las asignaturas utilizando el identificador de la asignatura para poder relacionar la información de las dos tablas.

De este modo en la nueva tabla resultante tendremos, por cada registro, una cualificación junto con toda la información de aquel alumno y la información de la asignatura, siendo mucho más fácil el manejo de los datos.

Anonimización de los datos

Debido a que los datos contenían información personal de los alumnos, antes de recibir los datos, el departamento de planificación y gestión académica aplicó un proceso de anonimizacíon sobre los datos, de tal manera que no se pudiera identificar a ningún alumno utilizando identificadores personales como el DNI o el NIUB. Éste proceso constó en eliminar la información personal de los alumnos que aparecen en el registro (por ejemplo: DNI, NIUB, domicilio, teléfono, correo electrónico etc.), e identificar a los alumnos dentro de los archivos de la base de datos mediante un identificador numérico y aleatorio. Así mismo, se desconoce la manera y el orden de asignar estos identificadores a los alumnos.

Teniendo en cuenta que los datos utilizados para este proyecto no serán difundidos al público y solo serán de uso interno, esta anonimización se ha hecho sin seguir el protocolo establecido por la Universidad de Barcelona. Si en un futuro se decidiera hacer públicos estos datos, supondría el estudio

de las normas vigentes sobre el tratamiento de los datos con carácter personal y la respectiva modificación de los datos para que la desanonimización fuese imposible.

3.2. Ciencia de los datos

En términos generales, la ciencia de los datos es la extracción de conocimientos de dichos datos siendo una continuación del área de minería de datos. Utiliza técnicas de muchas áreas como: matemáticas, estadística, la ciencia de la información, la programación informática, la visualización de datos, la ingeniería de datos, etc.

Cualquier proyecto que emplea la ciencia de los datos se divide en varias fases [2]:

- ***Preguntas sobre el proyecto*** ¿Qué es lo que quiero conseguir? ¿Tiene sentido el planeamiento del proyecto?
- ***Obtención de los datos necesarios*** ¿De qué manera obtengo los datos necesarios al proyecto? "Web Scraping", consultas a bases de datos.
- ***Descripción del proyecto*** ¿Mediante que técnicas podemos entender el contenido de los datos? Aquí encontramos englobados varios procesos.
  - ***Procesamiento*** ¿Cómo se tienen que limpiar y/o separar los datos? Implica la filtración, la identificación de los outliers, la redimensionalidad, el procesamiento de los valores que faltan, la extracción de características y la normalización.
  - ***Agregación*** ¿Cómo se tienen que recolectar y resumir los datos? Mediante estadística básica: media, desviación estándar, box plots, scatter plots, etc.
  - ***Enriquecimiento*** ¿Cómo se añade más información a los datos? Mediante la clusterización (¿Cómo se segmentan los datos para encontrar grupos relacionados?) o mediante la visualización (¿Hay alguna relación inesperada entre los datos?)
- ***Descubrimiento*** ¿Cómo se relacionan los datos entre sí? Mediante la búsqueda de otras fuentes de datos.
- ***Análisis*** ¿Cómo modelamos nuestros datos? ¿Cómo identificamos las variables concluyentes? (Selección de variables) ¿Cómo están las variables relacionadas entre sí? (Modelado probabilístico).
- ***Predicción*** ¿Qué información se puede predecir a partir de los datos? Mediante regresiones, clasificaciones o recomendaciones.
- ***Evaluación*** ¿Son los resultados genéricos y robustos? Mediante regresiones, clasificaciones o recomendaciones. Implica las pruebas estadísticas y el rendimiento del modelo.

3.3. Etapas del proyecto

3.3.1. Planteamiento de las preguntas

Ésta es la primera etapa del proyecto, que consistió en plantear las preguntas que buscábamos resolver y estudiar la información que podríamos extraer del conjunto de datos. Durante ésta etapa también pudimos enfocar que funcionalidades debíamos tener en la nueva aplicación. Sin duda alguna se trata de una de las etapas más importantes del proyecto, pues en ella se consiguen establecer metas y definir su transcurso y desarrollo.

3.3.2. Obtención de los datos

Al inicio de ésta etapa planteamos los datos que nos serían necesarios para así poder solicitarlos, con la finalidad de llevar a cabo el proyecto. En función de las cuestiones planteadas, se ha realizado un estudio del tipo de datos y la información necesaria para cada pregunta. Básicamente los datos planteados se dividen en tres partes:

- Información de los alumnos. Representa toda aquella información de registro relevante sobre los alumnos.
- Información de las asignaturas. Representa toda aquella información sobre las asignaturas de los dos grados.
- Información de las cualificaciones. Representa toda aquella información sobre las cualificaciones de los alumnos.

Una vez planteada y definida la lista de datos que resultarían necesarios para el desarrollo del proyecto, se formalizó una petición de datos al Departamento de Planificación Académico-Docente dentro del Vicerrectorado de Política Docente. Los datos nos fueron proporcionados en archivos *Excel*, estructurados en las tres categorías arriba indicadas.

3.3.3. Procesamiento y limpieza de los datos

Esta etapa consiste en limpiar y separar los datos, de manera que no existan valores nulos entre los diferentes campos del registro o datos erróneos.

El procesamiento consistió en los siguientes puntos:

- **Unión de los grados de Ingeniería Informática**. Al introducirse el grado de Ingeniaría Informática en el año 2009, el código identificador del grado era *G1041*. Debido a que a partir del año 2011, el código cambió a ser *G1077*, se tuvieron que reemplazar todas las ocurrencias en los datos del código *G1041* por el código *G1077*.

- **Eliminación de las cualificaciones del segundo semestre del año 2015**. Se ha observado que en las asignaturas que los alumnos han matriculado y que aún no han sido cursadas (es decir las correspondientes al segundo semestre del año 2015), el campo de las cualificaciones consistía en un 0. Para resolver éste problema se han tenido que eliminar todas aquellas entradas de la tabla que tenían como curso el segundo semestre del año 2015.

- **Cambio de formato de los datos**. Para poder manejar y cargar mejor los datos en los programas que se han desarrollado, se ha decidido pasar los datos del *hojas de cálculo de Excel* a archivos *CSV*

- **Cambio de nombre de las columnas**. Con el fin de unir las diferentes tablas y relacionar la información, los nombres de las columnas con los cuales se hace la unión tienen que tener el mismo nombre.

3.3.4. Agregación de los datos

Dado que los datos utilizados para este proyecto no serán difundidos al público y solo serán de uso interno, esta anonimización se ha hecho sin seguir el protocolo establecido por la Universidad de Barcelona. Por este motivo, se ha decido realizar una prueba para ver realmente si los datos son anónimos o no. En el caso de que se constate que los datos no son totalmente anónimos, hará falta aplicar un proceso de anonimización suplementario que constará en la aplicación de ***técnicas de agregación*** sobre los datos.

3.3.5. Enriquecimiento de los datos

Debido a que necesitábamos información adicional sobre las asignaturas de los dos grados, nos dirigimos a la Secretaria de la Facultad de Matemáticas para obtener los datos que nos hacía falta en el análisis de algunos puntos del trabajo. La información obtenida consiste en:

- Información sobre el curso de las asignaturas (Es decir: primero, segundo, tercero o cuarto).

- Información sobre el semestre de las asignaturas (Es decir: primer semestre o segundo semestre).

En un futuro se ha planteado la obtención de más datos curriculares sobre los alumnos a través del *Campus Virtual de la Universidad de Barcelona*. Se pretende obtener información sobre las notas de las prácticas de laboratorio, las notas de los diferentes parciales que se llevan a cabo, de las horas de dedicación de los alumnos para una determinada asignatura, etc. Estos datos ayudarán a mejorar el análisis tanto inicial como predictivo

y así entender mejor los hábitos y los diferentes perfiles de alumnos que pueden haber.

3.3.6. Análisis de los datos

Ésta etapa consiste en responder a las preguntas planteadas en el en apartado *Planteamiento de las preguntas (Apartado 3.4.1)*. Mediante diferentes métodos estadísticos y utilizando un lenguaje de programación, Python, por cada pregunta se analizan y se procesan los datos. Siguiendo una serie de pasos, se llega a ofrecer unos resultados tanto visuales, mediante figuras y/o gráficos, como cuantitativos, mediante métricas de los diferentes métodos estadísticos que utilizamos.

3.3.7. Evaluación de los resultados

Una vez obtenidos los resultados de las preguntas planteadas, se tendrán que verificar y validar utilizando diferentes pruebas estadísticas. Estas pruebas nos ayudarán a entender mejor los diferentes resultados tanto visuales como cuantitativos. Se utiliza la prueba T de Student en los casos donde hay que verificar si dos grupos de datos son estadísticamente equivalentes, es decir, tienen la misma distribución de datos. Utilizaremos también las métricas de un regresor lineal para verificar si la recta de regresión lineal obtenida en una de las preguntas se ajusta a los datos y si ésta regresión es concluyente y de calidad. Además, se extraerán conclusiones y se analizará si tienen sentido en el contexto tratado.

3.4. Preguntas o Problemas planteados

En este proyecto nos hemos planteado una serie de preguntas que se podrán resolver mediante el análisis de los datos disponibles. Las preguntas son las siguientes:

1. **¿Hay asignaturas "outliers" que tienen notas de los alumnos muy diferentes del resto?**

   Con este planteamiento queremos verificar si hay asignaturas en las cuales los alumnos tienen notas muy diferentes respecto a las demás asignaturas o si las notas de alguna asignatura tienen una baja correlación en comparación con las notas de las demás asignaturas.

   Para resolver esta pregunta utilizaremos medidas como el coeficiente de correlación de Spearman para averiguar como se relacionan las asignaturas entre si.

2. **¿Hay diferencias entre las notas del primer curso en función de la vía de acceso de los alumnos?**

   Para plantear la pregunta se han de especificar primero las diferentes vías de acceso que hemos tenido en cuenta a la hora de hacer las pruebas. Tenemos a los alumnos separados en cuatro vías de acceso:

   a) Bachillerato con PAU. Son aquellos alumnos que tienen como estudios previos bachillerato y que han realizado las pruebas de acceso a la universidad (PAU).

   b) FP2 / CFGS. Son aquellos alumnos que tienen como estudios previos formaciones profesionales de segundo nivel o ciclos formativos de grado superior.

   c) Diplomado / Licenciado. Son aquellos alumnos que ya tienen uno más títulos universitarios al empezar los actuales estudios.

   d) Universitarios (Bachillerato con PAU / FP2 / CFGS). Son aquellos alumnos que han iniciado los estudios universitarios en otra universidad y han hecho el traslado a la universidad actual. Estos alumnos pueden tener como estudios previos bachillerato, formaciones profesionales de segundo nivel o ciclos formativos de grado superior.

   Queremos saber si hay diferencias entre las notas de los alumnos en función de sus vías de acceso para poder así identificar a los alumnos que podrían tener problemas y guiarlos mejor en relación a sus estudios.

3. **¿Hay alguna relación entre la nota de acceso a la universidad y las notas del primer curso de los alumnos?**

   Con este planteamiento lo que queremos comprobar es si la nota de acceso de un alumno repercute en las notas de dicho alumno en las asignaturas del primer curso. Es decir, si la media de las notas del primer curso es estadísticamente equivalente a la nota de acceso a la universidad. De ser así, mediante un modelo de predicción, se podrían predecir las medias de un determinado alumno a lo largo de sus estudios universitarios y orientarlo adecuadamente.

   Para llevar a cabo este planteamiento utilizaremos técnicas de regresión lineal y pruebas estadísticas.

4. **¿Hay alguna diferencia entre las notas del primer semestre respecto al segundo semestre de todos los cursos universitarios?**

   En este apartado lo que queremos averiguar es si los alumnos suelen obtener notas más altas en el primer semestre en relación al segundo semestre o viceversa. Esta pregunta surge pensando en si la dificultad de las asignaturas de un semestre es más alta o más baja, teniendo en cuenta otro semestre del mismo año o si los alumnos suelen estar más motivados en el primer semestre en comparación con el segundo.

5. **¿Las notas en las asignaturas de Matemáticas y Física son más altas para los alumnos que tienen como vía de acceso Bachillerato/Ciclos Formativos?**

   Queremos verificar una tendencia en la cual se observa que las notas en las asignaturas de Matemáticas y Física de los alumnos que tienen como vía de acceso Bachillerato son más altas respecto a los alumnos que provienen de otras vías de acceso. Para verificarlo, separaremos los alumnos en dos grupos dependiendo de sus respectivas vías de acceso y comprobaremos las medias de las asignaturas de Matemáticas y Física.

6. **¿Afectan las convalidaciones en asignaturas de programación sobre las notas de las asignaturas relacionadas del segundo o tercer curso?**

   Esta pregunta surge debido a una tendencia que ha sido observada por el profesorado. Los alumnos que provienen de formaciones profesionales de segundo nivel o ciclos formativos de grado superior, y que

tienen las asignaturas de programación del primer curso convalidadas, suelen tener notas más bajas en las asignaturas de programación del segundo y tercer curso. Esto es debido a que en las asignaturas de programación del segundo y tercer curso, se tienen que aplicar los conocimientos adquiridos en las asignaturas del primer curso.

7. **¿Hay algún patrón temporal importante en las notas de las asignaturas a lo largo de los años?**

   Verificaremos si hay patrones concluyentes en la evolución de las notas de los alumnos en todas las asignaturas a lo largo de los años. Identificaremos cuales son, si es posible, los factores que han provocado fluctuaciones en las medias de los alumnos. Por ejemplo, la aparición de la reevaluación en el año 2011.

8. **¿Hay algún ranking concluyente de asignaturas en función de su dificultad?**

   Intentaremos encontrar distintos patrones y agrupaciones de asignaturas en función de las notas de los alumnos. Por ejemplo podríamos encontrar que en las asignaturas de programación los alumnos tienen peores notas y que ésta observación no cambie al largo de los años.

9. **¿Cómo se debería realizar la anonimización de los datos?**

   Actualmente la anonimización realizada por el departamento de gestión académica que nos proporcionó los datos consiste en cambiar el NIUB de los alumnos por un identificador único que no tiene ninguna relación alfa-numérica con el *NIUB*. Es decir, sabiendo solo los identificadores de los alumnos es imposible saber de quién se trata realmente debido a que no se conoce la manera ni el orden de asignar los nuevos identificadores.

   Debido a que los datos no se tienen que publicar actualmente, no se ha aplicado ninguna técnica de anonimización potente. Por este motivo queremos saber si, disponiendo de varios datos sobre un sujeto en particular, se podría llegar a identificarlo.

4. Planificación

4.1. Tareas

Las tareas propuestas para este trabajo final de grado son equivalentes a las etapas del proyecto que hemos definido en el *Subapartado 3.4 "Etapas del proyecto"*. Las únicas diferencias constan en la presencia de dos nuevas tareas que representan la formación y estudios previos al inicio del trabajo y el desarrollo de la memoria del trabajo final de grado. Así mismo, las tareas son:

1. ***Formación y estudios previos***
2. ***Planteamiento de las preguntas***
3. ***Procesamiento y limpieza de los datos***
4. ***Enriquecimiento de los datos***
5. ***Agregación de los datos***
6. ***Análisis de los datos***
7. ***Evaluación de los resultados***
8. ***Realización de la memoria***

4.2. Diagrama de Gantt

Para mostrar el tiempo de dedicación previsto a lo largo del trabajo para las tareas definidas, utilizaremos un diagrama de Gantt. A continuación se muestra un diagrama de Gantt (*Figura 1*) con la planificación prevista del tiempo necesario para cada tarea y otro diagrama de Gantt (*Figura 2*) con los tiempos reales que ha necesitado cada tarea.

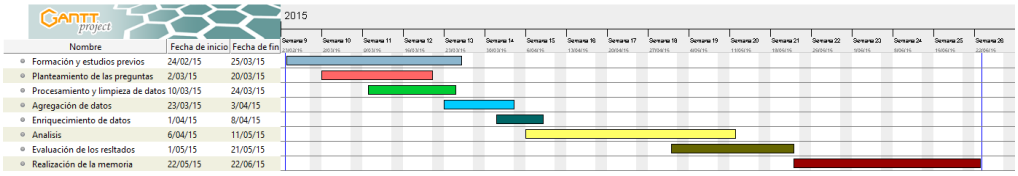

Figura 1: Planificación prevista

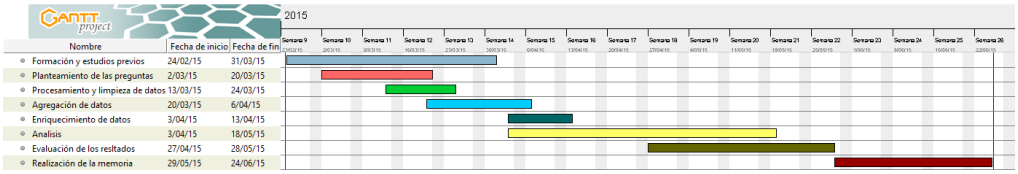

Figura 2: Planificación real

El grado de Ingeniería Informática tiene asignado para el Trabajo Final de Grado ***18 Créditos ECTS***. Debido a que en la Universidad de Barcelona cada crédito equivale a ***25 horas de dedicación***, el tiempo total que se ha de invertir es de ***450 horas***.

