UNIVERSIDAD DE GRANADA

TRABAJO FIN DE GRADO

DOBLE GRADO EN INGENIERÍA INFORMÁTICA Y MATEMÁTICAS

ASISTENTE PARA EL DESCUBRIMIENTO DE PROCESOS DE APRENDIZAJE OCULTOS DURANTE LA REALIZACIÓN DE PRÁCTICAS DE LABORATORIO

Autora

MARÍA ISABEL RUIZ MARTÍNEZ

Director

LUIS CASTILLO VIDAL

FACULTAD DE CIENCIAS

E.T.S. DE INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

Granada, a 9 de julio de 2023

Autorización

Yo, **María Isabel Ruiz Martínez**, alumna de la titulación Doble Grado en Ingeniería Informática y Matemáticas de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 75576979Z, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

En Granada a 9 de julio de 2023

Fdo: María Isabel Ruiz Martínez

Informe

D. **Luis Castillo Vidal**, Catedrático de Universidad del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

Informa:

Que el presente trabajo, titulado ***Asistente para el descubrimiento de procesos de aprendizaje ocultos durante la realización de prácticas de laboratorio***, ha sido realizado bajo su supervisión por **María Isabel Ruiz Martínez**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

En Granada a 9 de julio de 2023

Fdo: Luis Castillo Vidal

AGRADECIMIENTOS

Aquí van mis agradecimientos.

RESUMEN

Aquí va mi resumen.

PALABRAS CLAVE: analítica de aprendizaje, minería de procesos, teoría de grafos, grafo dirigido acíclico

SUMMARY

Aquí va mi resumen.

KEYWORDS: learning analytics, process mining, graph theory, directed acyclic graph

ÍNDICE GENERAL

Parte I

MOTIVACIONES

Introducción, Motivación, Objetivos y Estructura.

MOTIVACIÓN

0.1 INTRODUCCIÓN

La necesidad de comprender el proceso de aprendizaje y de personalizar la enseñanza para realizar una mejor adaptación a las necesidades del individuo ha motivado la *Analítica de Aprendizaje* o *Learning Analytics*, disciplina que consiste en la recogida de datos de un entorno de aprendizaje y el análisis de los mismos cuyo objetivo es asistir en el proceso de aprendizaje del alumnado.

Además, el uso de laboratorios virtuales y remotos en la enseñanza está en auge. Entre muchas de sus ventajas tenemos una mayor privacidad para el alumnado, accesos planficados a los mismos o soporte para reportar la actividad de los alumnos y la calificación de los mismos.

En este trabajo fin de grado se usarán datos de siete cursos académicos obtenidos en el laboratorio virtual para sistemas multiagente de la asignatura del cuarto curso académico Desarrollo Basado en Agentes del grado de Ingeniería Informática de la Universidad de Granada (España).

El laboratorio virtual diseñado para la asignatura recoge el trabajo diario de los alumnos almacenando las interacción entre los diferentes agentes y obteniendo así un extenso dataset que nos proporciona una base sólida para el uso de diversas analíticas de aprendizaje.

Así pues, se empleará un enfoque *"data-driven"* o *impulsado por datos*, tomando decisiones estratégicas basándose en el análisis de los datos y en la interpretación de los mismos.

0.2 MOTIVACIÓN

La principal motivación de este trabajo fin de grado es, precisamente, el análisis de los procesos de aprendizaje que siguen los alumnos para que el profesorado pueda asistirles mejor durante su proceso de aprendizaje y mejorar así su rendimiento académico.

0.3 OBJETIVOS

Los objetivos principales del proyecto serán:

- Identificar patrones de comportamiento indicativos de la evolución de los alumnos y del progreso de su aprendizaje, detectando, en las fases más tempranas posibles, comportamientos que pudiesen ser anómalos o que pudiesen indicar problemas de aprendizaje. Es decir, se pretende relevar, mediante la utilización de técnicas de minería de procesos, las posibles estrategias de los alumnos para cumplir los distintos objetivos de la asignatura así como desvelar su forma de trabajo habitual.
- Sugerir a estos alumnos las medidas necesarias para que recuperen un buen ritmo de aprendizaje y un progreso adecuado a lo que el profesorado de la asignatura espera.

0.4 ESTRUCTURA DEL TRABAJO FIN DE GRADO

Este trabajo fin de grado consta de cuatro partes, siete capítulos y otros elementos como la portada, la autorización para su ubicación en la biblioteca de la escuela, sendos resúmenes tanto en español como en inglés (con sus respectivas palabras clave), la sección de agradecimientos, los índices general, de figuras y de cuadros así como una bibliografía, un glosario de términos y un glosario de acrónimos.

A continuación se expone un breve esquema general del contenido de las partes y capítulos de este trabajo fin de grado:

- Parte I: Motivaciones.
  -
- Parte II: Estado del arte.
  - Capítulo 1:
  - Capítulo 2:
  - Capítulo 3:
  - Capítulo 4:
- Parte III: Análisis descriptivo.
  - Capítulo 5:
  - Capítulo 6:
  - Capítulo 7:
  - Capítulo 8:
- Parte IV: Planificación del proyecto.
  - Capítulo 9:
  - Capítulo 10:

- Parte V: Resultados obtenidos.
  - Capítulo 11:
  - Capítulo 12:
  - Capítulo 13:

Parte II

ESTADO DEL ARTE

Planteamiento del problema y Minería de Procesos.

1

PLANTEAMIENTO DEL PROBLEMA

En la asignatura Desarrollo Basado en Agentes los alumnos, organizados en grupos de 4 o 5 alumnos, se conectan a un Laboratorio remoto de la UGR que está siempre disponible para los mismos. La arquitectura del servidor remoto puede apreciarse en la Figura 1.

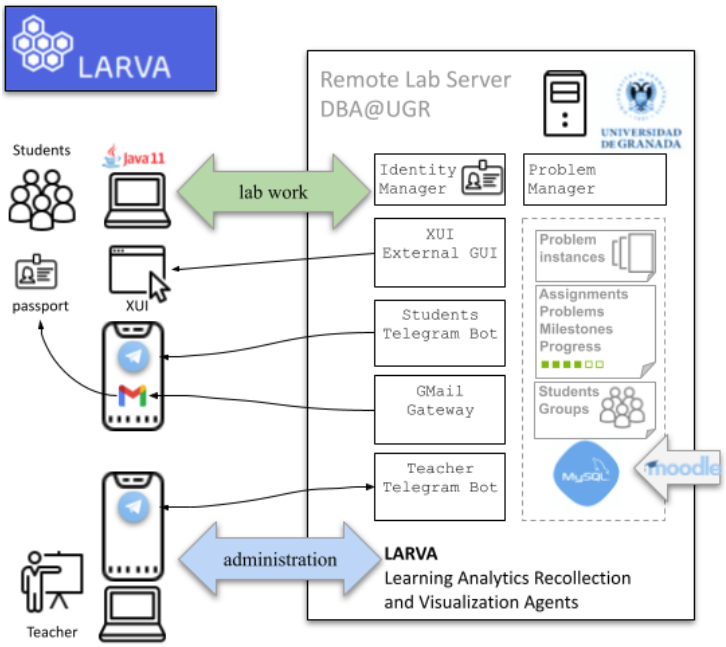


Figura 1: Arquitectura del Servidor Remoto. Por un lado, contiene el laboratorio virtual para sistemas multiagente distribuidos. Además, los alumnos también pueden consultar su progreso y el de sus compañeros a través de un Bot de Telegram. Por otro lado, el profesor también puede conocer el número de objetivos conseguidos por cada uno de sus grupos de alumnos.

Este servidor contiene varios mundos virtuales y se encarga de registrar y almacenar las interacciones con él Vidal (2016). Cada mundo virtual es una matriz cuadrada que representa espacios abiertos (en color blanco), obstáculos (en negro) y objetivos (en rojo) tal y como se muestra en la Figura 2. Los agentes de los alumnos deben entrar en uno de esos mundos

virtuales, percibir su vecindario, navegar a través de los espacios abiertos (empleando alguna clase de heurística exploratoria), evitar obstáculos y tratar de llegar al objetivo. En total, cada uno de los problemas planteados requieren de cinco pasos (o *milestones*) hasta su consecución.

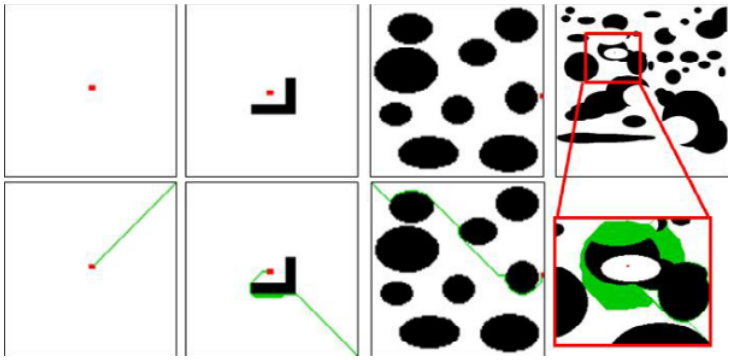
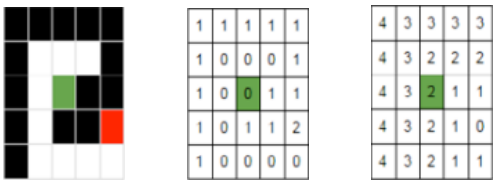


Figura 2: Alguno de los mapas que el alumnado debe resolver. Los agentes de los grupos de estudiantes deben acceder a uno de esos mundos y deben alcanzar los objetivos (coloreados en rojo) navengado a través del mundo y evitando los obstáculos (coloreados en negro). Alguno de los mundos no son resolubles porque el objetivo no se puede alcanzar con el objetivo de forzar a los agentes de los alumnos a razonar acerca de la irresolubilidad. Las posibles trayectorias están marcadas en verde.

La percepción del agente de su entorno es crítica para resolver estos mundos. En este laboratorio virtual los alumnos pueden configurar cuál de los siguientes sensores estarán enchufados en sus agentes (cualquier combinación de ellos):

- Un **GPS** que indica al agente sus coordenadas (x, y) en el mundo virtual.
- Un **sensor de batería**. Cada agente está alimentado con una batería cuya capacidad es limitada y cuya carga decrece conforme el agente realiza algún movimiento. La batería nunca debe ser vaciada por completo.
- Un **sensor radar** que informa al agente acerca de los tipos de celdas que lo rodean con una percepción local de 5x5 (observar Figura 3b).
- Un **sensor escáner** que actúa como *detector del objetivo* e indica al agente la distancia al objetivo medida desde cada una las celdas de su entorno 5x5 (observar Figura 3c).



(a) (b) (c)

Figura 3: Un agente (representado por una celda verde en el centro de cada figura) tiene una percepción local de su entorno: solamente percibe el entorno 5x5 de celdas colindantes. El Radar 3b muestra dicho entorno 5x5 que rodea al agente e informa de si una celda está vacía (valor 0), si hay un obstáculo (valor 1) o de si hay un objetivo (valor 2). El Escáner 3c muestra la distancia de cada una de las celdas colindantes al objetivo.

Basados en su percepción del mundo virtual, cada agente decidirá ejecutar alguna de las siguientes acciones en su entorno implementando cualquier heurística o proceso de búsqueda.

- LOGIN. Entrar en cualquiera de los mundos virtuales.
- MOVE. Mover al agente a una de las 8 celdas adyacentes y gastar una cierta cantidad de batería. Si la celda destino es un obstáculo o el agente se queda sin batería, el agente se rompe y se sale del mundo virtual.
- REFUEL. El agente recarga completamente su batería. A los agentes se les permite recargar su batería tantas veces como deseen.

2

MINERÍA DE PROCESOS

2.1 INTRODUCCIÓN A LA MINERÍA DE PROCESOS

Las transacciones de los agentes en estos mundos virtuales registradas en el servidor no sólo son importantes desde el punto de vista de la evaluación del alumnado, sino que también nos proporcionan información de cómo se han resuelto los problemas propuesto, hito por hito, y son un reflejo de la estrategia seguida por ccada uno de los equipos para intentar resolver todos los mundos.

Para tratar de desvelar estas estrategias ocultas se usarán técnicas de minería de procesos, considerando una estrategia como el proceso seguido por los estudiantes hasta llegar al objetivo. La minería de procesos puede definirse como la disciplina que tiene como objetivo descubrir, monitorear y mejorar procesos de negocio mediante el análisis de las transacciones del proceso que se han almacenado en algún sistema de información (Mayorga & García, 2015).

Actualmente, con el desarrollo y el creciente interés de las plataformas educativas y de toda la tecnología relacionada con las mismas, los sistemas de información nos permiten recoger todo tipo de información. Esto puede incluir desde información de bajo nivel (clicks del ratón) hasta información de alto nivel (realización de una actividad en particular dentro de la plataforma). Es decir, estos sistemas tienen la capacidad de almacenar datos temporales de diversa índole, como cadenas de clicks, registros de chats, históricos de modificación de documentos, registros de uso de los diferentes recursos educativos, etc. Bogarín et al. (2018). La minería de procesos puede usar todos estos logs para descubrir, monitorear y mejorar los procesos educativos. Surge así la denominada minería de procesos educacional (en inglés, *educational process mining*). No obstante, cabe destacar que, aunque en este trabajo fin de grado nos centraremos en la minería de procesos en el ámbito educativo, ésta también tiene númerosas aplicaciones en el área sanitaria, en el ámbito empresarial, institucional etc.

2.2 EXTRACCIÓN DE LOS PROCESOS CON DISCO

Para la extracción de los procesos ocultos se empleará el programa Disco. Disco es una herramienta de minería de procesos profesional que permite crear mapas visuales a partir de los registros en cuestión de minutos.

Para crear los diagramas de Disco, se extraerán los campos de información más importantes del dataset:

1. El identificador del caso, extraído de una clave aleatoria generada al principio de cada operación `LOGIN` y que distingue de manera unívoca cada sesión de trabajo de los estudiantes.
2. El agente, que se refiere al nombre del grupo de estudiantes.
3. La fecha y hora a la que se registró la transacción.
4. El campo actividad (`Activity`), que refleja la acción de los alumnos en el mundo virtual.
5. Varios campos de tipo recurso (`Resource`) que proporcionan información adicional que puede sernos de utilidad a la hora de filtrar los registros.

Así pues, se importarán el dataset de dos maneras diferentes, con el objetivo de estudiar tanto la frecuencia con que cada problema o mapa ha sido visitado como las acciones compuestas mapa-porcentaje superado. En la primera importación la `Activity` es el problema y el identificador del caso es el grupo. En la segunda importancia, por el contrario, la `Activity` es la composición del problema y el milestione y el identificador del caso son las sesiones. Las Figuras 4 y 5 muestran los diagramas obtenidos en cada uno de los casos.