4.3. Evaluación económica

Teniendo en cuenta que el tiempo teórico que se ha de invertir para el desarrollo del trabajo final de grado es de 450 horas, asignamos a cada tarea el tiempo necesario para su finalización. Podemos ver en el *Cuadro 1* una estimación de la dotación económica necesaria para este trabajo.

Tarea	**Horas**	**€/hora**	**Precio total (€)**
Formación y estudios previos	25	0	0
Planteamiento de las preguntas	25	10	250
Procesamiento y limpieza	15	20	300
Enriquecimiento de los datos	35	20	700
Agregación de los datos	20	20	400
Análisis de los datos	175	25	4375
Evaluación de los resultados	75	25	1875
Realización de la memoria	80	0	0
TOTAL	**450**	-	**7900 (€)**

Cuadro 1: Evaluación económica del trabajo

5. Desarrollo del proyecto

5.1. Herramientas

5.1.1. Herramientas de soporte

Las herramientas de soporte son aquellas herramientas que he utilizado al largo del desarrollo del proyecto como ayuda complementaria para de organizar y estructurar el trabajo de una manera más eficiente y menos vulnerables ante pérdida de datos o errores humanos.

GitHub

Github [4] [3] es una plataforma de desarrollo colaborativo de software para alojar proyectos online utilizando el sistema de control de versiones Git. El sistema de control de versiones distribuido Git nos permite mantener una gran cantidad de código a una gran cantidad de programadores eficientemente. Además de las funciones ofrecidas por el Git, GitHub dispone de varias herramientas en línea muy útiles para el trabajo en equipo. Entre cuales podemos destacar:

- Wiki
- Sistema de seguimiento de incidencias
- Interfaz gráfica para revisión/comparación de código
- Visor de ramas de desarrollo

Bitbucket

Igual que GitHub, Bitbucket [5] es una plataforma de desarrollo colaborativo de software pero que además dispone de otro sistema de control de versiones llamado Mercurial. Otra ventaja que favorece al BitBucket es el número ilimitado de repositorios privados de los que puede disponer cualquier usuario. Los repositorios privados sirven especialmente para proyectos de código cerrado y que no permite la participación o la visualización a las personas que no están incluidas en el proyecto.

Trello

Trello [6] es un administrador de proyectos que nos permite descomponer y organizar el trabajo de un proyecto en varias tareas. Estas tareas se distribuyen en varias listas en función de su estado de desarrollo. A cada elemento de una lista se le puede añadir más elementos: otras listas,

imágenes, vídeos, documentos, etc. Además, es extremadamente potente para uso colaborativo ya que las tareas se pueden asignar a una persona o a un grupo extenso de personas, pudiendo añadir fechas límites.

Se basa en el método Kanban para gestión de proyectos, y contiene tarjetas que se pueden mover por diferentes listas en función de su estado de desarrollo. Un ejemplo básico de listas podría ser: una lista de cosas por hacer (to do, o pendientes), que se están haciendo (doing, o en proceso) o hechas (done, o terminadas).

5.1.2. Herramientas de edición

Las herramientas de edición son aquellas herramientas que nos permiten la edición y creación tanto de documentos de texto como de scripts y programas informáticos.

TeXnicCenter

TeXnicCenter [13] es una herramienta de edición que permite escribir textos que posteriormente serán compilados por el sistema de composición de textos llamado Latex [14]. Latex permite la creación de documentos extremadamente personalizables de una alta calidad tipográfica. Se utiliza especialmente para generar documentos, artículos y libros científicos o del área tecnológico.

IPython notebook

IPython Notebook [12] es un entorno computacional interactivo basado en la web para la creación y edición de cuadernos (notebooks) IPython. Un cuaderno IPython es un documento JSON que contiene una lista ordenada de las células de entrada / salida que pueden contener código escrito en el lenguaje de programación Python, texto, expresiones matemáticas, etc. Es una manera fácil de poder estructurar y explicar diferentes partes del código de un programa informático, haciéndolo más fácil y cómodo de seguirlo y entenderlo. Se utiliza especialmente para presentar proyectos científicos que conllevan la visualización de gráficos y esquemas.

5.1.3. Herramientas de programación

Las herramientas de programación son aquellas herramientas que permiten llevar a cabo, mediante lenguajes de programación, la implementación y la resolución de las preguntas del proyecto que hemos planteado con anterioridad. Gracias a estas herramientas los datos podrán ser procesados y

manipulados para poder extraer y obtener las características que nos interesan de los datos. Una vez obtenidos los resultados deseados, éstos se representaran mediante: tablas, gráficos, figuras, etc.

Python

Python [7] es un lenguaje de programación interpretado de alto nivel cuya sintaxis favorece la legibilidad del código. Es un lenguaje multiplataforma y soporta la programación orientada a objetos.

Es idóneo para el desarrollo de proyectos de data science debido a que dispone de una serie de librerías que permiten la manipulación de una gran cantidad de datos de una manera eficiente, fácil y rápida.

Librearía Pandas

Pandas [8] es una biblioteca para el lenguaje de programación Python para la manipulación y análisis de datos. Ofrece una serie de estructuras de datos que nos permiten almacenar y manipular tablas y listas de datos especialmente numéricos.

Librearía Bokeh

Bokeh [10] es una librería para Python que permite realizar gráficos y figuras interactivas para representar con una mayor expresión, una gran cantidad de tipos de datos. Ofrece la posibilidad de crear tanto gráficos de alto nivel (diagramas de caja, histogramas, diagramas de barras, etc.) como gráficos de bajo nivel que permiten una alta personalización.

Librearía Seaborn

Seaborn [11] igual que Bokeh es una librería para Python que permite representar datos estadísticos mediante gráficos y figuras. Tiene como base la conocida librería matplotlib añadiéndole más funcionalidades y un mejor aspecto. Tiene una alta integración con varias librerías PyData y permite utilizar como datos de entrada estructuraras específicas de numpy y pandas.

Librearía NumPy

NumPy [9] es una biblioteca para el lenguaje de programación Python que proporciona nuevos objetos para poder almacenar una gran cantidad

de datos mediante vectores, matrices u otros objetos multidimensionales. Dispone también de una serie de funciones matemáticas de alto nivel para realizar operaciones rápidas y eficientes con dichos objetos multidimensionales.

Librearía D3

D3 [15] es una librería JavaScript que permite crear todo tipo de figuras y gráficos interactivos y altamente personalizables a partir de una serie de datos. Estos gráficos utilizan los navegadores de Internet para que el usuario pueda interaccionar y se basan en las tecnologías *SVG*, *HTML5*, *CSS* y *JavaScript*.

5.2. Técnicas, métricas y gráficas utilizadas

5.2.1. T-Test

En estadística, la prueba T de Student o T-test [16] [17] [18] se utiliza para determinar si la diferencia entre las medias de dos muestras procedentes de poblaciones independientes y normales es significativa. Para efectuar la prueba asumimos que las muestras tienen una distribución normal.

Formula para muestras con varianzas iguales:

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_{x_1 x_2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde $s_{x_1 x_2} = \sqrt{\frac{(n_1-1)s_{x_1}^2 + (n_2-1)s_{x_2}^2}{n_1+n_2-2}}$

Formula para muestras con varianzas distintas, conocido como el T-Test de Welch:

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_{\overline{x_1}-\overline{x_2}}}$$

donde $s_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

El test es dos tipos:

- **La prueba T de Student para muestras pareadas** que se utiliza típicamente para evaluar la diferencia entre las medias de dos muestras que provienen de una misma población y que han sido evaluadas en ocasiones diferentes.
- **La prueba T de Student para muestras independientes** que se utiliza para verificar si la diferencia entre dos muestras que provienen de poblaciones distintas es estadísticamente significativa.

En nuestro caso para comparar las diferencias entre grupos de notas de diferentes asignaturas o años utilizaremos la T-Test para muestras independientes, ya que las notas son de poblaciones diferentes. Para proceder con el test primero se tiene que especificar el nivel de la probabilidad que estamos dispuestos a aceptar (0.05 es un valor común que se utiliza). Esta probabilidad definirá la aceptación o el rechazo de la hipótesis nula y de la hipótesis alternativa, que se explicarán a continuación. Para poder seguir con la evaluación de las dos muestras se han de definir las siguientes hipótesis:

- **La hipótesis nula** representa que las medias de las dos muestras son iguales.
- **La hipótesis alternativa** representa que las medias de las dos muestras no son iguales.

A continuación se realiza la prueba que nos devolverá el valor T y el valor P. El valor T nos muestra la diferencia de media que hay entre las dos muestras, mientras que el valor P nos dice si ésta diferencia es estadísticamente significativa. Una vez realizado el test, comparamos el valor P resultante con el nivel de la probabilidad que establecimos antes de empezar el test. Por lo tanto:

- **Si el valor P es mayor** que el nivel de probabilidad establecido, se acepta la hipótesis nula
- **Si el valor P es menor** que el nivel de probabilidad establecido, se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

5.2.2. Regresión lineal

En el campo de la estadística, la regresión lineal [19] es una técnica matemática utilizada para estudiar la relación entre dos o más variables.

Tanto si se tratan de dos variables (regresión simple) como de más variables (regresión múltiple), el análisis de una regresión lineal se utiliza para poder cuantificar y explorar las relaciones que pueden haber entre una variable que se llama dependiente (Y) y una o más variables que se llaman independientes o predictivas ($X_1, X_2, ..., X_k$), así como para implementar un modelo o ecuación lineal con fines predictivos sobre las variables dependientes. El análisis de una regresión lineal dispone además de varias métricas y procedimientos de diagnósticos (el coeficiente de determinación R^2, el análisis de residuos, los puntos de influencia, etc.) que nos proporcionarán datos sobre la idoneidad del análisis y calidad y el ajustamiento de la regresión sobre los datos.

¿Pero cómo podríamos describir y analizar las relaciones que puede haber entre estas variables?

Mediante una recta de regresión (*Figura 3*). Las restas de regresión son aquellas rectas que mejor se ajustan a la nube de puntos (llamados diagrama de dispersión de puntos).

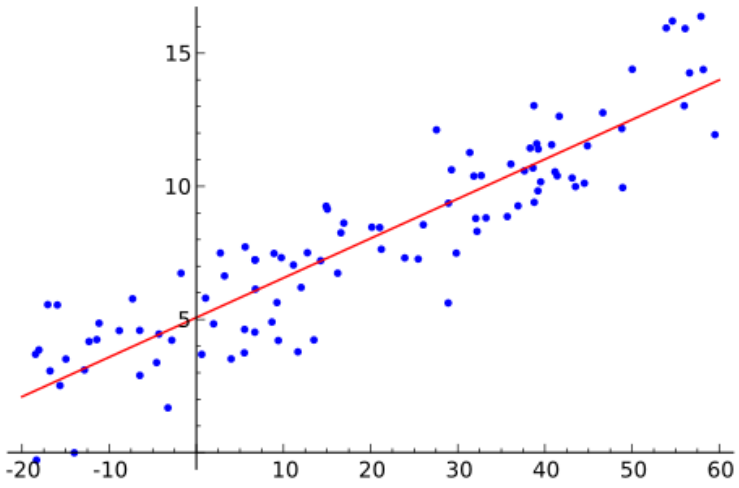

Figura 3: La nube de puntos y una posible recta de regresión lineal que se ajusta a estos puntos.

Para calcular la recta de regresión son posibles dos cálculos de máximo ajuste:

- La recta de regresión del eje Y sobre el eje X

$$y = \overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x})$$

- La recta de regresión del eje X sobre el eje Y

$$x = \overline{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \overline{y})$$

5.2.3. OLS - Mínimos cuadrados ordinarios

En el campo estadístico, los mínimos cuadrados ordinarios [20] son un modelo estadístico que forman parte de un grupo llamado Modelos de regresión y que tiene como objetivo estimar y/o predecir la media o el valor promedio poblacional de las variables dependientes en función de los valores ya conocidos llamados variables independientes. El método consiste en minimizar la suma de los cuadrados residuales, es decir, encontrar los estimadores que permitan que esta suma sea lo más pequeña posible.

Sea el modelo de la regresión lineal simple para estimar una variable dependiente Y_i en función de la variable independiente X_i:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

donde:

$i = 1, 2, 3, ..., n$
Y_i: Es la variable dependiente o estimada
X_i: Es la variable independiente
β_0: Es el parámetro que representa la intercesión o el termino constante
β_1: Es el parámetro respectivo a la variable independiente X_i
ε_i: Es el error asociado a la medición del valor X_i de modo que $\varepsilon_i \sim N(0, \sigma^2)$

Así mismo el método de los mínimos cuadrados ordinarios trata de seleccionar valores de los coeficientes β_0 y β_1 que resuelvan el problema:

$$(Minimizar_{\hat{\beta}_0, \hat{\beta}_1}) SCR = \sum_{i=1}^{N} \hat{u}_i^2$$

Nótese que el residuo asociado a cada observación $i, i = 1, 2, 3, ..., n$ depende de los valores de los coeficientes escogidos, porque:

$$\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

La solución a esta problema de optimización de errores residuales en función de los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$ se denomina estimador de Mínimos Cuadrados Ordinarios del modelo de regresión lineal simple. El estimador OLS elige, de entre todas las posibles rectas de regresión, la recta que tiene mínima la suma de los cuadrados de las distancias entre cada punto de la nube.

5.2.4. Diagrama de caja

Un diagrama de caja [16] [23] es un gráfico que pertenece a las herramientas de la estadística descriptiva y mediante el cual se pueden visualizar un conjunto de indicadores sobre una población. Éste tipo de gráfico es muy útil para detectar simetrías (si la mediana no está en el centro del rectángulo, la distribución no es simétrica) y para comparar dos o más variables. Suministra información sobre la distribución de los datos utilizando los siguientes descriptores y que se pueden observar en la *Figura 4*:

- el valor mínimo (el inicio de la semirrecta)
- el cuartil inferior (Q1 - el inicio de la caja)
- la mediana (la raya que aparece en la caja)
- el cuartil superior (Q3 - el final de la caja)

- el valor máximo (el final de la semirrecta)
- los valores atípicos de la distribución también llamados *outliers*

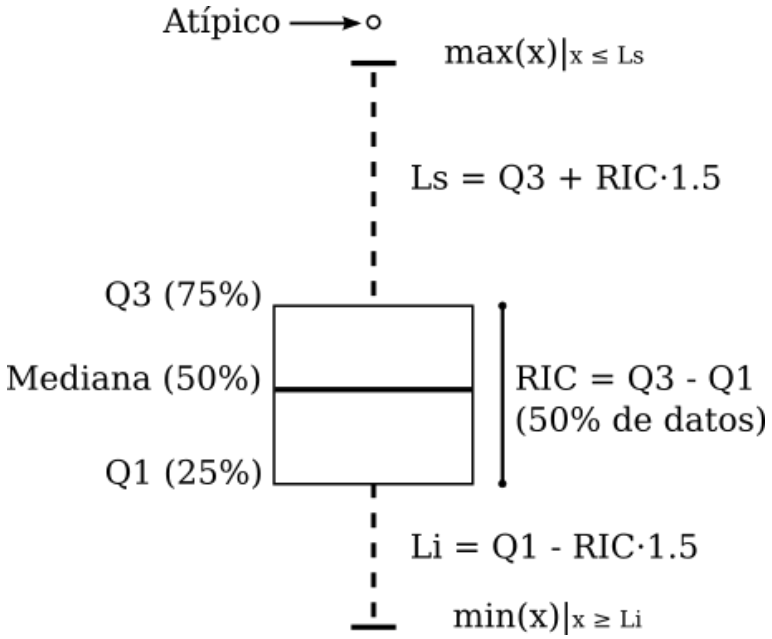

Figura 4: Un ejemplo de diagrama de caja y la indicación de los elementos presentes.

5.2.5. Diagrama de barras

El diagrama de barras [21] (o gráfico de barras) es un gráfico que se utiliza para representar datos de variables cualitativas o discretas. Está formado por barras rectangulares cuya altura es proporcional a la frecuencia de cada uno de los valores de la variable.

Existen varios tipos de gráficos de barras según las series de datos y como están estas representadas. Destacaremos dos de ellos:

Gráfico de barras sencillo (*Figura 5*). Representa los datos de una única serie o conjunto de datos.