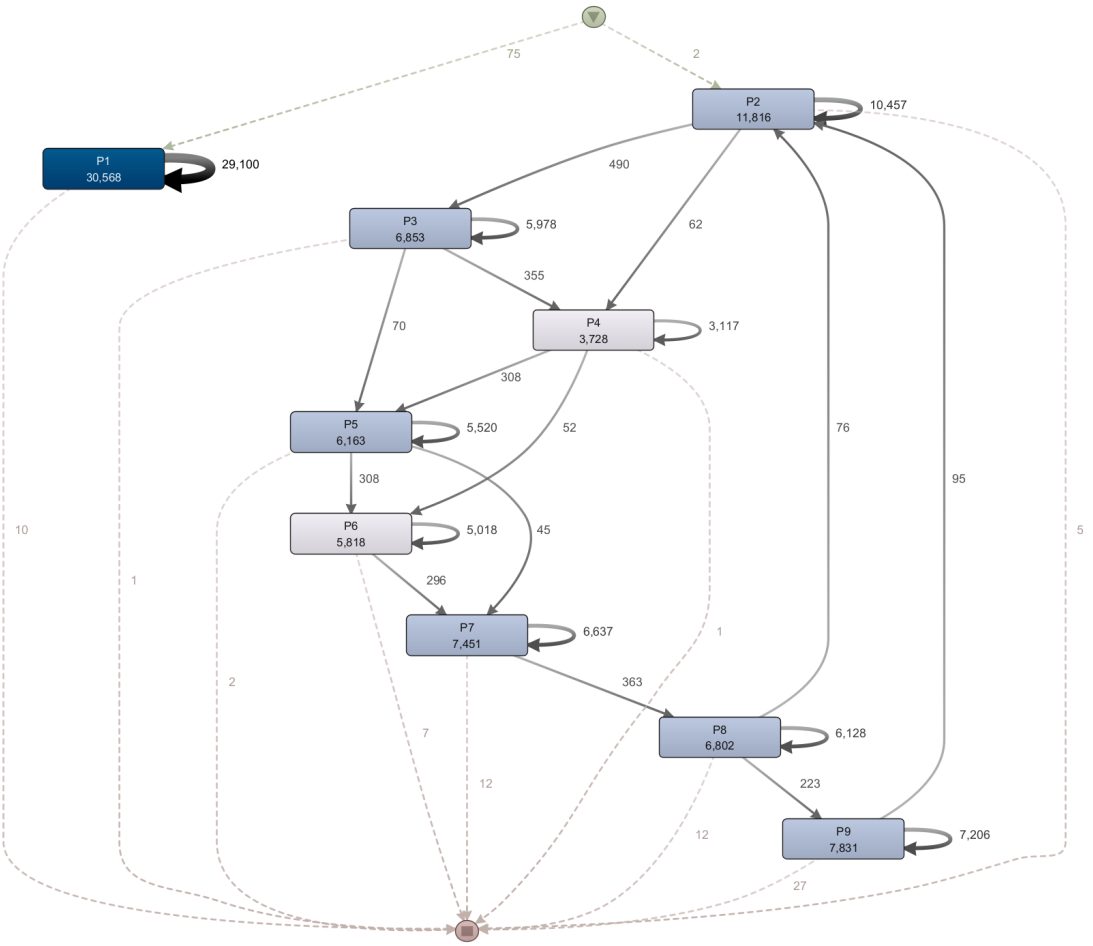


Figura 4: Análisis de procesos del dataset (`Activity` problema y `CaseId` grupo). Contiene el 100 % de las actividades y el 80 % de los caminos.

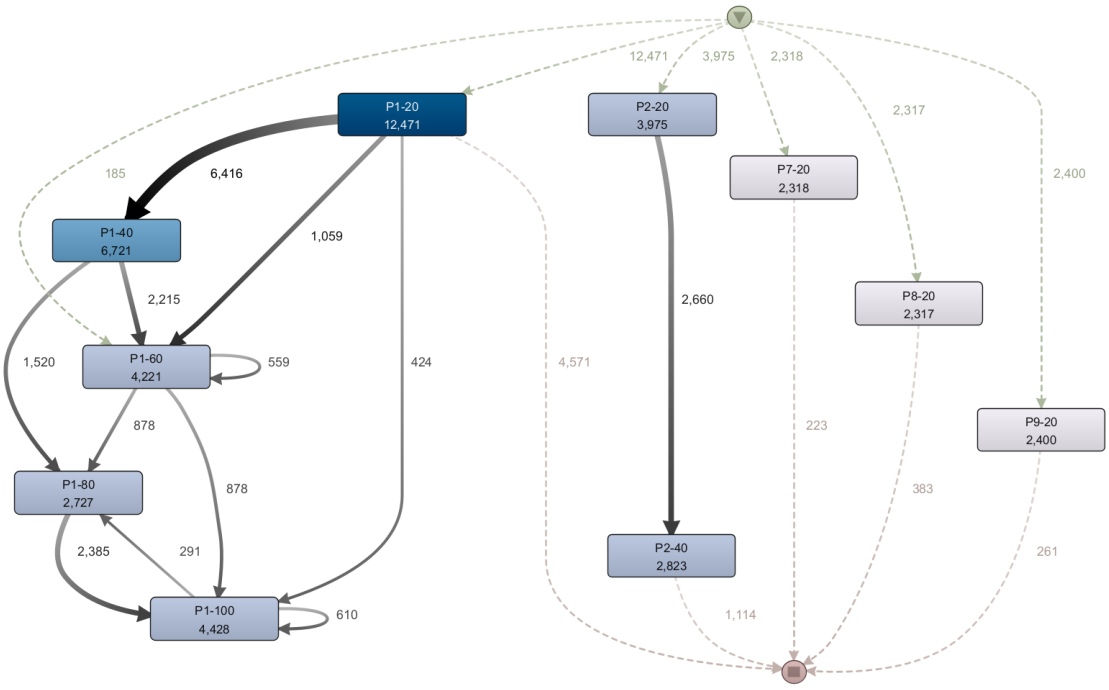


Figura 5: Análisis de procesos del dataset (`Activity` problema-milestone y `CaseId` sesión). Contiene el 20 % de las actividades y el 20 % de los caminos.

No obstante, a pesar de que tener una visión global del comportamiento de todos los grupos puede ayudar, nuestro objetivo final es poder caracterizar los comportamientos de los grupos y poder discernir, usando los datos del diagrama, si. Es por esto que nos será más interesante segmentar por grupos. Así pues, filtrando por el grupo `DBA 1516 P2 GA`, obtenemos los diagramas de las Figuras 6 y 7.

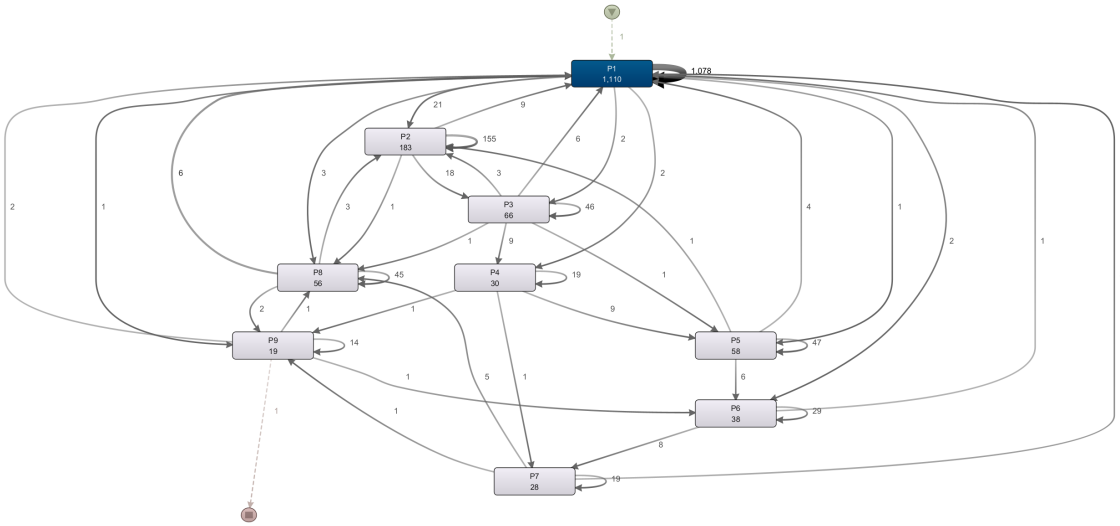


Figura 6: Análisis de procesos del grupo `DBA 1516 P2 GA` (`Activity` problema y `CaseId` grupo). Contiene el 100 % de las actividades y el 100 % de los caminos.