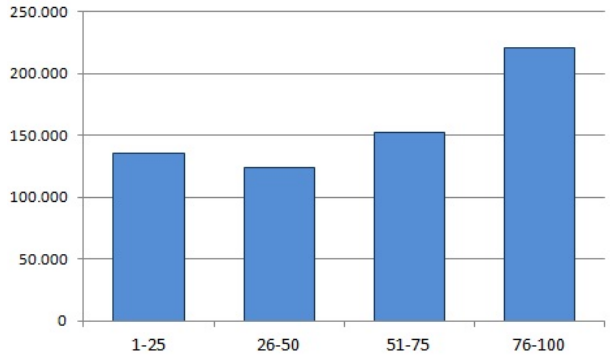

Figura 5: Un ejemplo de gráfico de barras sencillo.

Gráfico de barras agrupado (*Figura 6*).

- Representa los datos de dos o más series o conjuntos de datos.
- Cada serie se representa en un mismo color.
- Las barras se colocan una al lado de la otra por categoría de la variable para comparar las series de datos.

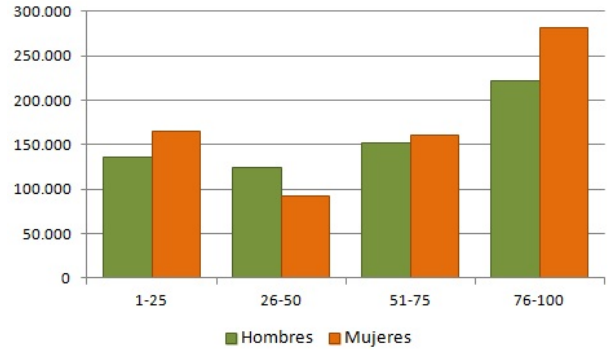

Figura 6: Un ejemplo de gráfico de barras agrupado.

5.2.6. F-Test

El F-Test [22] o la prueba de Fisher consiste en hacer un test sobre la igualdad de las varianzas de dos muestras, comparando un cociente ponderado de sus varianzas con los cuartiles de la ley de Fisher. Se utiliza para evaluar hipótesis que involucran múltiples parámetros. En una regresión lineal, la prueba sirve para determinar si una recta de regresión lineal encaja con el modelo de puntos y así saber si la regresión lineal calculada es viable.

$$Razón F = \frac{s_x^2}{s_w^2} = \frac{n s_{\overline{x}}^2}{(s_1^2 + s_2^2 + s_3^2 + ... + s_k^2)/k}$$

donde $s_k^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n-1}$, $s_{\overline{x}}^2 = \frac{\sum(\overline{x} - \overline{\overline{x}})}{n-1}$

k = grados de libertad y n = número de muestras

5.2.7. Correlación de Spearman

En el campo de la estadística, el *coeficiente de correlación de Spearman* o *rho de Spearman* [26] es una medida de la dependencia estadística entre dos variables y evalúa lo bien que se relacionan entre si las dos variables. El coeficiente determina si las dos variables están correlacionadas, es decir, si los valores de una variable tienden a ser más altos o más bajos para valores más altos o más bajos de la otra variable.

El coeficiente de correlación de Spearman se calcula utilizando la siguiente ecuación:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

donde $d_i = x_i - y_i$ es la diferencia entre los elementos de las dos listas ordenadas de modo creciente.

El resultado del coeficiente de correlación se encuentra en el rango [-1, 1] y se interpreta de la siguiente manera:

- Cercano a -1: correlación negativa
- Cercano a 0: sin correlación linear
- Cercano a 1: correlación positiva

5.2.8. Desviación típica o estándar

En estadística la desviación típica o estándar [25] es una medida de dispersión para variables de razón (cuantitativas o racionales) y de intervalo. Esta métrica nos dice cuanto tienden a alejarse los valores del promedio en una distribución. Se define como la raíz cuadrada de la varianza de la variable.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

donde $\overline{x}$ representa el promedio o la media artimetica de la distribución $x_1, x_2, x_3, ..., x_n$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

5.2.9. Grafo circular interactivo

Es un tipo de grafo que se utiliza para visualizar relaciones entre los diferentes puntos disponibles. Todos los puntos se posicionan en el mismo radio de un círculo y las relaciones son representadas de líneas curvas que pueden unir dos puntos. Es un gráfico interactivo con lo cual, se nos permite seleccionar cualquier punto del grafo para poder ver información más detallada sobre sus relaciones con los demás puntos. De este modo, si seleccionamos un punto podemos tener las siguientes 3 relaciones:

- El color de la línea es verde y el punto de destino es verde. En este caso la correlacion va desde el punto seleccionado hacia el punto de destino.
- El color de la línea es verde y el punto de destino es rojo. En este caso la correlación va tanto desde el punto seleccionado hacia el punto de destino como desde el punto de destino hacia el punto seleccionado. Se tiene que remarcar también que en este caso, el coeficiente de correlación es más alto desde el punto seleccionado
- El color de la línea es roja y el punto de destino es verde. En ese caso la correlacion va desde el punto de destino hacia el punto seleccionado.

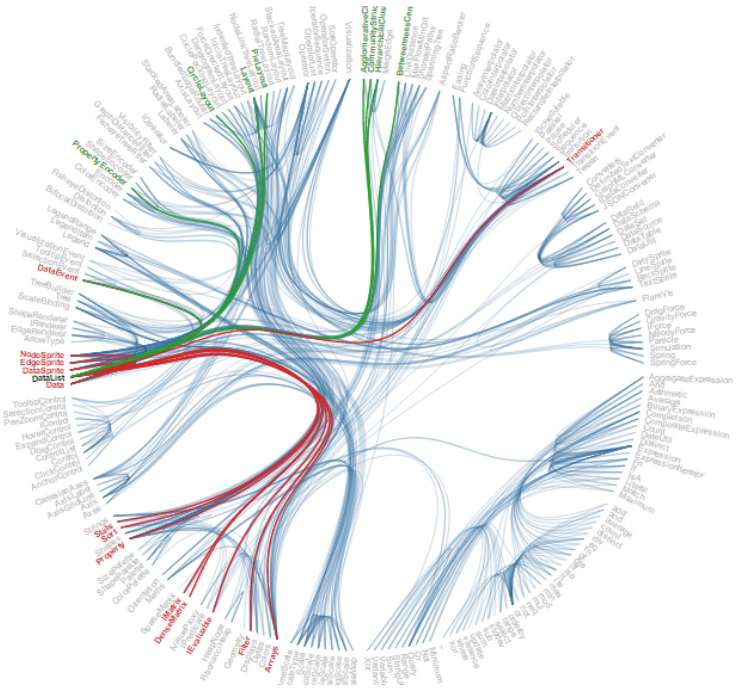

Figura 7: Un ejemplo de grafo circular interactivo.

6. Experimentos y resultados

6.1. Relación entre asignaturas y asignaturas outliers

En este aparatado se estudiarán las relaciones que pueden haber entre las asignaturas del grado de Ingeniería Informática. El estudio se hará a partir de las notas obtenidas por los alumnos en las diferentes asignaturas. Observaremos si las notas de los alumnos tienen distribuciones parecidas para diferentes asignaturas, o si al contrario, existen asignaturas con notas distribuidas de forma diferente al resto de asignaturas. También servirá para encontrar desproporcionalidades en las notas de una asignatura, es decir, si la gran mayoría de las notas se acumulan en la parte baja del rango de notas (notas muy bajas), o en la parte alta del rango de notas (notas muy altas).

Para empezar con la descripción del proceso, indicaremos las asignaturas implicadas:

- **Asignaturas del primer curso**: Programación I, Diseño Digital Básico, Introducción a los Ordenadores, Álgebra, Cálculo, Matemática Discreta, Física, Algorítmica y Programación II, Estructura de Datos
- **Asignaturas del segundo curso**: Probabilidades y Estadística, Empresa, Electrónica, Algorítmica avanzada, Introducción a la Computación Científica, Diseño de Software, Proyecto Integrado de Software, Estructura de Computadores, Programación de Arquitecturas Empotradas y Sistemas Operativos I
- **Asignaturas del tercer curso**: Sistemas Operativos II, Redes, Lógica y Lenguajes, Base de datos, Software Distribuido, Inteligencia Artificial, Visión Artificial, Taller de Nuevos Usos de la Informática, Factores Humanos y Computación y Gráficos y Visualización de Datos
- **Asignaturas del cuarto curso**: Ética y Legislación e Ingeniería del Software

Como se puede observar, para el cuarto año tenemos solo dos asignaturas. Esto es debido a que se han seleccionado solo las asignaturas obligatorias descartando las asignaturas optativas. Se ha tomado esta decisión debido a que para obtener el coeficiente de *correlación* entre las notas de las asignaturas, hacen falta grupos de datos con el mismo número de elementos. Es decir, para cada asignatura, tienen que haber el mismo número de notas de los mismos sujetos.

Se utilizará la ***correlación de Spearman*** para medir las relaciones entre las notas de las asignaturas. Éste tipo de correlación nos indicará lo bien que se relacionan dos variables en función de sus valores. Es decir, si tenemos dos listas de notas en orden creciente de dos asignaturas, cuando un valor de una lista aumenta, y el valor de la otra lista también aumenta, se obtendrá un buen coeficiente de correlación.

Para entender mejor como se obtiene el coeficiente de correlación de Spearman entre las notas de las asignaturas, haremos la prueba con las asignaturas del primer curso del grado de Ingeniería Informática.

	Prog. I	**Diseño D. B.**	**Álg.**	**Calc.**	**Algorít.**
Programación I	-	0.69	0.72	0.60	0.78
Diseño Digital Básico	0.69	-	0.67	0.65	0.73
Álgebra	0.72	0.69	-	0.74	0.71
Calculo	0.60	0.65	0.74	-	0.66
Algorítmica	0.79	0.73	0.71	0.66	-

Cuadro 2: Matriz de correlación entre las asignaturas del primer semestre del primer curso

Viendo los coeficientes de correlación presentados en el *Cuadro 2* podemos extraer alguna información sobre la relación entre las notas de las asignaturas. Por lo tanto:

- Se puede observar que las asignaturas de *Programación I* y *Algorítmica* tienen un coeficiente de correlación bastante alto, de 0.78 entre sí. Este puede ser debido a que las dos son asignaturas de programación y la distribución de las notas es parecida.
- Se puede observar también que las asignaturas de *Álgebra* y *Calculo* tienen un coeficiente de correlación bastante alto, de 0.74 entre sí. Este puede ser debido a que las dos son asignaturas de matemáticas y la distribución de las notas es parecida.

Se tiene que tener en cuenta que el rango de valores posibles del coeficiente de correlación es [-1, 1].

Una vez entendido el coeficiente de correlación, pasamos a realizar el proceso utilizando todas las asignaturas. Los resultados se presentarán mediante un grafo en un gráfico interactivo y que se construirá de la siguiente manera:

- Cada asignatura se relacionará con las 4 asignaturas más parecidas en función de sus coeficientes de correlación.
- Habrá un límite del coeficiente de correlación por debajo del cual la asignatura se considera *outlier*, es decir, no tiene ninguna relación con el resto de asignaturas. Teniendo en cuenta el promedio de todos los coeficientes de correlación, este límite se ha establecido en 0.5.

Figura 8: Figura que el grafo con las relaciones entre asignaturas

Como podemos observar en la *Figura 8*, una vez seleccionada la asignatura de Algorítmica Avanzada, nos aparecen tres tipos de relaciones:

- Las relaciones que tienen como color el ***rojo*** y que representan las asignaturas hacia las cuales existe una relación **desde** la asignatura seleccionada. En este caso, la asignatura seleccionada fue ***Algorítmica Avanzada*** y está directamente relacionada con las asignaturas: ***Programación 2***, ***Estructura de datos*** y ***Introducción a los Ordenadores***.

- Las relaciones que tienen como color el ***verde*** y que representan las asignaturas desde las cuales existe una relación **hacia** la asignatura seleccionada. En este caso, la asignatura ***Algorítmica*** está directamente relacionada con la asignatura seleccionada ***Algorítmica Avanzada***.

- Las relaciones que tienen como color el verde y el punto de destino es rojo. En este caso la correlación va tanto **desde** el punto seleccionado hacia el punto de destino como desde el punto de destino **hacia** el punto seleccionado. Se tiene que remarcar también que en este caso, el coeficiente de correlación es más alto desde el punto seleccionado. En este caso, el ejemplo está dado por la relación bidireccional entre ***Algorítmica Avanzada*** y ***Sistemas Operativos II***

También se puede observar la apariencia de algunas *asignaturas outliers*: ***Redes***, ***Factores Humanos y Computación***, ***Física***, ***Programación I*** y ***Trabajo Final de Grado***. La falta de relaciones puede ser causada por varios motivos. Vamos a ejemplificar dos de los casos:

- *Asignatura outlier* ***Trabajo Final de Grado***. Esta asignatura tiene una baja correlación con las demás asignaturas debido a que la distribución de las notas es diferente respecto al resto. La mayoría de las notas se encuentran en la parte alta del rango de evaluación [0, 10], es decir la mayoría de las notas son altas. También es debido a que prácticamente no hay alumnos con suspensos.

- *Asignatura outlier* ***Redes***. Esta asignatura tiene una baja correlación con las demás asignaturas debido a que la distribución de las notas es diferente respecto al resto. La mayoría de las notas se encuentran en la parte baja del rango de evaluación [0, 10], es decir la mayoría de los alumnos que han aprobado la asignatura, la han aprobado con notas muy cercanas al 5.

El gráfico interactivo se puede encontrar en la siguiente página web:
http://danielurdas.github.io/

6.2. Notas en función de la vía de acceso

Queremos saber si los alumnos del grado de Ingeniería Informática tienen mejores o peores notas en función de su vía de acceso, es decir, en función de sus estudios previos. Para realizar la prueba hemos separado los alumnos en función de sus vías de acceso de la siguiente manera:

1. Bachillerato con PAU. Son aquellos alumnos que tienen como estudios previos bachillerato y que han hecho las pruebas de acceso a la universidad (PAU).

2. FP2 / CFGS. Son aquellos alumnos que tienen como estudios previos formaciones profesionales de segundo nivel o ciclos formativos de grado superior.

3. Diplomado / Licenciado. Son aquellos alumnos que ya tienen uno más títulos universitarios al empezar los actuales estudios.

4. Universitarios (Bachillerato con PAU / FP2 / CFGS). Son aquellos alumnos que han iniciado los estudios universitarios en otra universidad y han hecho el traslado a la universidad actual. Estos alumnos pueden tener como estudios previos bachillerato, formaciones profesionales de segundo nivel o ciclos formativos de grado superior.

Una vez separadas las cualificaciones de los alumnos en función de la vía de acceso, creamos un diagrama de caja para observar e interpretar las diferentes semejanzas o diferencias entre los grupos de alumnos.

Como podemos ver en el diagrama de caja de la *Figura 9*, los estudiantes que poseen otro título universitario, tienden a tener notas más altas y con una distribución con menos varianza, es decir, la mayoría de las notas son altas y se acumulan en un intervalo más reducido. Al contrario, el resto de alumnos que tienen otras vías de acceso, tienen una distribución de las notas muy amplia y eso implica un promedio de la notas más bajo. En especial, se puede observar que los alumnos que provienen de Bachillerato tienen la distribución de las notas muy amplia, el 95 % de las notas se encuentran en el intervalo [≈1.5, 10].