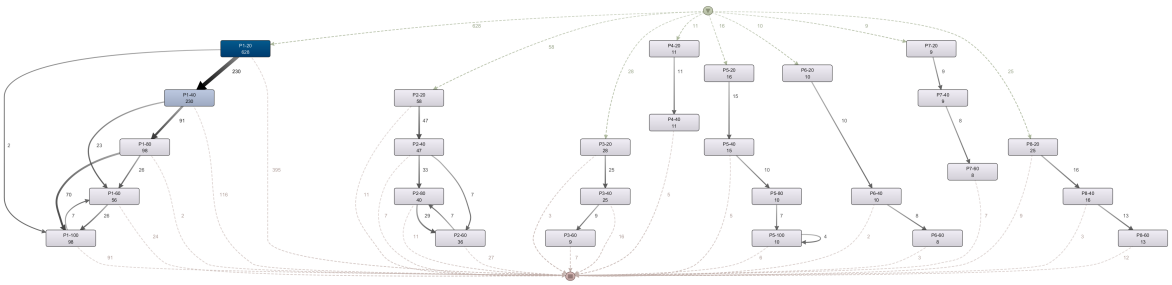


Figura 7: Análisis de procesos del grupo `DBA 1516 P2 GA` (`Activity` problema-milestone y `CaseId` sesión). Contiene el 60 % de las actividades y el 80 % de los caminos.

Sin embargo, tras una larga experimentación con el programa Disco, se empiezan a ver sus limitaciones. En primer lugar, aunque Disco permite el filtrado de datos, si se quiere segmentar por grupos y extraer los procesos ocultos de cada uno de los grupos, hay que seleccionar el correspondiente grupo en el filtro, extraer los diagramas correspondientes e ir cambiandolo manualmente. Dado que tenemos un total de 77 grupos de alumnos en los siete cursos académicos que forman parte del estudio (que pueden consultarse en las Tablas 17 y 18), es inviable seguir usando el programa.

Así pues, en este trabajo fin de grado se ha implementado nuestra propia versión del programa, personalizada y adaptada a las necesidades del problema.

3

IMPLEMENTACIÓN DE LA HERRAMIENTA DE MINERÍA DE PROCESOS

Comparando las Figuras 8 y 6 vemos que el diagrama de la implementación propia y el original obtenido con Disco coinciden.

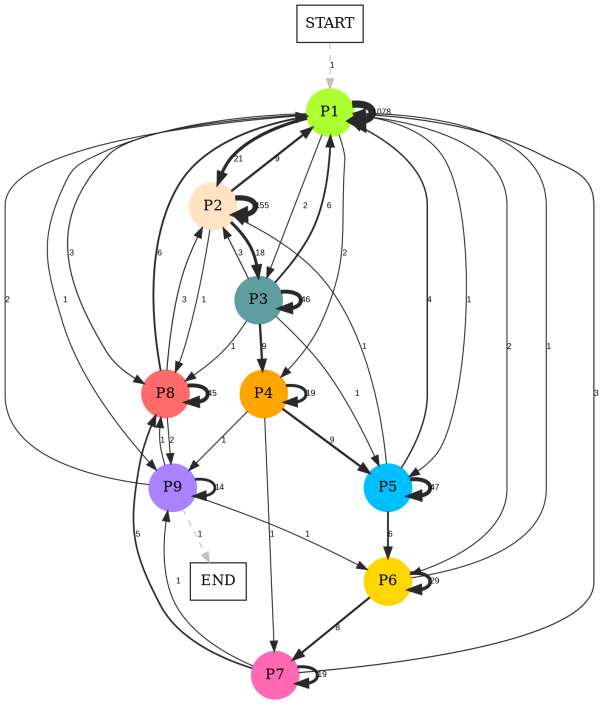


Figura 8: Análisis de procesos del grupo `DBA 1516 P2 GA` (`Activity` problema y `CaseId` grupo) obtenido con la implementación propia.

Además, como podemos ver en la Figura 9, hemos obtenido el mismo diagrama que en el de la Figura 7 con la salvedad de que hemos impedido el retorno a un estado anterior (el motivo se verá más adelante). Es decir, se han eliminado ciclos.

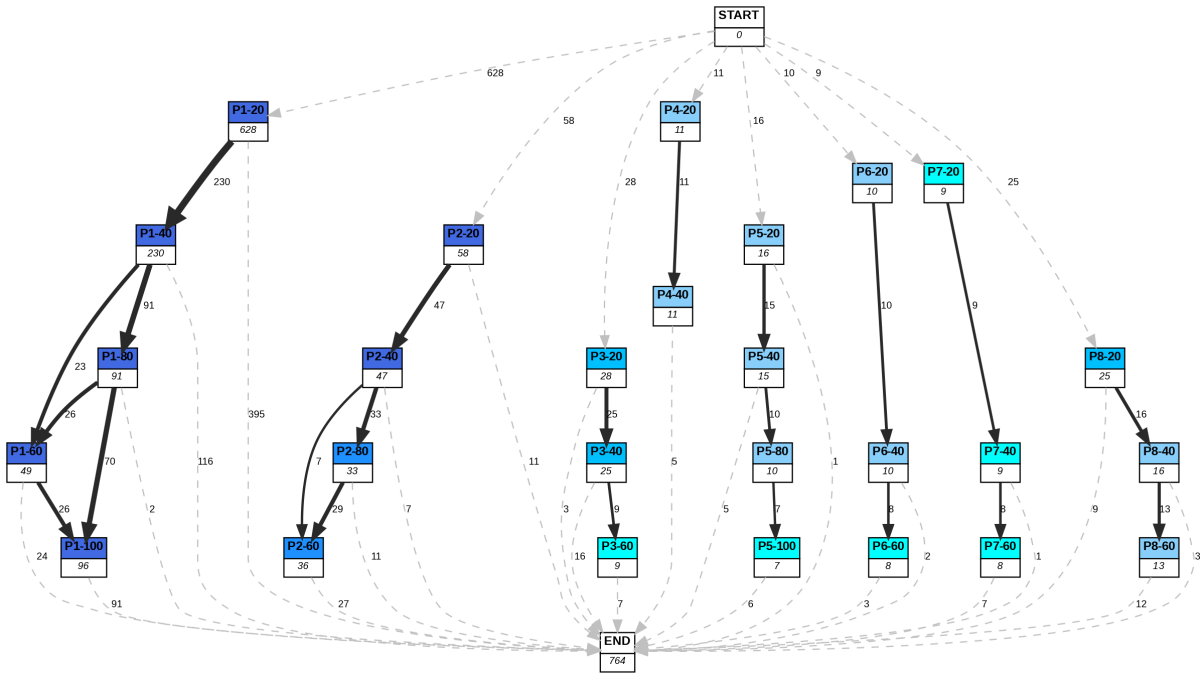


Figura 9: Análisis de procesos del grupo `DBA 1516 P2 GA` (`Activity` problema-milestone y `CaseId` sesión) obtenido con la implementación propia.

A partir de ahora, estos diagramos tendrán la consideración de grafos. En particular, serán grafos dirigidos y operaremos con ellos como tales. En el siguiente capítulo se expondrán los conceptos básicos de grafos y principales resultados matemáticos que se usarán en este trabajo fin de grado.

4

TEORÍA DE GRAFOS

En el ámbito de las matemáticas y las ciencias de la computación, se emplea el término *grafo* (del griego *grafos* que significa *dibujo* o *imagen*) para referirse a un conjunto de objetos llamados *vértices* o *nodos*, los cuales están unidos por enlaces conocidos como *aristas* o *arcos*. Estas conexiones representan las relaciones binarias que existen entre los elementos de un conjunto, y son objeto de estudio de la teoría de grafos.

4.1 GRAFOS

En esta sección se introducirán las deficiones básicas que forman parte de la teoría de grafos.

Definición 1. Matemáticamente, un *grafo* $G = (V, E)$ es una tupla de vértices V y aristas E que relacionan dichos vértices. Denominaremos *orden* del grafo al número de vértices del mismo ($|V|$). Por supuesto, siempre tendremos que $V \neq \emptyset$.

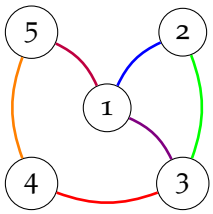


Figura 10: Ejemplo de grafo simple.

Ejemplo 1. El grafo dado en la Figura 10 tiene conjunto de vértices $V = \{1, 2, 3, 4, 5\}$ y conjunto de aristas $E = \{(1, 2), (1, 3), (2, 3), (3, 4), (4, 5), (5, 1)\}$.

Definición 2. Un *vértice* o *nodo* es la unidad fundamental de las que se componen los grafos. Los vértices en sí mismos se tratan como objetos indivisibles y sin propiedades. No obstante, pueden tener asociados una semántica dependiendo del contexto de aplicación del grafo. Por ejemplo, en el grafo 9 un nodo representa la consecución de un objetivo de un problema.

Definición 3. Una *arista* representa una relación entre dos vértices de un grafo. Las aristas se denotan por $(u, v) \in E$ donde $u, v \in V$. Visualmente, se representan como las líneas que unen los vértices que forman parte de la definición de la misma.

Definición 4. Un *grafo podenderado* es un grafo cuyas aristas tienen un peso o valor asociado.

Formalmente, se puede definir como un trío ordenado $G = (V, E, W)$ donde $V = \{v_1, \ldots, v_n\}$ es un conjunto de vértices, $E = \{e_1, \ldots, e_m\}$ y $W = \{w_1, \ldots, w_m\}$ es el conjunto de pesos asociados a cada arista.

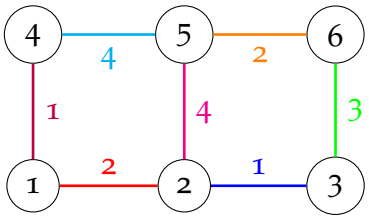


Figura 11: Ejemplo de grafo ponderado.

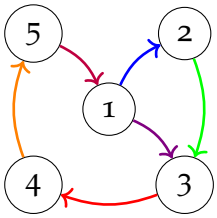


Figura 12: Ejemplo de grafo dirigido.

Definición 5. Un *grafo no dirigido* es un grafo cuyas aristas representan relaciones simétricas y carecen de sentido definido. Es decir, la arista (u, v) es idéntica a la arista (v, u). Es decir, las aristas no son pares ordenados sino conjuntos $\{u, v\}$ (o 2-multiconjuntos) de vértices.

Un grafo no dirigido podrá tener, a lo más, $\frac{|V|^2}{2}$ aristas.

Definición 6. Se denomina *grafo dirigido* o *digrafo* a aquellos grafos cuyas aristas tengan un sentido definido. En un digrafo, cada arista se representa como un par ordenado de dos vértices. Por ejemplo, (u, v) denota la arista que va de u hacia v (desde el primer vértice hasta el segundo vértice).

Los grafos no dirigidos se pueden ver como un caso particular de los grafos dirigidos en tanto que son grafos dirigidos simétricos.

Mientras que en un grafo no dirigido se tiene que $E \subseteq \{x \in \mathcal{P}(V) : |x| = 2\}$ (es decir, E es un conjunto de pares no ordenados de elementos de V), cuando el grafo es dirigido se tiene que E es un conjunto de pares ordenados $(i, j) \in V \times V$.

Ejemplo 2. En la Figura 12 se muestra un ejemplo de grafo dirigido mientras que en la Figura 10 tenemos un ejemplo de grafo no dirigido.

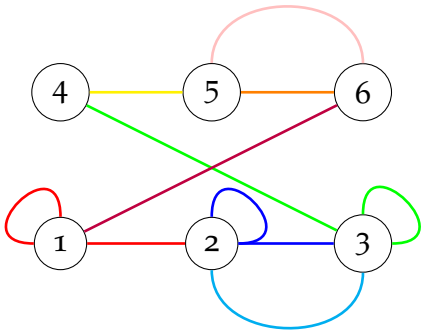


Figura 13: Ejemplo de multigrafo.

Definición 7. Un *grafo conexo* es un grafo en que todos sus vértices están conectados por un camino o por un semicamino dependiendo de si el grafo es no dirigido o dirigido.

De lo contrairo, si algún grafo no cumple la propiedad anterior se dirá que es *disconexo*.

Definición 8. Un *bucle* es una arista que relacionado un vértice consigo mismo.

Definición 9. En un grafo $G = (V, E)$, se dice que dos aristas son *paralelas* o *múltiples* si el vértice inicial y el vértice final de las mismas coinciden.

Los grafos que permiten la existencia de bucles y aristas múltiples se denominan *multigrafos*. Por el contrario, los grafos sin bucles y sin aristas paralelas se denominarán *simples*.

Ejemplo 3. En la Figura 11 tenemos un ejemplo de grafo ponderado.

En la Figura 10 se muestra un ejemplo de grafo simple. Por otro lado, en la Figura 13 podemos ver un multigrafo.

Definición 10. En un grafo $G = (V, E)$ dos vértices se dirán *adyacentes* (o *vecinos*) si están relacionados por al menos una arista. Es decir, dos vértices $u, v \in V$ son adjacentes si $\exists e \in E$ tal que $e = (u, v)$.

La *matriz de adyacencia* de un grafo es una matriz cuadrada de dimensión $|V| \times |V|$ que se utiliza como forma de representar las relaciones binarias entre los nodos del mismo. La denotaremos por $A = (a_{ij})_{1 \leq i,j \leq |V|}$.

Si tenemos que G es un grafo no dirigido, entonces $a_{ij} = 1$ y $a_{ji} = 1$ si el vértice v_i es adyacente al vértice v_j y $a_{ij} = a_{ji} = 0$ en caso contrario. Si el grafo G es dirigido, entonces tendremos que $a_{ij} = 1$ si y sólo si existe $e \in E$ tal que $e = (v_i, v_j)$ y $a_{ij} = 0$ en caso contrario.

Por último, si tenemos un grafo ponderando, entonces se sustiuirá en valor de 1 en los casos anteriores por el peso de las aristas correspondientes.

Ejemplo 4. Tenemos que la matriz de adyacencia del grafo de la Figura 10 es:

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{1}$$

Definición 11. Sea $G = (V, E)$ un grafo no dirigido y sea $v \in V$ un vértice suyo. Se denomina grado del vértice v al número de aristas incidentes al vértice y se denotará de ahora en adelante por $\deg(v)$.

Al conjunto de todos los vértices adyacentes a un vértice dado se le denominará *vecindad* del vértice en cuestión. Formalmente, la vecindad de un vértice $v \in V$ es el conjunto

$$N(v) = \{u \in V | \{v, u\} \in E\} \tag{2}$$

Así pues, el grado de un vértice $v \in V$ puede definirse como el módulo de su vecindario: $\deg(v) = |N(v)|$.

En el caso de los grafos dirigidos se distingue entre el *grado de entrada* $\deg^-(v)$ (número de aristas que tienen a v como el vértice final) y el *grado de salida* $\deg^+(v)$ (número de ariastas que tienen a v como vértice inicial).

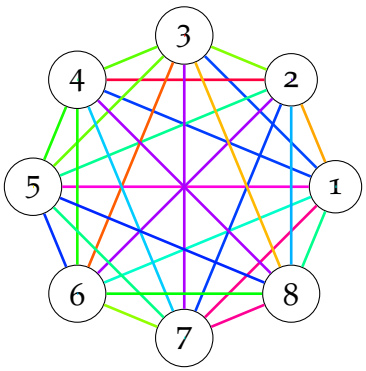


Figura 14: Ejemplo de grafo completo.

Definición 12. Un grafo en el que todos sus vértices tienen el mismo grado (de entrada, en el caso de los grafos dirigidos) se denomina *regular*. Además, un grafo con vértices de grado k se llamará k-regular.

Definición 13. Un *grafo completo* $G = (V, E)$ es un grafo no dirigido simple en el que para cada par de vértices $u, v \in V$ existe una arista $e \in E$ tal que $e = \{u, v\}$.