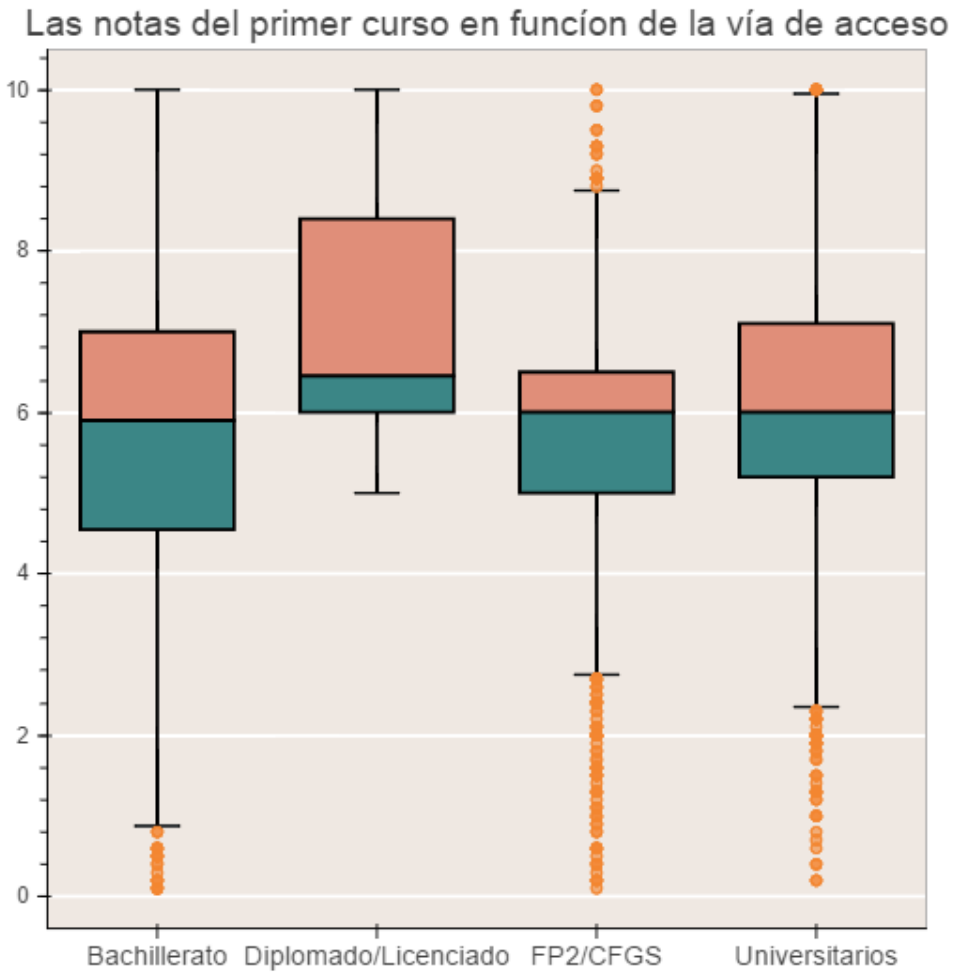

Figura 9: Figura que muestra un diagrama de caja con las cualificaciones de los alumnos en función de las vías de acceso

Más abajo en el *Cuadro 3* podemos ver diferentes métricas (*Media*, *Mediana*, *Varianza* y *Desviación estándar*) que nos ayudarán a entender mejor el diagrama de caja. Tal y como hemos comentado, se puede observar que mientras los alumnos que poseen otro título universitario tienen una media más alta y una varianza más baja, los alumnos procedentes de otras vías, en especial los que provienen de bachillerato, tienen medias más bajas y una varianza mucho más alta.

Vía de acceso	Media	Mediana	Varianza	Desv. estándar
Bachillerato	5.64	5.90	4.11	2.03
Diplomado/Licenciado	7.05	6.45	2.54	1.59
FP2/CFGS	5.62	6.00	3.20	1.79
Otra Universidad	6.03	6.00	2.78	1.67

Cuadro 3: Tabla que muestra diferentes métricas para las notas de los alumnos en función de sus vías de acceso

A continuación, utilizando la prueba estadística *T-Test*, verificaremos si las diferencias entre las notas de los alumnos en función de sus vías de acceso, son estadísticamente relevantes. Verificaremos primero las notas de los alumnos procedentes de Bachillerato respecto a las notas de los alumnos procedentes de Formaciones Profesionales y Ciclos Formativos de Grado Superior. Para continuar con la prueba, se definen dos hipótesis:

- **Hipótesis nula**: La media de la muestra de notas de los alumnos que tienen como vía de acceso Bachillerato es estadísticamente equivalente a la media de la muestra de notas de los alumnos que tienen como vía de acceso Ciclo formativo o formación profesional.
- **Hipótesis alternativa**: La media de la muestra de notas de los alumnos que tienen como vía de acceso Bachillerato es estadísticamente distinta a la media de la muestra de notas de los alumnos que tienen como vía de acceso Ciclo formativo o formación profesional.

Seguidamente efectuamos la prueba y obtenemos los siguientes resultados:

- ***T Estadístico*: 0.22**
- ***Valor P*: 0.83**

Como podemos ver, debido a que el ***Valor P* es igual a 0.83**, es decir, **mayor que 0.05**, se acepta la hipótesis nula y se confirma que la media de las dos muestras es estadísticamente equivalente.

Pasamos a ver un ejemplo donde la distribución de las notas de dos grupos es diferente tal y como se puede observar en el diagrama de caja, y aplicaremos la prueba *T-Test* para confirmarlo. Se aplicará la prueba sobre las notas de los alumnos que ya poseen un título universitario, respectivamente de los alumnos que han iniciado los estudios en otra universidad. Definimos las dos hipótesis necesarias para la prueba:

- **Hipótesis nula**: La media de la muestra de notas de los alumnos que ya poseen un título universitario es estadísticamente equivalente a la media de la muestra de notas de los alumnos que han iniciado los estudios en otra universidad.
- **Hipótesis alternativa**: La media de la muestra de notas de los alumnos que ya poseen un título universitario es estadísticamente distinta a la media de la muestra de notas de los alumnos que han iniciado los estudios en otra universidad.

Seguidamente efectuamos la prueba y obtenemos los siguientes resultados:

- ***T Estadístico*: 0.22**
- ***Valor P*: 0.04**

Como podemos ver debido a que el ***Valor P* es igual a 0.04**, es decir, **menor que 0.05**, se rechaza la hipótesis nula y se acepta la hipótesis alternativa y se confirma que la media de las dos muestras es estadísticamente distinta.

6.3. Relación entre la nota de acceso a la universidad y las notas del primer curso universitario

Queremos comprobar si la nota de acceso de un alumno repercute en las notas de dicho alumno en las asignaturas del primer curso. Es decir, si la media de las notas del primer curso es estadísticamente equivalente a la nota de acceso a la universidad. De ser así, mediante un modelo de predicción basado en una regresión lineal, se podrían predecir las medias de un determinado alumno al largo de sus estudios universitarios.

Por lo tanto para continuar con la prueba hemos seleccionado todas las notas de las asignaturas del primer curso, para el grado de Ingeniería Informática.

Debido a que las notas de acceso a la universidad tienen un rango de [0, 14] y las notas de la asignaturas un rango de [0, 10], se han tenido que rebajar a 10 todas aquellas notas de acceso a la universidad mayores que 10. Hemos decidido tomar esta acción para tener los mismos rangos para los dos ejes y así poder obtener correctamente regresión lineal.

Una vez obtenidas las muestras con las notas de acceso a la universidad y las notas de las asignaturas de primer curso, procedemos a calcular una recta de regresión lineal. Cada punto de la nube de puntos representa a un alumno con su respectiva nota de acceso y su media de las asignaturas del primer curso. A continuación podemos observar el resultado de la regresión lineal y la recta calculada en la *Figura 10*.

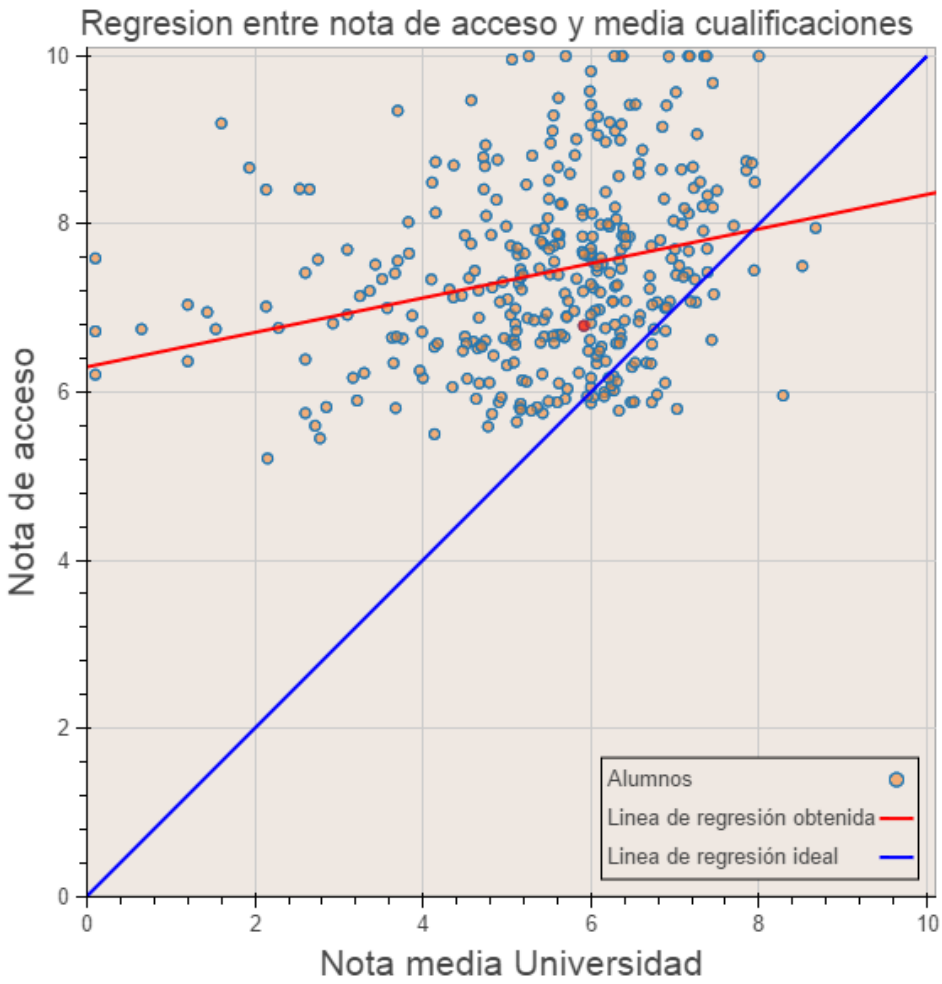

Figura 10: Regresión lineal entre la nota de acceso y la media de las cualificaciones del primer curso

Como podemos observar la nube de puntos no sigue a una recta, y visualmente se puede observar que la recta de regresión lineal es inviable teniendo en cuenta los puntos dados. Una recta de regresión ideal sería aquella donde cada alumno tendría la misma media de las notas que la nota de acceso a la universidad. Ésta recta de regresión ideal se puede ver en la *Figura 10* de color azul.

También podemos observar que la mayoría de los puntos se encuentran por encima de la recta de regresión ideal. Eso quiere decir que una gran parte de los alumnos obtienen una media de las asignaturas más baja que la nota de acceso a la universidad. En la *Figura 11* se puede ver que el 90.5 % de los alumnos realmente obtienen una media de las asignaturas más baja que la nota de acceso a la universidad.

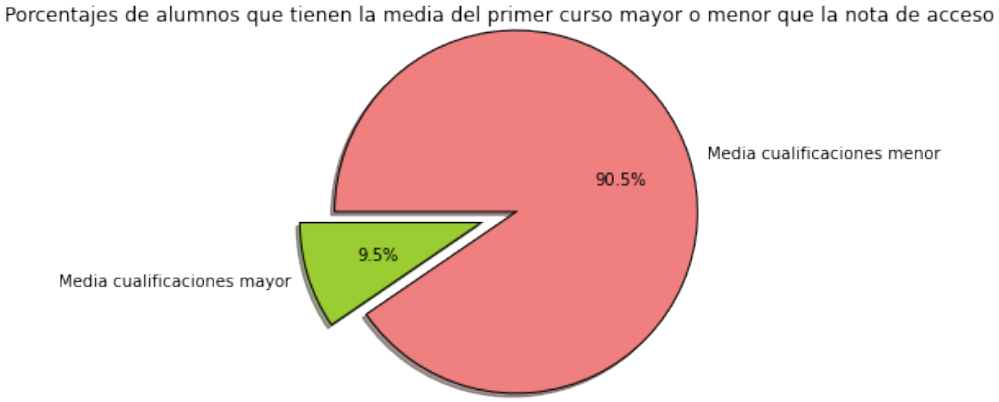

Figura 11: Porcentajes de alumnos que tienen la media del primer curso mayor o menor que la nota de acceso

A continuación, para confirmar la inviabilidad de la regresión lineal, realizaremos la prueba F-Test sobre las dos muestras de datos. Para continuar con la prueba se definen las dos hipótesis:

- **Hipótesis nula**: La media y la desviación estándar de los dos grupos de datos son estadísticamente equivalentes.
- **Hipótesis alternativa**: La media y la desviación estándar de los dos grupos de datos son estadísticamente distintos.

Seguidamente efectuamos la prueba y obtenemos los siguientes resultados:

- ***F Estadístico*: 4779**
- ***Valor P*: $3.29 \cdot 10^{-204}$**

Como podemos ver debido a que el ***Valor P* es igual a $3.29 \cdot 10^{-204}$**, es decir, **menor que 0.05**, se rechaza la hipótesis nula y se acepta la hipótesis alternativa y se confirma que la media y la desviación estándar de las dos muestras es estadísticamente distinta.

6.4. Comparación de notas entre el primero y segundo semestre

Queremos averiguar si los alumnos suelen obtener notas más altas en el primer semestre respecto al segundo semestre o viceversa. Este planteamiento surge pensando en si la dificultad de las asignaturas de un semestre es más alta o más baja teniendo en cuenta otro semestre del mismo año o si los alumnos suelen estar más motivados en el primer semestre en comparación con el segundo.

Para realizar las pruebas se han separado los alumnos del grado de Ingeniería Informática de los alumnos del grado de Matemáticas para que las pruebas tengan más coherencia.

A continuación podemos ver en la *Figura 12* un diagrama de barras con la comparación entre las medias de las notas del primer y segundo semestre de todos los cursos del grado de Ingeniería Informática.

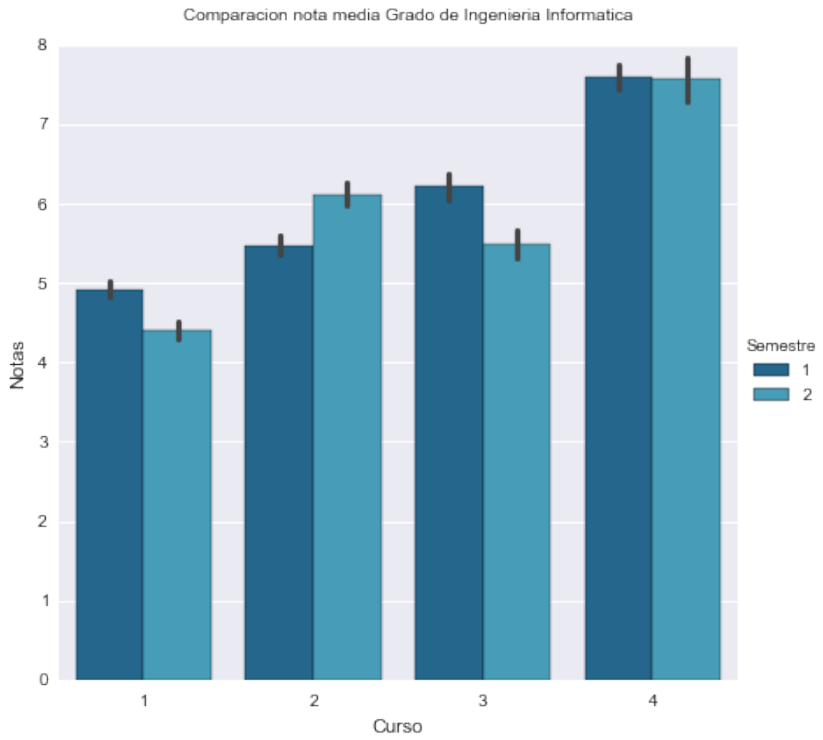

Figura 12: Comparación entre las medias de las notas del primer y segundo semestre de todos los cursos del grado de Ingeniería Informática

Como podemos observar en la *Figura 12*, no hay una diferencia importante entre las medias de las notas de los dos semestres de los alumnos de Ingeniería Informática. Las únicas diferencias notables se encuentran en las

medias de los cursos *1*, *2* y *3*, pero estas diferencias nos superan el *0.5*.

Pasamos a ver un diagrama de cajas en la *Figura 13* para obtener más información sobre la distribución de las notas. Éste diagrama nos proporcionará información sobre la dispersión de las notas, pudiendo así entender mejor las posibles diferencias que pueden haber entre los dos semestres de todos los cursos.

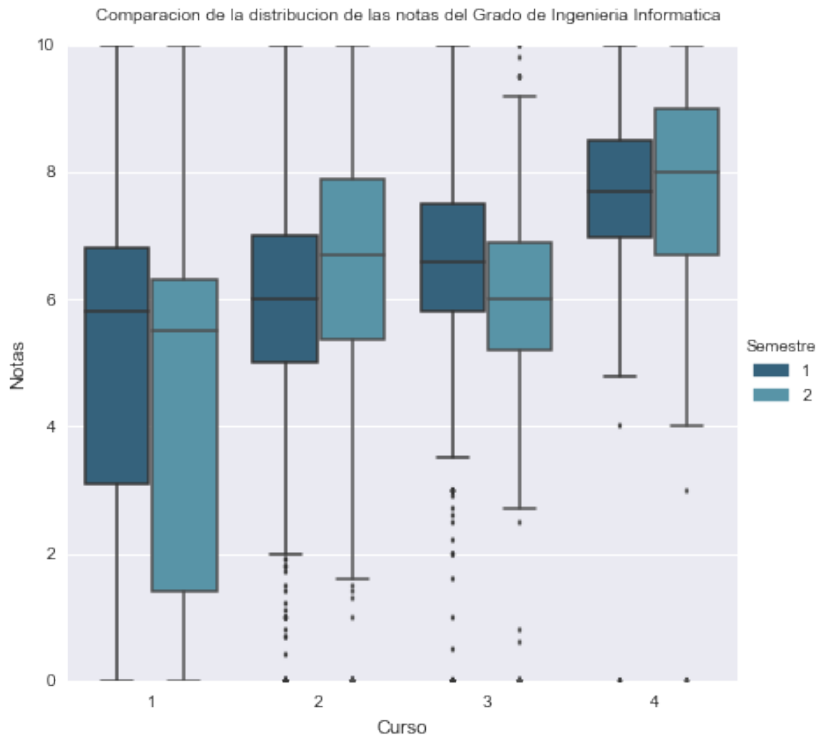

Figura 13: Comparación de la distribución de las notas para el grado de Ingeniería Informática

Observando la *Figura 13* podemos deducir que aunque las medias de los dos semestres son muy parecidas, la distribución de las notas es bastante distinta ya que, como se puede observar en el ejemplo del primer curso, mientras que el 50 % de las notas del primer semestre se encuentran en el intervalo $\approx$[3.1, 6.7], en el segundo semestre se encuentran en el intervalo $\approx$[1.5, 6.2].

Mirando el *Cuadro 4* se puede observar que las desviaciones estándar de las notas del segundo semestre son siempre mayores que las desviaciones estándar de las notas del primer semestre. Eso quiere decir que la dispersión de las notas es más extensa en el segundo semestre.

Curso y semestre	**Media**	**Desviación estándar**
Primer curso - primer semestre	4.92	2.79
Primer curso - segundo semestre	4.40	2.83
Segundo curso - primer semestre	5.48	2.39
Segundo curso - segundo semestre	6.12	2.58
Tercer curso - primer semestre	6.21	2.11
Tercer curso - segundo semestre	5.49	2.39
Cuarto curso - primer semestre	7.59	1.48
Cuarto curso - segundo semestre	7.57	2.12

Cuadro 4: Medias y desviaciones estándar para las notas según el curso y el semestre del grado de Ingeniería Informática.

Pasamos a observar los mismos estudios y diagramas pero ahora para el grado de Matemáticas. Como podemos observar en la *Figura 14*, las medias de los dos semestres es casi igual entre sí. Esto se aplica a todos los cursos excepto para el primer curso donde la media del segundo semestre es considerablemente más baja respecto al primer semestre.

Si miramos la *Figura 15* que representa el diagrama de cajas, podemos observar que para el segundo semestre del primer curso, el 75 % de las notas se encuentran en el intervalo $\approx[0, 6]$. Esto puede ser debido a la gran dificultad de las asignaturas de este semestre o del gran abandono de los alumnos durante éste.

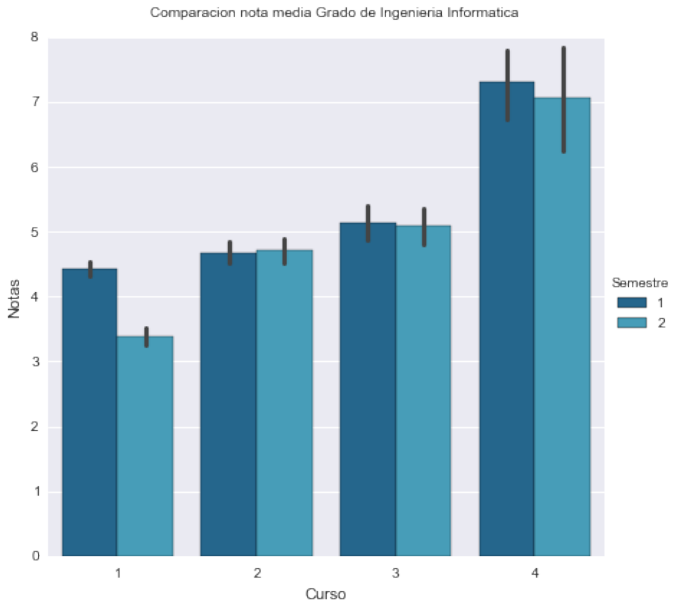

Figura 14: Comparación entre las medias de las notas del primer y segundo semestre de todos los cursos del grado de Matemáticas.

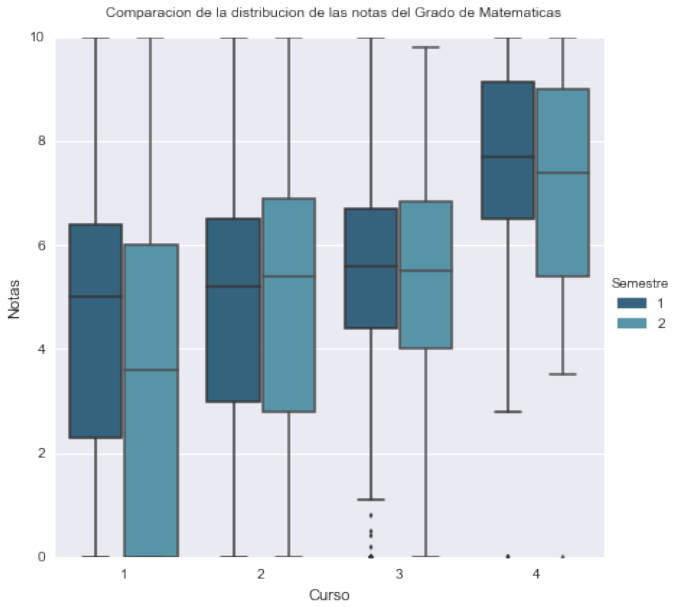

Figura 15: Comparación de la distribución de las notas para el grado de Matemáticas.

Curso y semestre	**Media**	**Desviación estándar**
Primer curso - primer semestre	4.43	2.76
Primer curso - segundo semestre	3.38	3.06
Segundo curso - primer semestre	4.68	2.68
Segundo curso - segundo semestre	4.71	2.86
Tercer curso - primer semestre	5.13	2.45
Tercer curso - segundo semestre	5.09	2.51
Cuarto curso - primer semestre	7.31	2.52
Cuarto curso - segundo semestre	7.07	2.22

Cuadro 5: Medias y desviaciones estándar para las notas según el curso y el semestre del grado de Matemáticas.

6.5. Notas más altas en las asignaturas de Matemáticas y Física para los alumnos de provienen de Bachillerato

Queremos verificar una tendencia en la cual se observa que las notas en las asignaturas de Matemáticas y Física de los alumnos que tienen como vía de acceso Bachillerato son más altas respecto a los alumnos que provienen de otras vías de acceso. Para verificarlo, separaremos los alumnos en dos grupos dependiendo de sus respectivas vías de acceso y comprobaremos las medias de las asignaturas de Matemáticas y Física.

Ésta tendencia se debe a que los alumnos que estudian Bachillerato cursan asignaturas de Matemáticas y Física en sus estudios pre-universitarios, mientras que los que cursan un Ciclo formativo de grado superior o una Formación profesional no lo hacen. Dado este motivo, es de esperar que los alumnos provenientes de Grados superiores y Formaciones profesionales tengan más dificultades y peores calificaciones en las asignaturas de Matemáticas y Física de los estudios universitarios.

Para ver las posibles diferencias, utilizaremos un diagrama de cajas para interpretar las notas de las asignaturas de Matemáticas y Física de los dos grupos de alumnos. En la *Figura 16*, podemos observar que realmente hay una diferencia notable entre las notas de los dos grupos de alumnos. Se puede ver que las notas de los alumnos que provienen de *CFGS* y *FP2* son más bajas y que mientras que el 50 % de las notas de los alumnos que provienen de Bachillerato se encuentran en el intervalo $\approx$[3.8, 7.2], en el caso de los alumnos que provienen de *CFGS* y *FP2*, el 50 % de las notas se encuentran en el intervalo $\approx$[2.2, 6.5]. Nótese también que un número considerable de notas del intervalo del 50 % se encuentran por debajo de la mediana.

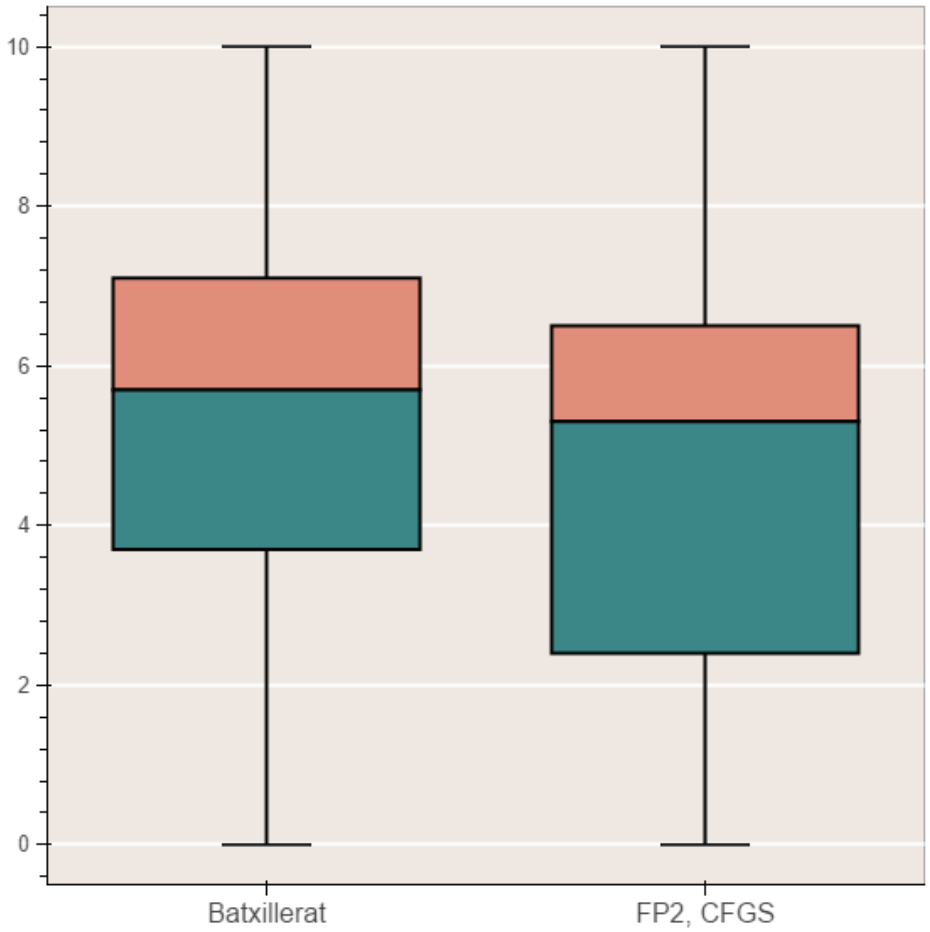

Figura 16: Diagrama de caja con las calificaciones de las asignaturas de Matemáticas y Física

A continuación realizaremos la prueba estadística T-Test para verificar si las medias de las notas de los dos grupos de alumnos son estadísticamente distintas y así confirmar la tendencia explicada al principio del apartado. Para continuar con la prueba se definen las dos hipótesis:

- **Hipótesis nula**: La media de la muestra de notas de los alumnos que provienen de *Bachillerato* es estadísticamente equivalente a la media de la muestra de notas de los alumnos que provienen de *CFGS* y *FP2*.
- **Hipótesis alternativa**: La media de la muestra de notas de los alumnos que provienen de *Bachillerato* es estadísticamente distinta a la media de la muestra de notas de los alumnos que provienen de *CFGS* y *FP2*.

Seguidamente efectuamos la prueba y obtenemos los siguientes resultados:

- ***T Estadístico*: 4.26**
- ***Valor P*: $2.26 \cdot 10^{-5}$**

Como podemos ver debido a que el ***Valor P* es igual a $2.26 \cdot 10^{-5}$**, es decir, **menor que 0.05**, se rechaza la hipótesis nula y se acepta la hipótesis alternativa y se confirma que la media de las dos muestras es estadísticamente distinta.

Más abajo en el *Cuadro 6* podemos ver las diferentes métricas que nos ayudarán entender mejor el diagrama de cajas y las explicaciones hechas.

Métrica	**Alumnos Bachillerato**	**Alumnos CFGS y FP2**
Media	5.21	4.54
Desviación estándar	2.66	2.72
Mínimo	0.00	0.00
Cuartil 25 %	3.70	2.40
Cuartil 50 %	5.70	5.30
Cuartil 75 %	7.10	6.50
Máximo	10.00	10.00

Cuadro 6: Tabla que muestra diferentes métricas para las notas de los alumnos en función de sus vías de acceso

6.6. Impacto de las convalidaciones en las notas de las asignaturas de programación del segundo o tercer curso

Se quiere validar una tendencia que se ha observado indicando que los alumnos que provienen de formaciones profesionales de segundo nivel o ciclos formativos de grado superior, y que tienen las asignaturas de programación del primer curso convalidadas, suelen tener notas más bajas en las asignaturas de programación del segundo y tercer curso. Esto se cree que es debido a que en las asignaturas de programación del segundo y tercer curso, se tienen que aplicar los conocimientos adquiridos en las asignaturas del primer curso. Dado que muchos de los alumnos que convalidan asignaturas de programación, tienen conocimientos escasos debidos a estas convalidaciones, a la hora de aplicar los conocimientos necesarios, tienen dificultades en seguir el plano docente de la asignatura.

En este estudio nos centraremos en la asignatura de ***Proyecto Integrado de Software*** del grado de Ingeniería Informática, ya que para llevar a cabo las prácticas de laboratorio de esta asignatura hace falta aplicar todos los conocimientos adquiridos previamente en las asignaturas: ***Programación I***, ***Programación II*** y ***Diseño de Software***.

A continuación, veremos en el *Cuadro 7* algunas métricas que confirmarán el hecho de que los alumnos con asignaturas de programación convalidadas tienen peores notas en ***Proyecto Integrado de Software***, en comparación con los alumnos que hayan estudiado las asignaturas de programación previas en la universidad.

Métrica	**No Convalidados**	**Convalidados**
Media	8.10	6.81
Desviación estándar	1.39	1.18
Mínimo	3.00	5.50
Cuartil 25 %	7.50	6.00
Cuartil 50 %	8	6.00
Cuartil 75 %	9	7.50
Máximo	10.00	10.00

Cuadro 7: Tabla que muestra diferentes métricas para las notas de los alumnos en función de si tiene las asignaturas de programación convalidadas o no.

Podemos remarcar las diferencias en cuanto a la media y a los diferentes

cuartiles que nos indican que aunque los alumnos con asignaturas de programación convalidadas aprueban ***Proyecto Integrado de Software***, lo hacen un 75 % de las veces con notas inferiores a 7.50, mientras que el resto de alumnos tienen notas inferiores a 9.

Vamos a ver lo explicado anteriormente, expuesto en el diagrama de cajas de la *Figura 17*.

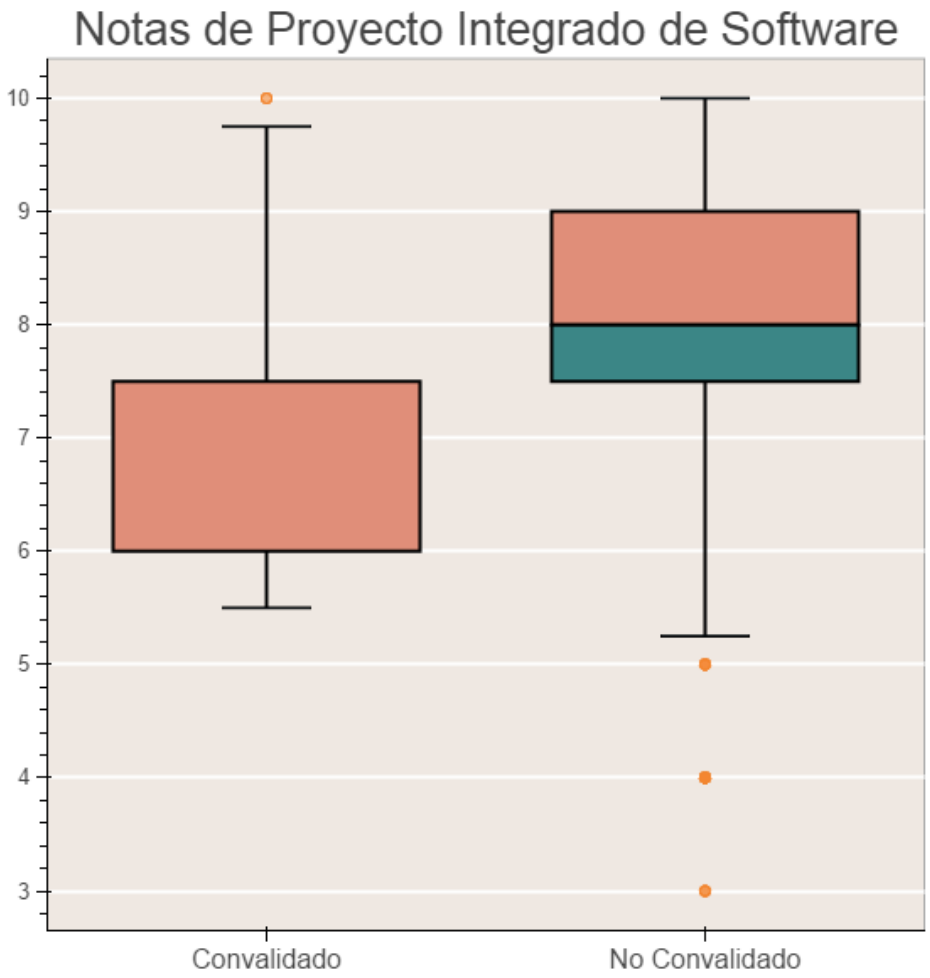

Figura 17: Diagrama de caja con las cualificaciones de Proyecto Integrado de Software

A continuación realizaremos la prueba estadística T-Test para verificar si las medias de las notas de los dos grupos de alumnos son estadísticamente distintas y así confirmar la tendencia explicada al principio del apartado. Para continuar con la prueba se definen las dos hipótesis:

- **Hipótesis nula**: La media de la muestra de notas de ***Proyecto Integrado de Software*** de los alumnos que tienen las asignaturas básicas de programación convalidadas es estadísticamente equivalente a la media de la muestra de notas de ***Proyecto Integrado de Software*** de los alumnos que no tienen las asignaturas básicas de programación convalidadas.