El *grafo completo de n vértices* se denotará por K_n. Así pues, K_n tendrá $fracn \cdot (n-1)2$ aristas y es un grafo regular de grado $n-1$.

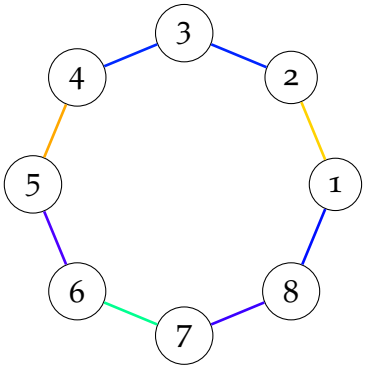


Figura 15: Ejemplo de grafo ciclo.

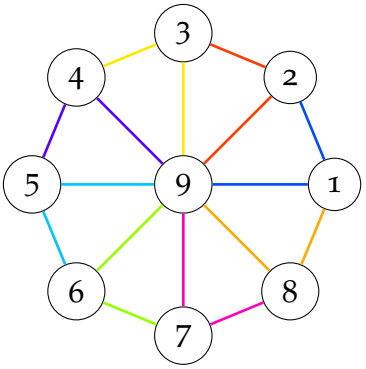


Figura 16: Ejemplo de grafo rueda.

Definición 14. Un *grafo ciclo* o simplemente un *ciclo* es un grafo que consiste en un camino simple cerrado. Esto es, hay un único camino en el que no se repite ningún vértice salvo el primero con el último.

Denotaremos a un grafo ciclo de n vértices por C_n. Si consideramos que es un grafo no dirigido, cada vértice tendrá un vecindario de tamaño 2 y, por tanto, será un grafo 2-regular. Por el contrario, si tenemos un grafo dirigido, será un grafo 1-regular.

Definición 15. Un grafo rueda es un grafo de n vértices (denotado usualmente por W_n) es un grafo que se obtiene al añadir un único vértice a un grafo ciclo de $n-1$ vértices, conectando el nuevo el vértice a todos los ya existentes. Es decir, el nuevo vértice será adyacente a todos los vértices del grafo C_{n-1}.

Ejemplo 5. En las Figuras 14, 15 y 16 podemos ver un grafo completo, un grado ciclo y un grafo rueda respectivamente.

Definición 16. Diremos que un grafo es *cíclico* si contiene al menos un grafo ciclo. Por el contrario, se dirá que un grafo es *acíclico* si no contiene ningún ciclo.

No obstante, en este trabajo fin de grado nos centraremos en los llamados *grafos dirigidos acíclicos* o *DAG* (*Directed Acyclic Graphs*, en inglés) que no son más que grafos dirigidos desprovistos de ciclos.

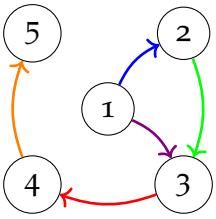


Figura 17: Ejemplo de grafo acíclico dirigido con 5 nodos.

Ejemplo 6. En la Figura 10 tenemos un grafo dirigido con ciclos o cíclico (contiene, por ejemplo, el 1-3-4-5). Sin embargo, eliminado una de las aristas del mismo obtenemos el grafo de la Figura 17, que es acíclico.

Definición 17. Un grafo conexo acíclico no dirigido se denominará *árbol*. Por otro lado, un *árbol orientado* o *poliárbol* será un grafo dirigido acíclico cuyo grafo no dirigido subyacente es un árbol. De otra manera, si cambiamos sus aristas dirigidas por no diridas, se obtendía un grafo no dirigido conexo y acíclico.

Definición 18. Un *árbol de expansión* de un grafo conexo no dirigido G es un subgrafo suyo que es árbol y que contiene a todos sus vértices.

El *número de árboles de expansión* de un grafo conexo G, habitualmente denotado por $t(G)$, es un invariante importante en la teoría de grafos. Éste puede obtenerse mediante el denominado *Teorema de Kirchhoff*. Este teorema demuestra que el número de árboles de expansión de un grafo puede obtenerse en tiempo polinómico a partir del determinante de una submatriz de la *matriz Laplaciana* del grafo. Más aún, nos dice que éste número es igual a cualquier cofactor de la matriz Laplaciana. El Teorema de Kirchhoff es una generalización de la *fórmula de Cayley*, que proporciona el número de árboles de expansión en el caso de un grafo completo y que veremos a continuación.

Proposición 1. *Dado un grafo completo $K_n = (V, E)$ con $V = \{v_1, v_2, \ldots, v_n\}$, la fórmula de Cayley establece que el número de árboles de expansión del mismo es $t(K_n) = n^{n-2}$.*

En 1918, el alemán H. Prüfer obtuvo una elegante correspondencia biyectiva entre árboles etiquetados con n vértices y sucesiones de longitud $n-2$, denominadas *códigos de Prüfer*.

Definición 19. La definición del *código de Prüfer* de un árbol $T = (V, E)$ no trivial, denotado por $P(T)$, es recursiva. Si $|V| = 2$ entonces T consiste de una sola arista y $P(T) = \emptyset$. Supongamos ahora que el código de Prüfer de cualquier árbol con n vértices está definido y sea $T = (V = \{v_1, v_2, \ldots v_n, v_{n+1}\}, E)$ un árbol con $n+1$ vértices. Sea

$$v = \text{mín}\{i \in \{1, 2, \ldots, n, n+1\} \mid \deg(v_i) = 1\} \tag{3}$$

y sea u el único vértice adyacente a v en T. Por lo tanto, $T' = T - v$ es un árbol con n vértices y $P(T-v)$ está bien definido por hipótesis de inducción. El código de Prüfer de T se definirá de la siguiente forma:

$$P(T) = (u, P(T')) \tag{4}$$

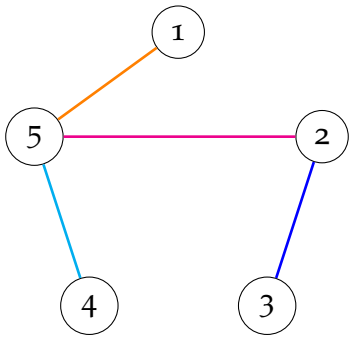


Figura 18: Ejemplo de árbol con 5 nodos.

Ejemplo 7. Consideremos el árbol de la Figura 18. Tenemos que el vértice de grado uno con la numeración más pequeña es el vértice 1. Este vértice únicamente es adayacente al vértice número 5. Así pues, $P(T) = (5, P(T-1))$. Cuando eliminamos el primer vértice, obtenemos que el vértice de grado uno de menor numeración es el 3, cuyo único vértice adyacentes es el 2, lo que conduce a que $P(T) = (5, 2, P(T - \{1,3\}))$. Eliminando ahora este tercer vértice, tenemos que el vértice 2 es el de menor numeración cuyo grado es 1 y su único nodo adyacente es el número 5. Esto hace que $P(T) = (5, 2, 5, P(T - \{1,3,2\}))$. Como $T - \{1,3\}$ es un árbol con dos vértices, $P(T - \{1,3\}) = \emptyset$ y $P(T) = (5,2,5)$.

Este ejemplo pone de manifiesto que no es necesario que todos los vértices aparezcan en el códgo de Prüfer y que pudiera ocurrir que un mismo vértice aparezca más de una vez en los mismos. De hecho, el número de veces que un vértice aparece en el código de Prüfer depende del grado de dicho vértice. Este resultado se verá en el siguiente Lema.

Proposición 2. *Sea c_i el número de veces que aparece el número i en el código de Prüfer de un árbol $T = (V = \{1, \ldots, n\}, E)$ con $n \geq 3$ vértices. Entonces $deg(i) = c_i + 1$.*

Demostración. Si $n = 3$ entonces $P(T)$ consiste de un solo número, correspondiente al vértice de grado dos.