- **Hipótesis alternativa**: La media de la muestra de notas de ***Proyecto Integrado de Software*** de los alumnos que tienen las asignaturas básicas de programación convalidadas es estadísticamente distinta a la media de la muestra de notas de ***Proyecto Integrado de Software*** de los alumnos que no tienen las asignaturas básicas de programación convalidadas.

Seguidamente efectuamos la prueba y obtenemos los siguientes resultados:

- ***T Estadístico*: -7.39**

- ***Valor P*: $8.89 \cdot 10^{-11}$**

Como podemos ver debido a que el ***Valor P* es igual a $8.89 \cdot 10^{-11}$**, es decir, **menor que 0.05**, se rechaza la hipótesis nula y se acepta la hipótesis alternativa y se confirma que la media de las dos muestras es estadísticamente distinta.

6.7. Patrones temporales importantes en las notas de las asignaturas al largo de los años

Queremos verificar si hay patrones concluyentes en la evolución de las notas de los alumnos en todas las asignaturas a lo largo de los años. Identificaremos cuales son los posibles factores que han provocado fluctuaciones en las medias de los alumnos.

Para realizar las pruebas, se han separado las notas de los alumnos en función del curso de las asignaturas y del año en el que se han cursado. Así mismo, se hará un estudio sobre la evolución de las notas por cada uno de los cuatro cursos universitarios. Los datos se representarán en una figura en función de la media de las asignaturas, mostrando al mismo tiempo el cambio entre los años.

Pasamos a ver la evolución de las notas para las asignaturas del grado de Ingeniería Informática en: la *Figura 18*, la *Figura 19*, la *Figura 20* y la *Figura 21*.

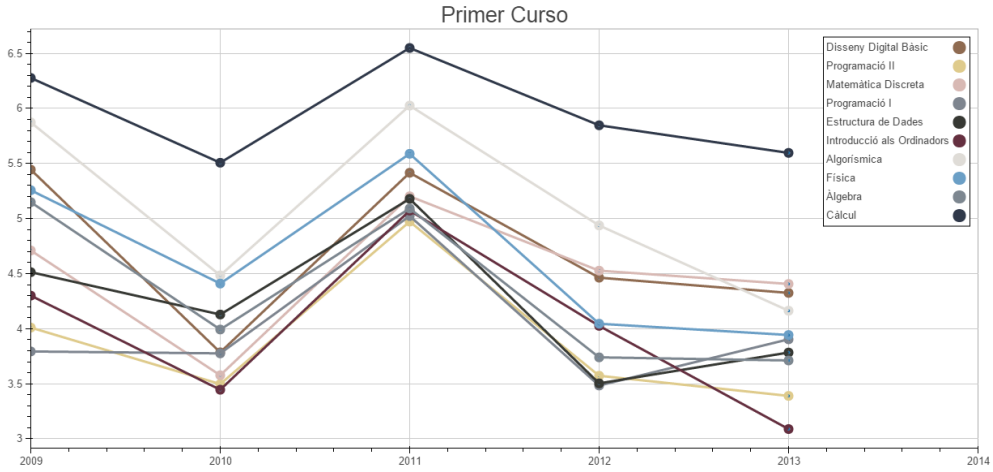

Figura 18: Evolución de las medias para las asignaturas del primer curso de Ingeniería Informática

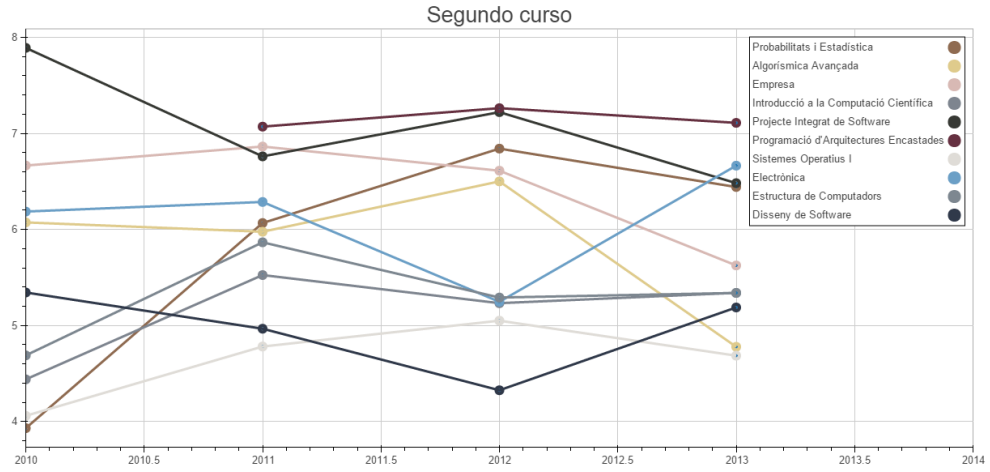

Figura 19: Evolución de las medias para las asignaturas del segundo curso de Ingeniería Informática

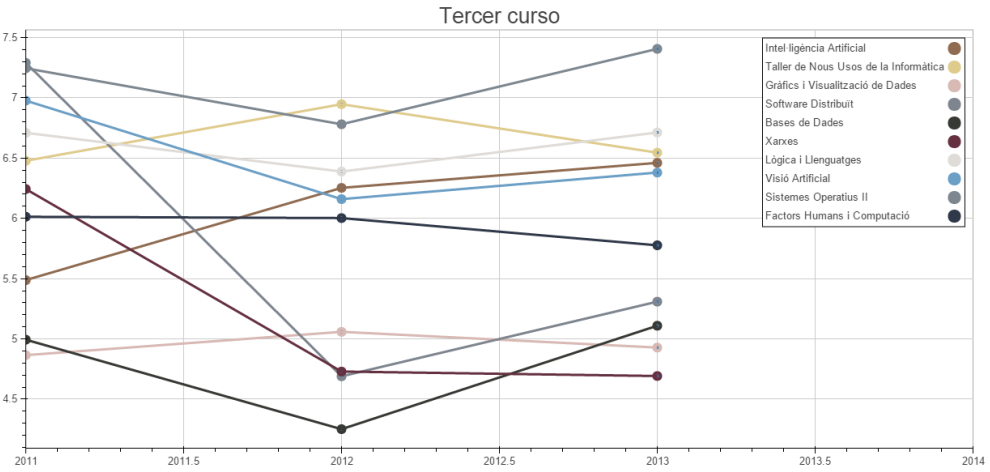

Figura 20: Evolución de las medias para las asignaturas del tercer curso de Ingeniería Informática

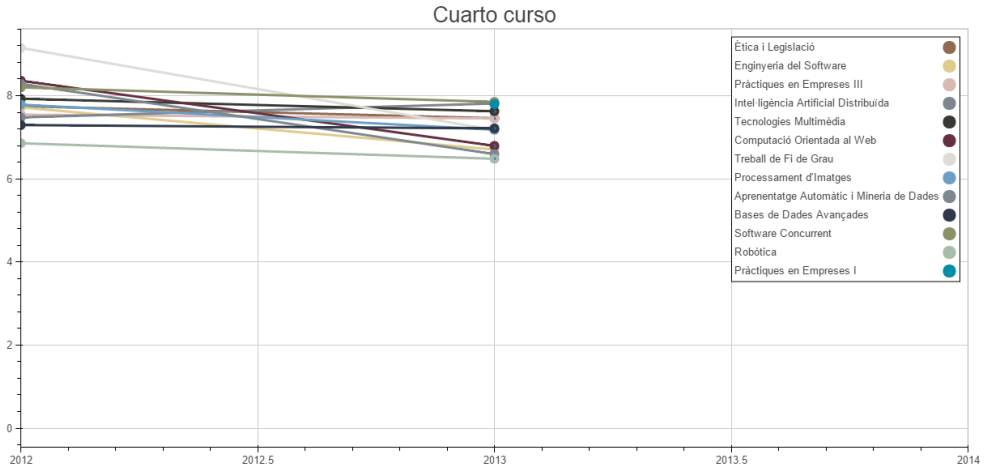

Figura 21: Evolución de las medias para las asignaturas del cuarto curso de Ingeniería Informática

Podemos observar en la *Figura 18*, que corresponde a las medias de los alumnos de las asignaturas del primer curso de Ingeniería Informática, que hay un importante cambio sobre las medias del año ***2011*** respecto a las medias del año ***2010***. Sabemos que ***el año 2011*** fue cuando se adoptaron los exámenes de ***reevalución*** para los alumnos del grado de Ingeniaría Informática. Tal y como se puede observar, esta reevaluación tuvo un impacto positivo sobre las notas de los alumnos. En el *Cuadro 8* podemos ver información sobre las medias y las desviaciones estándar de los notas de las asignaturas de los dos años.

Año	**Media**	**Desviación estandar**
2010	4.03	2.85
2011	5.42	2.45

Cuadro 8: Medias y desviaciones estándar para las medias de las asignaturas del primer curso de Ingeniería Informática

Realizaremos la prueba estadística ***T-Test*** sobre las notas de los alumnos del primer curso del año ***2010***, respectivamente ***2011***, para verificar si esta diferencia es estadísticamente relevante. Para continuar con la prueba se definen las dos hipótesis:

- **Hipótesis nula**: La media de la muestra de notas del primer curso del ***año 2010*** es estadísticamente equivalente a la media de la muestra de notas de del primer curso del ***año 2011***.
- **Hipótesis alternativa**: La media de la muestra de notas del primer

curso del ***año 2010*** es estadísticamente distinta a la media de la muestra de notas de del primer curso del ***año 2011***.

Seguidamente efectuamos la prueba y obtenemos los siguientes resultados:

- ***T Estadístico*: -11.97**
- ***Valor P*: $1.88 \cdot 10^{-31}$**

Como podemos ver debido a que el ***Valor P* es igual a $1.88 \cdot 10^{-31}$**, es decir, **menor que 0.05**, se rechaza la hipótesis nula y se acepta la hipótesis alternativa y se confirma que la media de las dos muestras es estadísticamente distinta.

Otro patrón interesante que podemos observar en la *Figura 21* es sobre las medias de las asignaturas del cuarto curso ya que, entre los dos años en los que se han llegado a cursar estas asignaturas, la diferencia es muy baja. Además, las medias de las asignaturas son muy parecidas.

En cuanto a las asignaturas del grado de Matemáticas, no se ha observado ningún patrón importante sobre la evolución de las notas para cada curso. Podemos ver en la *Figura 22*, *Figura 23*, *Figura 24* y *Figura 25*, las medias de las asignaturas en función del curso y del año.

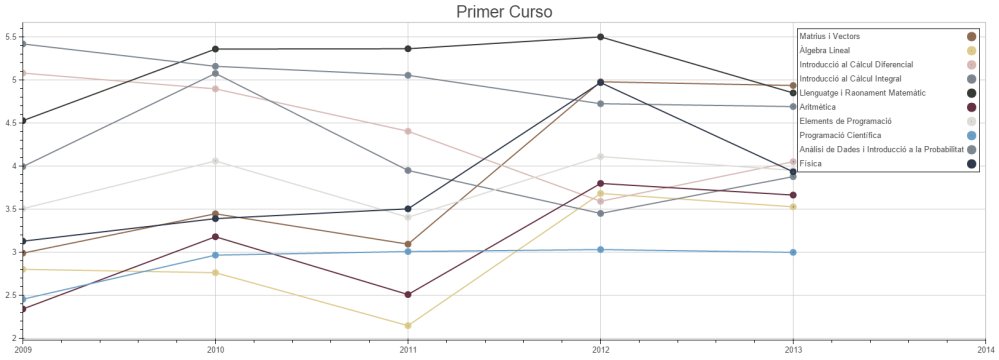

Figura 22: Evolución de las medias para las asignaturas del primer curso de Matemáticas

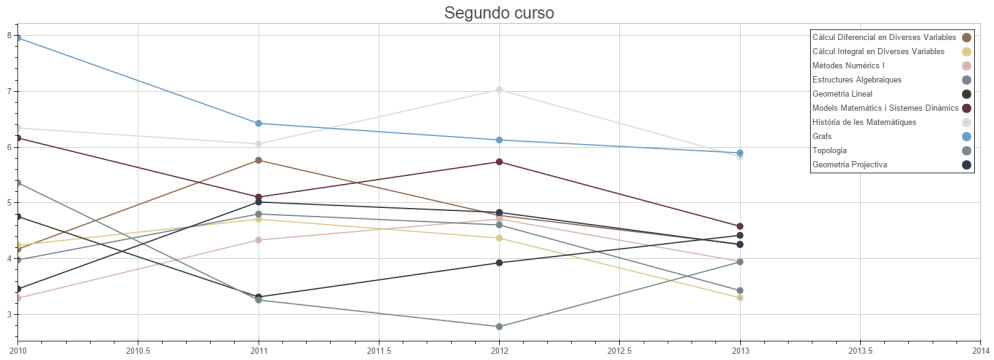

Figura 23: Evolución de las medias para las asignaturas del segundo curso de Matemáticas

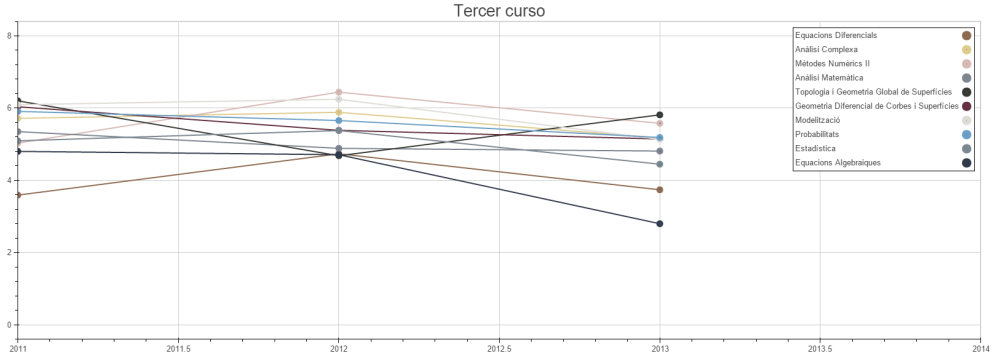

Figura 24: Evolución de las medias para las asignaturas del tercer curso de Matemáticas

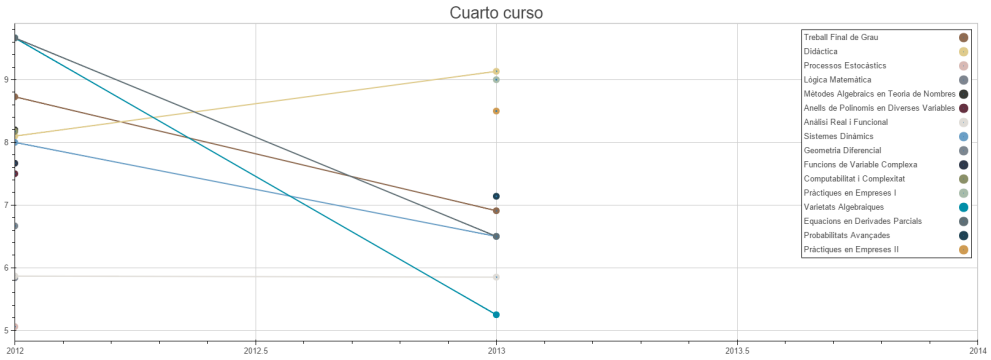

Figura 25: Evolución de las medias para las asignaturas del cuarto curso de Matemáticas

6.8. Ranking en función de la dificultad de las asignaturas

Queremos obtener el ranking de las asignaturas en función de la dificultad, es decir, en función de las notas de los alumnos. Esperaremos ver si las asignaturas se agruparan de algún modo en el ranking y así ver cuáles son las asignaturas en las que los alumnos obtienen peores o mejores notas.

A continuación, observamos la *Figura 26* que representa el ranking de las asignaturas para las calificaciones del **año 2012**. Como se puede ver, todas las asignaturas del cuarto curso se agrupan en la parte de arriba del ranking, es decir, los alumnos obtienen mejores notas en las asignaturas del cuarto curso en comparación con el resto de los cursos. También podemos observar que las asignaturas del primer curso suelen aparecer en la parte baja del ranking. Aunque se tiene que tener en cuenta la desviación estándar ya que, como se puede apreciar en el gráfico, está marcada como una línea roja por encima y debajo de los puntos. Resulta ser mucho más alta para las asignaturas de primero, siendo las notas muy dispersas. Una de las causas de la desviación estándar elevada es la alta tasa de abandono de los estudios.

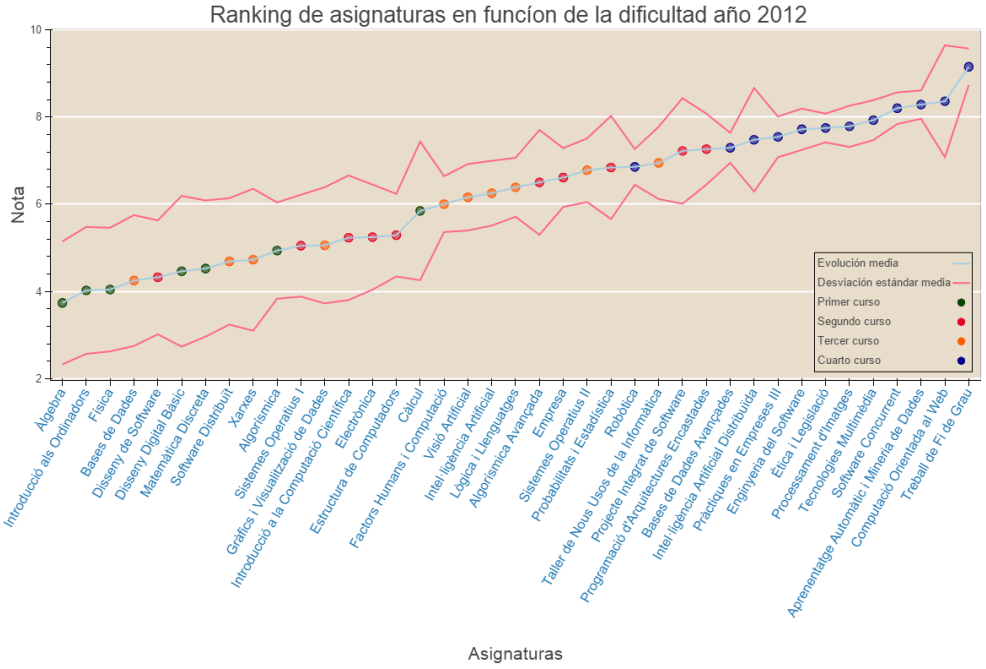

Figura 26: Ranking de las asignaturas de Ingeniería Informática en función de la dificultad para el año 2012. Se muestra la nota media de cada asignatura. El color indica su curso y la desviación estándar, marcada por las líneas rojas por encima y debajo de los puntos.

Observando la gráfica del ranking de las asignaturas para el *año 2013 Figura 27*, encontramos la misma tendencia.

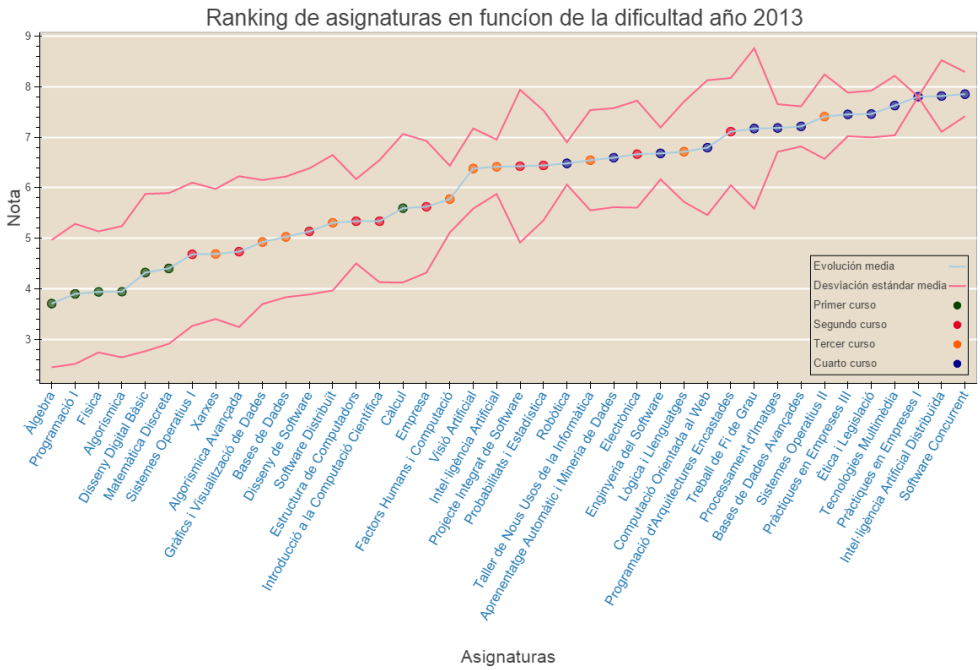

Figura 27: Ranking de las asignaturas de Ingeniería Informática en función de la dificultad para el año 2013. Se muestra la nota media de cada asignatura. El color indica su curso y la desviación estándar, marcada por las líneas rojas por encima y debajo de los puntos.

6.9. La anonimización de los datos

Como se ha explicado en el *Capitulo 2*, la anonimización de los datos [27] [28] [29] es importante antes de hacer público cualquier conjunto de datos. Para que un conjunto de datos sea realmente anónimo, su identificación y su relación con un sujeto ha de ser completamente irreversible. Es decir, podemos decir que un dato es anónimo si la posibilidad de vinculación con la persona a la que hubiera identificado el dato es nula, teniendo en cuenta que se pueden utilizar otros datos y recursos y que estos pueden aumentar el riesgo de identificación con el tiempo.

Así mismo las técnicas de anonimización deben ser las adecuadas teniendo en cuenta tanto el entorno o el contexto de los datos como el objetivo y la utilización. El error más común que se comete a la hora de anonimizar los datos es creer que remplazar un atributo de un registro por otro diferente hará imposible su recuperación o la identificación de la persona a la que identifica ese atributo.

Para asegurarnos de que la anonimización se hace correctamente, es importante elegir bien las técnicas de anonimización para reducir los tres principales riesgos que son inherentes a cualquier proceso de anonimización.

- El primer riesgo llamado *Singling out* que representa la posibilidad de aislar datos que permiten identificar a un sujeto dentro de un conjunto.
- El segundo riesgo llamado *Linkability* que representa la capacidad de relacionar, por lo menos, dos datos que hacen referencia a la misma persona o grupo de personas (indiferentemente de si los datos se encuentran en la misma base de datos o no)
- El tercer riesgo llamado *Inference* que representa la posibilidad de deducir el valor de un atributo a partir de otros conjuntos de atributos.

Conociendo los riesgos existentes, el contexto de los datos y sus usos, se podrá utilizar una o varias técnicas de anonimización para lograr el propósito deseado sin poner en peligro la privacidad de los sujetos.

Entre estas técnicas destacamos los siguientes dos grupos:

1. Técnicas que alteran la veracidad de los datos para eliminar el vínculo entre los datos y el sujeto. Si los datos son lo suficientemente inciertos, entonces ya no se pueden referenciar a una persona en específico.

Dentro de este grupo destacamos las técnicas de **adicción de ruido** que modifican o alteran los datos dentro de un conjunto de datos de tal manera que los datos serán menos precisos pero que conservarán su distribución general. Otra técnica destacada es la **permutación** de los datos que mezclan algunos valores de un mismo atributo para que estén vinculados a diferentes individuos. Esta técnica resulta especialmente útil cuando se quiere mantener la distribución exacta dentro de un tipo de datos.

2. Técnicas de generalización de datos que consisten en impedir que un cierto sujeto pueda ser identificado dentro de un grupo de datos, agrupándolo con otros sujetos. Dentro de este campo se conocen las técnicas de **agregación** y **k-anonymity**.

Teniendo en cuenta todo esto y dado el riesgo residual de la anonimización que siempre va a existir, se tendrá que tener en cuenta la identificación, supervisión y control de los riesgos tantos actuales como nuevos, y evaluar constantemente si los controles que existen sobre la anonimización son suficientes.

Ahora bien, vamos a verificar la anonimización que se ha hecho sobre los datos de los alumnos que estamos utilizando para llevar a cabo la investigación. Actualmente la anonimización hecha por el departamento de gestión académica que nos proporcionó los datos consiste en cambiar el NIUB de los alumnos por un identificador único que no tiene ninguna relación alfanumérica con el NIUB. Es decir, sabiendo solo los identificadores de los alumnos es imposible saber de quién se trata realmente ya que no se sabe la manera ni el orden de asignar los nuevos identificadores.

Debido a que prácticamente no se ha aplicado ninguna técnica de anonimización (a parte del remplazo del niub por un identificador aleatorio), si se dispone de varios datos sobre un sujeto en particular se podría llegar a identificarlo. Se ha de especificar también que a la hora de anonimizar los datos se ha tenido en cuenta el hecho de que los datos no serán difundidos públicamente y que solo tendrán acceso los desarrolladores del proyecto.

Por ejemplo, se exponen los siguientes casos hipotéticos que podrían llevar a la identificación de un sujeto:

- Conociendo el año de nacimiento, la comarca donde realizó sus estudios secundarios, el sexo, si ha hecho un ciclo formativo o no, o la nota que obtuvo en las pruebas de los PAU, cualquiera de las combinaciones hechas a partir de estos datos podrían llevarnos a identificar a un sujeto en particular

- Conociendo algunas de las notas que ha sacado, se podría llegar a identificar el sujeto.
- Conociendo el año en el que empezó los estudios más la nota de la pruebas PAU se podría también llegar a identificar el sujeto.

Como los 3 ejemplos mostrados con anterioridad se podrían crear varias combinaciones con los datos que uno podría disponer y poder así identificar a una persona. A continuación, ejemplificaré algunos casos en concreto.

Para poder continuar con los experimentos se ha de especificar que el número total de alumnos dentro de la base de datos es 961.

Ejemplo 1

Supongamos que disponemos de la siguiente información acerca de una persona:

- Año de nacimiento: **1986**
- Sexo: **hombre**
- Población del centro de los estudios de secundaria: **Castelldefels (comarca Baix Llobregat)**

A continuación se mostrara que con saber esta información podríamos identificar a esta persona dentro de la base de datos y acceder a la información relacionada.

1. Primero filtramos los registros por el año de nacimiento de las personas: 1986 de modo que se eliminarán todas aquellas personas que no hayan nacido el año 1986. Aplicando esta acción, el número de registros baja de **961** a **22**.
2. A continuación filtraremos los registros según el sexo de la personas de tal manera que nos quedemos solo con los varones. Aplicando esta acción, el número de registros baja de **22** a **21**.
3. Por ultimo filtraremos los registros según la comarca donde han cursado la secundaria, en este caso eliminaremos todas aquellas personas que no hayan estudiado en la comarca de *Baix Llobregat*. Aplicando esta acción, el número de registros baja de **21** a **1**.

Como se puede observar, con estos tres datos se ha podido identificar a una única persona dentro de la base de datos. Así mismo se tendría que aplicar un proceso de anonimización adicional para bajar los riesgos de

identificación. Se plantean las siguientes soluciones para dos de los campos implicados en el experimento (el año de nacimiento y la comarca del centro de los estudios de secundaria):

- Para el año de nacimiento se ha pensado utilizar la técnica de agrupación de los datos. Es decir, agrupar los años de nacimiento en función de la edad del alumno. De este modo tendríamos:

  1. Alumnos con una edad **inferior a 25 años**
  2. Alumnos con una edad **entre 25 y 35 años**
  3. Alumnos con una edad **superior a 35 años**

- Para la comarca del centro de estudios de secundaria se ha pensado utilizar la técnica de agrupación de los datos de tal manera que las comarcas se agrupen en grupos de 3 o 4 en función de la proximidad. Otra solución posible sería subir el nivel del área administrativa a nivel de provincias (en vez de disponer de la comarca: *Baix Llobregat*, tener la provincia de los estudios de secundaria: *Provincia de Barcelona*).

Ejemplo 2

Supongamos que disponemos de la siguiente información acerca de una persona:

- Nota de las pruebas PAU (Nota de acceso a la universidad): **6.915**

A continuación se mostrara que con saber esta información podríamos identificar a esta persona dentro de la base de datos y acceder a la información relacionada.

- Filtramos los registros por la nota de acceso a la universidad, de tal manera que se eliminen todos aquellos registros que tengan el campo de la nota de acceso diferente a 6.915. Aplicando esta acción, el número de registros baja de **961** a **1**.

Como se puede observar, con un único dato se ha podido identificar a una única persona dentro de la base de datos. Por lo tanto se tendría que aplicar un proceso de anonimización adicional para bajar los riesgos de identificación. Se plantea la siguiente solución para el campo que representa la nota de acceso a la universidad:

- Para eliminar el riesgo de identificación de un sujeto utilizándonos de la nota de acceso a la universidad, una posible solución sería quedarnos con solo un decimal. En este caso, la nota pasaría de **6.915** a **6.9**

Ejemplo 3

Supongamos que disponemos de la siguiente información acerca de una persona:

- Notas de las asignaturas cursadas: **5.8**, **8.9**, **5.5**, **6.2**, **7.8**

A continuación se mostrara que con saber esta información podríamos identificar a esta persona dentro de la base de datos y acceder a la información relacionada.

- Filtramos las cualificaciones según la lista de notas más arriba indicada, de tal manera que se eliminan todos aquellos alumnos que no tengan todas las notas de la lista: **5.8**, **8.9**, **5.5**, **6.2**, **7.8**. Aplicando esta acción, el número de registros baja de **961** a **1**.

Como se puede observar, conociendo algunas notas de un alumno, se ha podido identificar a una única persona dentro de la base de datos. Por lo tanto se tendría que aplicar un proceso de anonimización adicional para bajar los riesgos de identificación. Se plantean las siguientes soluciones para el campo que representa la cualificación para eliminar el riesgo de identificación de un sujeto utilizándose de las notas obtenidas:

- Una posible solución sería redondear la nota hacia la nota más cercana. Por ejemplo, **8.9 se quedaría en 9** y **6.2 se quedaría en 6**.

- Añadir un pequeño ruido a las notas sin cambiar la distribución de los datos. Por ejemplo, **8.9 se quedaría en 8.92** y **6.2 se quedaría en 6.17**.

Ejemplo 4

Supongamos que disponemos de la siguiente información acerca de una persona:

- Nota de la asignatura cursada de **Álgebra (364291): 8.5**

- Nota de la asignatura cursada de: **Electrónica (364305): 7.7**

A continuación se mostrará que con saber esta información podríamos identificar a esta persona dentro de la base de datos y acceder a la información relacionada.

- Filtramos las cualificaciones según la lista de notas más arriba indicada, de tal manera que se eliminan todos aquellos alumnos que no tengan un 8.5 en Álgebra y un 7.7 en Electrónica. Aplicando esta acción, el número de registros baja de **961** a **1**.

Como se puede observar, conociendo dos notas de un alumno, se ha podido identificar a una única persona dentro de la base de datos. Por lo tanto se tendría que aplicar un proceso de anonimización adicional para bajar los riesgos de identificación. Se plantean las siguientes soluciones para el campo que representa la cualificación para eliminar el riesgo de identificación de un sujeto utilizándose de las notas obtenidas:

- Una posible solución sería redondear la nota hacia la nota más cercana. Por ejemplo, **8.5 se quedaría en 9** y **7.7 se quedaría en 8**.
- Añadir un pequeño ruido a las notas sin cambiar la distribución de los datos. Por ejemplo, **8.5 se quedaría en 8.51** y **7.7 se quedaría en 7.68**.

¿Pero cuántos alumnos se podrían identificar realmente conociendo algunos de los datos?

Nos planteamos esta pregunta pensando en cuántos alumnos se podrían identificar conociendo uno o más datos sobre el mismo. Una base de datos se considera vulnerable si a partir de uno o más datos se podría llegar a textbfidentificar a menos de 5 sujetos, en este caso, alumnos.