Supongamos ahora que el resultado es cierto para todo árbol $T = (V = \{1, \ldots, n\}, E)$ con $n \geq 3$ vértices. Sea $T' = (V' = \{1, \ldots, n, n+1\}, E')$, $v = \min\{i \in V' | \deg(i) = 1\}$ y sea u el único vértice adyacente a v en T'. Así pues, $P(T') = (u, P(T' - v))$. Para cada $i \in V'$ sea b_i el número de veces que aparece i en $P(T' - v)$. Por hipótesis de inducción, $\deg_{T'-v}(i) = b_i + 1$.

Además, si $i \neq u$ entonces $c_i = b_i$ y $\deg_{T'}(i) = \deg_{T'-v}(i)$. Por lo tanto, en este primer caso, $\deg_{T'}(i) = c_i + 1$. Por otra parte, si $i = u$ entonces $c_i = b_i + 1$ y $\deg_{T'}(i) = \deg_{T'-v}(i) + 1$. Por lo tanto, $\deg_{T'}(i) = b_i + 2 = c_i + 1$. □

El siguiente resultado muestra que el código de Prüfer define una función inyectiva del conjunto de árboles generadores con n vértices al conjunto de palabras de longitud $n-2$ del alfabeto $\{1, 2, \ldots, n\}$.

Proposición 3. *Si T y T' son dos árboles con $n \geq 3$ vértices numerados tales que $P(T) = P(T')$, entonces $T = T'$.*

Demostración. Si $n = 3$, entonces $P(T)$ consiste de un solo número, correspondiente al único vértice de grado dos de T. Como $P(T) = P(T')$, este vértice también es el único vértice de grado dos de T' y tenemos que $T = T'$.

Supongamos ahora que el resultado es cierto para cualesquiera dos árboles con n vértices numerados. Sean ahora T y T' dos árboles con $n+1$ vértices numerados tales que $P(T) = P(T')$. Sea v el mínimo elemento de $\{1, \ldots, n, n+1\}$ que no aparece en $P(T)$. Por el lema anterior $\deg_T(v) = 1$ y $\deg_{T'} = 1$. Así pues, existe un único u tal que u es adyacente a v en T y existe un único u' tal que u' es adyacente a v en T'. De ahí obtenemos que $P(T) = (u, P(T-v))$ y que $P(T') = (w, P(T-v'))$. Como $P(T) = P(T')$, se sigue que $u = u'$ y, por lo tanto, $P(T-v) = P(T'-v')$. Por lo que, por hipótesis de inducción, $T - v = T' - v$. Concluimos así que $T = T'$. □

A continuación veremos que a cada palabra de longitud $n-2$ del alfabeto $\{1, 2, \ldots, n\}$ le corresponde un árbol cuyo código de Prüfer es esa palabra.

Proposición 4. *Sea $n \geq 3$. Si $L = (u_1, u_2, \ldots, u_{n-2})$ es una lista cuyos elementos pertenecen al conjunto $V = \{1, \ldots, n\}$, entonces existe un árbol con vértices numerados (o etiquetado) T tal que $P(T) = L$.*

Demostración. Si $n = 3$ entonces $L = (u_1)$ y T es la trayectoria de longitud dos cuyo vértice interno es u_1.

Supongamos ahora que el resultado es cierto para toda lista de longitud $n-2$. Sea $L = (u_1, u_2, \ldots, u_{n-1})$ una lista de longitud $n-1$. Sea v el elemento más pequeño de V que no aparece en L. Por hipótesis de inducción existe un árbol $T' = (V', E')$ con conjunto de vértices $V_{T'} = \{1, \ldots, n+1\} - \{v\}$, tal que $P(T') = (u_2, u_3, \ldots, u_{n-1})$. Sea e la arista que une a v con u_1 y sea $T = (V' + v, E' + e)$. Tenemos que T es un árbol y que $P(T) = (u_1, P(T')) = L$ tal y como se quería demostrar. □

El siguiente ejemplo pone en práctica el procedimiento descrito en el lema anterior para construir un árbol generador cuyo código de Prüfer sea igual a una lista dada.

Ejemplo 8. Consideremos la lista $(3,5,3,1)$. La longitud de esta lista es 4, por lo que corresponde a un árbol con conjunto de vértices $V = \{1,2,3,4,5,6\}$. El primer vértice que no aparece en la lista es 2, el cual deber ser adyacente al vértice 3 (Figura 19). Consideremos ahora la sublista $(5,3,1)$, el primer vértice del conjunto $\{1,3,4,5,6\}$ (obtenido al eliminar el vértice 2) que no aparece en la lista es el 4, que debe ser adyacente al vértice 5. Análogamente, el primer vértice del conjunto $\{1,3,5,6\}$ que no aparece en la sublista $(3,1)$ es el 5, el cual debe ser adyacente al vértice 3, y el primer vértice del conjunto $\{1,3,6\}$ que no aparece en la sublista (1) es el 3, el cual debe ser adyacente al vértice 1. Finalmente, obtenemos la lista vacía y el conjunto $\{1,6\}$, lo cual indica que el vértice 1 debe ser adyacente al vértice 6. La Figura **??** es el árbol asociado a la lista $(3,5,3,1)$.

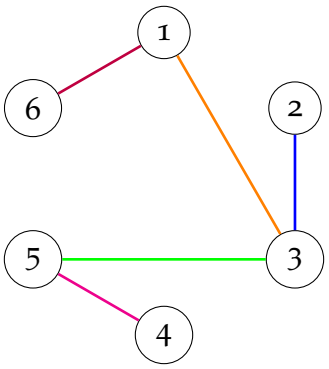


Figura 19: Ejemplo de árbol con 6 nodos.

Finalmente, se procederá a demostrar la fórmula de Cayley 1:

Demostración. Si $n = 1$ o $n = 2$ el resultado es trivialmente cierto. Si $n \geq 3$ de las Proposiones 3 y 4 se sigue que existe una correspondencia biyectiva entre el conjunto de árboles generadores con n vértices y el conjunto de palabras de longitud $n-2$ del alfabeto $\{1,2,\ldots,n\}$. Como el número de palabras de longitud $n-2$ de un alfabeto con n elementos es n^{n-2}, entonces hay n^{n-2} árboles de expansión distintos de un grafo completo con n vértices. □

A continuación, se introducirá el concepto de matriz laplaciana de un grafo que, junto con el Teorema de Kirchhoff nos permitirá calcular el número de árboles de expansión de un grafo arbitrario.

Definición 20. La matriz laplaciana (también conocida como matriz de admitancia o matriz de Kirchhoff) es una representación matricial de un grafo muy utilizada en la Teoría espectral de grafos, cuyo objetivo es el estudio de las propiedades de los grafos en relación de los polinomios caracterísicos, valores y vectores propios de las matrices asociadas a los mismos.

Para un grafo simple G con vértices $V = (v_1, \ldots, v_n)$ los elementos de la matrix laplaciana $L_{n\times n}$ se definen como sigue:

$$L_{i,j} = \begin{cases} \deg(v_i) & \text{si } i = j \\ -1 & \text{si } i \neq j \text{ y } v_i \text{ es adyacente a } v_j \\ 0 & \text{en cualquier otro caso} \end{cases} \tag{5}$$

Equivalentemente, se tiene que $L = D - A$ donde D es la matriz de grados del grafo (matriz diagonal cuyos elementos no nulos son los grados de cada uno de los vértices) y A es la matriz de adyacencia del grafo.

Ejemplo 9. Se tiene que la matriz laplaciana del grafo simple de la Figura 10 es la siguiente:

$$\begin{pmatrix} 3 & -1 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{pmatrix} \tag{6}$$

Este concepto se puede generalizar al caso de grafos ponderados, donde las matrices de adyacencia pueden contener números naturales distintos de ceros y unos. Además, también se puede generalizar a grafos dirigidos, utilizando en vez de la matriz de grados del grafo la matriz de grados de entrada o la matriz de grados de salida dependiendo de la aplicación que se esté considerando.