Nos centraremos en la tabla que representa el registro de los alumnos ya que contiene información que muchas personas podría conocer sobre un sujeto, tal como: año de nacimiento, grado estudiado, año del comienzo de los estudios, etc.

A continuación realizaremos un experimento para verificar cuantas personas se pueden aislar a partir de un único dato. Cogeremos todos los valores únicos de cada descriptor de los registros (año nacimiento, nota de acceso a la universidad, grado estudiado, etc.) y filtraremos los registros para cada uno de estos valores únicos.

Por ejemplo, supongamos que tenemos los siguientes valores únicos para el campo de la nacionalidad:

1. España
2. Perú
3. Rumanía
4. Italia

Así mismo, se filtraran los registros para cada uno de estos valores únicos para observar cuantos alumnos se pueden aislar.

Para verificar cuantos alumnos se pueden aislar a partir de los datos, se ha creado un código en *python* que, de manera automática, verificará para cada valor de los descriptores del registro, cuántos alumnos se pueden aislar. Recordemos que la base de datos se considera vulnerable si se puede llegar a aislar a menos de 5 alumnos. Más abajo podemos ver en el (*Cuadro 9*) los porcentajes de aislamientos para cada descriptor.

Observación: Cuando nos referimos al ***aislamiento*** o al ***porcentaje de aislamiento*** queremos decir lo siguiente:

- Supongamos que el campo ***año de nacimiento*** dispone de ***20 valores únicos***.
- Si con cada valor único de este campo, llegamos a identificar a ***menos de 5 alumnos***, decimos que con el valor único *hemos aislado* a los sujetos.
- Por lo tanto, si de los 20 valores únicos, con 15 podemos ***identificar a menos de 5 alumnos***, decimos que ***el porcentaje de aislamiento*** es de 75 %.

Descriptor	**Porcentaje de aislamiento**
Lugar CFGS	100 %
Nota de acceso	99.34 %
País sistema extranjero	87.50 %
Lugar estudios de secundaria	81.48 %
Nacionalidad	66.66 %
Universidad de procedencia	60.71 %
Año de la pruebas PAU	52 %
Año de nacimiento	51.51 %
Tipo lugar estudios de secundaria	50 %

Cuadro 9: Porcentajes de aislamiento

Como podremos observar, a partir de la nota de acceso se podrían identificar el 99.34 % de los grupos de menos de 5 alumnos y aunque éste descriptor tenga muchos valores, al tener 3 decimales, las notas son en la mayoría de los casos diferentes entre los distintos alumnos. Por éste motivo, se tendría que aplicar un proceso de anonimización, para que las notas de acceso tengan un solo decimal.

En el siguiente gráfico (*Figura 28*), podemos observar el porcentaje total de aislamiento para cada descriptor del número total de aislamientos posibles.

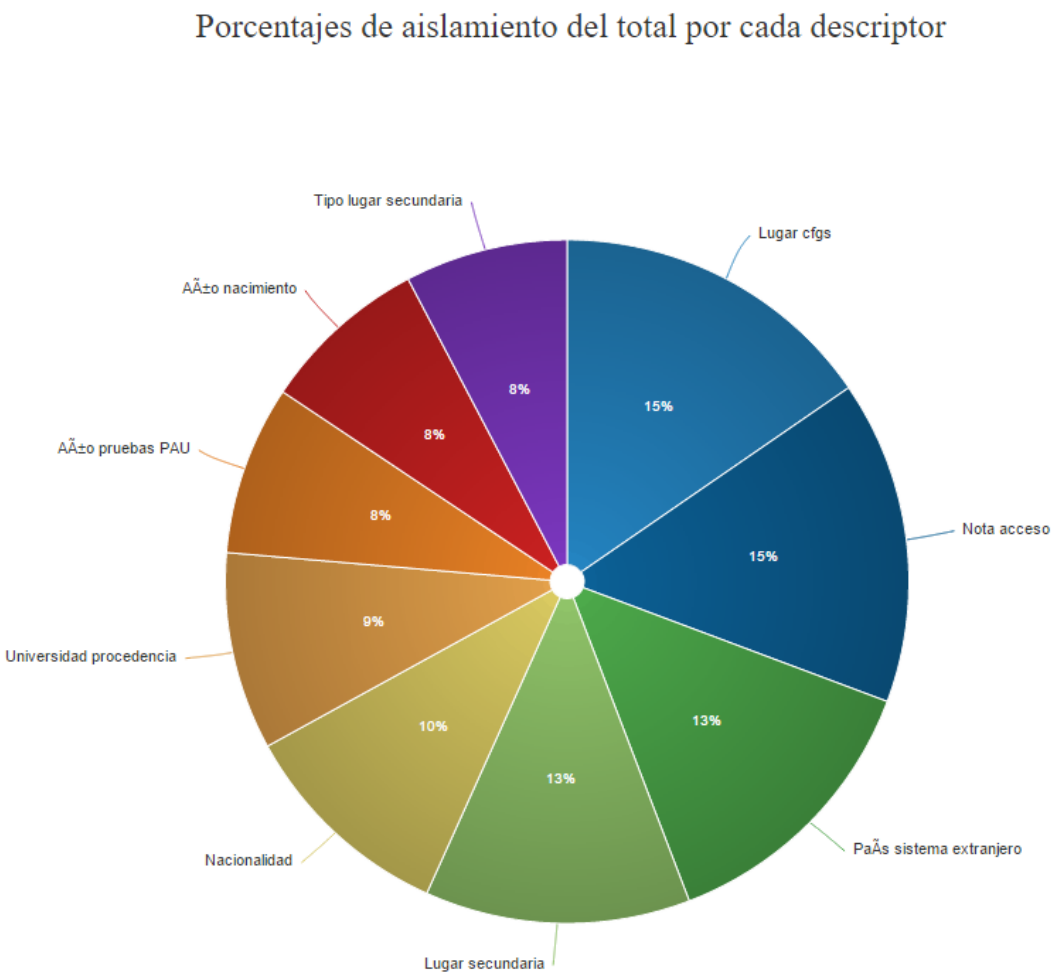

Figura 28: La distribución de los porcentajes de aislamiento del total por cada descriptor

A continuación realizaremos el mismo experimento pero esta vez verificaremos cuantas personas se pueden aislar a partir de dos datos emparejados. Cogeremos todos los valores únicos de cada descriptor de los registros (año nacimiento, nota de acceso a la universidad, grado estudiado, etc.), crearemos todas las combinaciones posibles de dos entre estos valores y filtraremos los registros para cada una de estas parejas. Por ejemplo, supongamos que tenemos los siguientes valores únicos para el campo de la nacionalidad:

1. España

2. Perú

3. Rumanía

4. Italia

Y los siguientes valores únicos para el campo del sexo:

1. Masculino

2. Femenino

A partir de estos valores se crearán 8 parejas distintas: España - Masculino, Perú - Masculino, Rumanía - Masculino, Italia - Masculino, España Femenino, Perú - Femenino, Rumanía - Femenino, Italia - Femenino.

Para verificar cuantos alumnos se pueden aislar a partir de los datos, se ha creado un código en *python* que, de manera automática, verificará por cada pareja de valores de los descriptores del registro cuántos alumnos se pueden aislar. Recordemos que la base de datos se considera vulnerable si se puede llegar a aislar a menos de 5 alumnos. Más abajo en el (*Cuadro 10*) podremos ver una tabla con el porcentaje de aislamientos por cada pareja de descriptores.

Descriptor	**Porcentaje de aislamiento**
Lugar CFGS y Tipo lugar CFGS	100 %
Escuela pública/privada CFGS y Lugar CFGS	75 %
ID grado y Lugar CFGS	75 %
Lugar estudios de secundaria y Sexo	62.96 %
Lugar estudios de secundaria y Becado	62.96 %
Lugar CFGS y Sexo	62.5 %
Lugar CFGS y Simultaneidad de grados	62.5 %
País sistema extranjero y Sexo	62.5 %
Lugar estudios de secundaria y Becado	62.5 %
ID grado y Lugar País sistema extranjero	62.5 %
ID grado y Nacionalidad	60.42 %
.	.
.	.
.	.
Lugar estudios de secundaria y Nota acceso	0.46 %
País sistema extranjero y Nota acceso	0.28 %
Lugar CFGS y Nota acceso	0.23 %

Cuadro 10: Porcentajes de aislamiento por pares

Anonimización de los datos mediante la *Agregación* y pruebas de identificación

Debido a que los datos de los que disponemos no se les han aplicado el proceso de anonimización establecido por las normativas vigentes de la Universidad de Barcelona, procedemos a explicar algunas de las técnicas de *Agregación* que podríamos aplicar sobre los descriptores de un alumno para hacer imposible la identificación conociendo algunos de los datos. Reiteramos que la anonimización no se ha hecho debido a que los datos serán de uso interno y no se harán públicas y en un futuro si se decidiera la publicación, los datos tendrían que pasar por un proceso riguroso de anonimización.

Ejemplo 1. Agregación de las notas de acceso a la universidad

Como pudimos ver en las pruebas de desanonimización, conociendo la nota de acceso a la universidad de un alumno, se podía llegar a identificarlo en la mayoría de los casos. En consecuencia, redondearemos las notas para hacer la identificación imposible a través de este campo.

- Redondeando los valores utilizando dos decimales se han podido identificar a menos de 5 alumnos un 788 de veces de las 966 posibles
- Redondeando los valores utilizando un decimal se han podido identificar a menos de 5 alumnos un 51 de veces de las 966 posibles
- Redondeando los valores sin utilizar decimales, no se podido identificar a ningún alumno. Así mismo, para hacer anónimo este campo se tienen que eliminar todos los decimales de la nota.

Ejemplo 2. Agregación de la edad

Como pudimos ver en las pruebas de desanonimización, conociendo la edad de un alumno, se podía llegar a identificarlo en algunos de los casos. En consecuencia, agregamos la edad de los alumnos para hacer imposible la identificación. Para conseguirlo agregamos la edad de modo que cada alumno estará en una de las siguientes 3 categorías:

- inferior a 25 años
- superior a 25 años e inferior a 35 años
- superior a 35 años

Aplicando esta acción, el número de alumnos que se puede identificar por cada categoría es el siguiente:

1. Número de alumnos con edad inferior a 25 años: 633

2. Número de alumnos con edad superior a 25 años e inferior a 35 años: 308

3. Número de alumnos con edad superior a 35 años: 25

Podemos ver que por cada categoría no se pueden identificar a menos de 5 alumnos que es el límite para considerar los datos anónimos. Aunque se han elegido estas tres categorías, se podría intentar agregar los datos utilizando otras categorías.

7. Conclusiones y trabajo futuro

7.1. Conclusiones

Éste trabajo final de grado fue una gran oportunidad para poder aplicar todos los conocimientos adquiridos a lo largo de los estudios universitarios tanto a nivel de programación, como a nivel estadístico y del área de la ciencia de los datos. Utilizando las técnicas de un proyecto típico de la ciencia de los datos, se ha podido extraer a partir de los datos una gran cantidad de información valiosa e interesante sobre las relaciones que pueden haber entre los alumnos y las asignaturas, los diferentes factores que pueden influir en la calidad docente de las asignaturas y los conocimientos que el alumno puede adquirir a lo largo de sus estudios.

Analizando los datos se ha podido observar y descubrir nuevas tendencias y problemas en cuanto a las notas de los alumnos y a la interacción con las asignaturas. Éste trabajo sirve para entender mejor cómo se comportan los alumnos en relación con las asignaturas y que se puede hacer una vez identificados los problemas y las causas. Creo que los resultados del trabajo podrán ayudar tanto al jefe de estudios como a los tutores de estudios a entender mejor los problemas de los alumnos y a tomar las decisiones adecuadas para hacer la docencia más efectiva y ayudar a los alumnos con problemas a superarlos.

7.2. Trabajo futuro

Debido a que este trabajo final de grado es parte del proyecto de innovación docente *Sistema inteligente de soporte al tutor de estudios* [1], el trabajo futuro consiste en la continuación del desarrollo del proyecto. Es decir, se continuará con la implementación del sistema añadiéndole nuevas funcionalidades. En concreto se seguirá con las siguientes dos fases:

- **Fase 4**: Desarrollo del sistema inteligente. Ésta fase representa la implementación del sistema inteligente que será una herramienta de soporte con gráficos y datos cuantitativos y cualitativos, tanto para el tutor de estudios como para el jefe de estudios. Además ayudará en la toma de decisiones sobre las acciones de mejora en las asignaturas y en los planos docente y tutorial.
- **Fase 5**: Evaluación del sistema. En ésta fase se analizarán los resultados del sistema, la eficacia del mismo, y la identificación de los posibles errores producidos.

8. Bibliografía

Referencias

[1] "Proyecto de Innovación Docente" [En linea].
Pagina Web: http://mid.ub.edu/webpmid/content/sistema-intel%E2%80%A2ligent-de-suport-al-tutor-d%E2%80%99estudis

[2] "Capstone Projects" - Data Science & Big Data Postgraduate Course - Universitat de Barcelona

[3] GitHub Guides [En linea].
Pagina Web: http://guides.github.com/

[4] Pagina web de Github [En linea].
Pagina Web: http://github.com

[5] Pagina web de Bitbucket [En linea].
Pagina web: http://bitbucket.org

[6] Pagina web de Trello [En linea].
Pagina web: http://trello.com

[7] Pagina web de Python [En linea].
Pagina web: http://www.python.org

[8] Pagina web de Pandas [En linea].
Pagina web: http://pandas.pydata.org

[9] Pagina web de Numpy [En linea].
Pagina web: http://www.numpy.org

[10] Pagina web Bokeh [En linea].
Pagina web: http://bokeh.pydata.org

[11] Pagina web Seaborn [En linea].
Pagina web: http://stanford.edu/~mwaskom/software/seaborn

[12] Pagina web de Ipython Notebook [En linea].
Pagina web: http://ipython.org/notebook.html

[13] Pagina web de TeXnicCenter [En linea].
Pagina web: http://http://www.texniccenter.org/

[14] Pagina web de Latex (Miktex - Dostribución para Windows) [En linea].
Pagina web: http://miktex.org/

[15] Pagina web de D3 [En linea].
Pagina web: http://d3js.org/

[16] "Probabilitat i Estadística" - Mireia Besalú, Carles Rovira

[17] Pagina web de Social Research Methods [En linea].
Pagina web: http://www.socialresearchmethods.net/kb/stat_t.php

[18] Pagina web de Wikipedia - T Test [En linea].
Pagina web: http://en.wikipedia.org/wiki/Student%27s_t-test

[19] Pagina web de Wikipedia - Regresión lineal [En linea].
Pagina web: http://en.wikipedia.org/wiki/Linear_regression

[20] Pagina web de DataRobot - Ordinary Least Squares in Python [En linea].
Pagina web: http://www.datarobot.com/blog/ordinary-least-squares-in-python/

[21] Pagina Web Universo Formulas - Diagrama de Barras [En linea].
Pagina web: http://www.universoformulas.com/estadistica/descriptiva/diagrama-barras/

[22] Pagina Web Monografias.com - Prueba F de Fisher [En linea].
Pagina web: http://www.monografias.com/trabajos91/prueba-hipotesis-f-fisher-empleando-excel-y-winstats/prueba-hipotesis-f-fisher-empleando-excel-y-winstats.shtml

[23] Pagina web de Wikipedia - Diagrama de caja [En linea].
Pagina web: https://es.wikipedia.org/wiki/Diagrama_de_caja

[24] Pagina Web Monografias.com - Correlación de Spearman [En linea].
Pagina web: http://www.monografias.com/trabajos85/coeficiente-correlacion-rangos-spearman/coeficiente-correlacion-rangos-spearman.shtml

[25] Pagina web de Wikipedia - Desviación estándar [En linea].

Pagina web: https://es.wikipedia.org/wiki/Desviaci%C3%B3n_t%C3%ADpica

[26] Pagina Web Monografias.com - Correlación de Spearman [En linea].

Pagina web: http://www.monografias.com/trabajos85/coeficiente-correlacion-rangos-spearman/coeficiente-correlacion-rangos-spearman.shtml

[27] Pagina Web Noticias Juridicas - Anonimización de los datos [En linea].

Pagina web: http://noticias.juridicas.com/conocimiento/articulos-doctrinales/4922-iquest;existe-de-verdad-la-anonimizacion-el-grupo-del-articulo-29-de-proteccion-de

[28] Pagina Web Privacidad Logica - Tecnicas de anonimizar [En linea].

Pagina web: http://www.privacidadlogica.es/2014/04/22/tecnicas-de-anonimizar-datos-personales-segun-las-autoridades-europeas-de-p

[29] Pagina Web Comisión Europea - Dictamen sobre tecnicas de anonimización [En linea].

Pagina web: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_es.pdf