Ejemplo 10. La matriz laplaciana del grafo simple de la Figura 11 será la siguiente:

$$\begin{pmatrix} 3 & -2 & 0 & -1 & 0 & 0 \\ -2 & 7 & -1 & 0 & -4 & 0 \\ 0 & -1 & 4 & 0 & 0 & -3 \\ 0 & -4 & 0 & -4 & 10 & -2 \\ 0 & 0 & -3 & 0 & -2 & 5 \end{pmatrix} \tag{7}$$

Ejemplo 11. Las matrices lapacianas del grafo dirigido de la Figura 12 de entrada y de salida se muestran en las ecuaciones **??** y **??** teniendo en cuenta la matrices de adyacencia (Ecuación 8), del grado de entrada y del grado de salida de dicho grafo.

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{8}$$

Teorema 1. *El Teorema de Kirchhoff. Sea un grafo conexo G con n vértices numerados y sean* λ_1, λ_2, $\ldots$, λ_{n-1} *los valores propios no nulos de su matriz laplaciana. Se tiene entonces que el número de árboles de expansión del grafo G es*

$$t(G) = \frac{1}{n} \cdot \lambda_1 \cdots \lambda_{n-1} \tag{9}$$

Como podemos ver, se trata de una generalización de la fórmula de Cayley que además de muestra que el número de árboles de expansión de cualquier grafo se puede calcular en tiempo polinómico a partir del determinante de una submatriz de la matriz laplaciana. Específicamente, el número de árboles de expansión de un grafo conexo coincide con cualquier cofactor de su matriz laplaciana.

4.2 MEDIDAS DE COMPLEJIDAD DE PROPÓSITO GENERAL

Esta sección es una recopilación de varias de las medidas de complejidad aplicadas a grafos dirigidos acíclicos ponderados $G = (V, E)$ a partir de sus matrices de adyacencia ponderadas. Se denotará a las matrices de adyacencia por $\mathcal{A}$ y se tendrá que el número de vértices es $v = |V| = 2 \cdot n + 1$, $n \in \mathbb{N}$. Es decir, tendremos un número impar de vértices.

Todas ellas tratan de medir la complejidad de un grafo en términos de la densidad de las conexiones entre los vértices utilizando diferentes perspectivas.

4.2.1 *Peso del comportamiento*

Detecta cuánto esfuerzo han dedicado los alumnos a sus tareas de laboratorio en términos del número total de sesiones abiertas en el servidor. Matemáticamente se define como sigue:

$$We(\mathcal{A}) = \sum_r \sum_c \mathcal{A}[r][c] \tag{10}$$

4.2.2 *Eficacia*

Detecta cuántos problemas se han resuelto. Se formula como sigue:

$$Ef(\mathcal{A}) = \left| \left\{ p_i \in P, \sum_r \mathcal{A}[r][p_i^s] > 0 \right\} \right| \tag{11}$$

4.2.3 *Densidad*

Es el cociente entre el peso del grafo y el número máximo de aristas permitido:

$$De(\mathcal{A}) = \frac{W(\mathcal{A})}{v \cdot (v-1)} \tag{12}$$

4.2.4 *Grado medio*

El grado de un nodo es el número de aristas entrantes y salientes o, en el caso de los grafos ponderados, la suma de los pesos entrantes y salientes.

$$Dm(\mathcal{A}) = \frac{1}{v} \cdot \sum_i (\sum_r \mathcal{A}[r][i] + \sum_c \mathcal{A}[i][c]) \cdot f \tag{13}$$

4.2.5 *Longitud del camino característico*

Si interpretamos los pesos (frecuencias) como distancias, ya que ambos se refieren a una especie de esfuerzo para llegar a una solución, podríamos aplicar cualquier algoritmo de camino más corto como el de Dijkstra (**incluir referencia**) a la matriz característica $\mathcal{A}$ para propagar las frecuencias y obtener un cierre de la matriz original $\hat{\mathcal{A}}$. Es decir, obtendremos una aproximación de cuántas sesiones habría costado llegar desde cualquier nodo del grafo a cualquier otro nodo. Esto es,

$$Le(\mathcal{A}) = \frac{1}{E(\hat{\mathcal{A}})} \cdot \sum_r \sum_c \hat{\mathcal{A}}[r, c] \tag{14}$$

4.2.6 *Diámetro del grafo*

Sólo trata de medir el número de aristas entre los nodos más distantes del grafo.

$$Di(\mathcal{A}) = \max \hat{\mathcal{A}}[r, c] \tag{15}$$

4.2.7 *Conectividad*

Número medio de nodos conectados a un determinado nodo.

$$Co(\mathcal{A}) = \frac{1}{v} \cdot \sum_{i} (\sum_{r} \mathcal{A}'[r][i] + \sum_{c} \mathcal{A}'[i][c]) \tag{16}$$

4.2.8 *Betweenness*

Es muy conocido en las redes sociales e identifica cuál de los problemas ha recibido más atención e influencia que los demás.

Parte III

ANÁLISIS DESCRIPTIVO

Análisis estadístico de los datos.

5

LOS REGISTROS EXISTENTES

En el laboratorio remoto se ha registrado la actividad de 7 años consecutivos (desde el curso académico 1516 al 2122). Un ejemplo de las interacciones almacenadas puede verse en la Tabla 1.

Cuadro 1: Muestra de los datos que se recopilan en el servidor.

Year	Group	SessionID	Date	Problem	Step
1819	Keid	493252533735	28/10/2018 20:23:35	P1	1
1819	Keid	493252533735	28/10/2018 20:23:40	P1	3
1819	Keid	389034076811	7/11/2018 19:01:49	P2	1
1819	Cerastes	487544594557	27/10/2018 13:05:11	P1	1
1819	Cerastes	487544594557	27/10/2018 13:10:57	P1	3
1819	Jabbah	550676318711	20/12/2018 22:22:42	P8	1
1819	Cerastes	336303012053	17/12/2018 13:28:50	P9	1
1819	Keid	563159878397	25/10/2018 12:41:43	P8	1

5.1 NÚMERO DE GRUPOS CADA AÑO

El número de grupos puede variar en cada curso en función del número de alumnos matriculados en la asignatura ese año. Así pues, se muestran a continuación en las Tablas 17 y 18 los grupos por curso académico. El número de grupos por año puede consultarse también en la Figura 20.

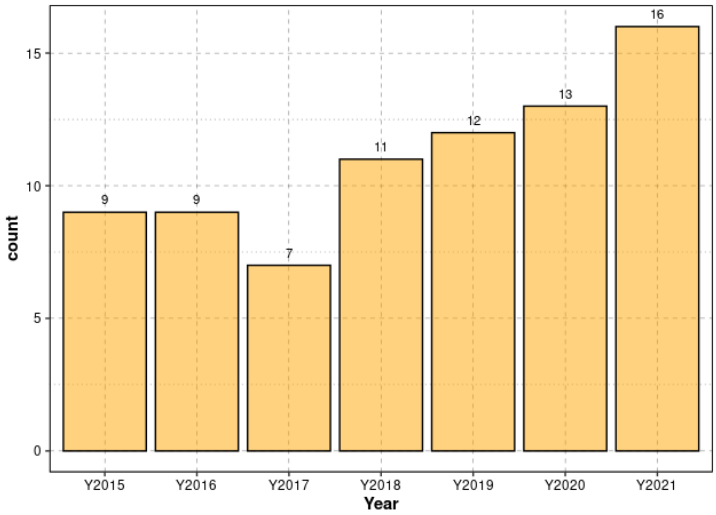


Figura 20: Número de grupos por curso académico estudiado.

5.2 EL PERIODO DE TIEMPO ANALIZADO CADA AÑO

En la Tabla 2 se muestra el número de días que dura la práctica cada año. Se puede apreciar que la duración de la práctica que estamos considerando puede variar en función del curso académico.

Cuadro 2: Número de días que dura la práctica cada año.

Year	Length (days)
Y2015	33
Y2016	24
Y2017	30
Y2018	18
Y2019	28
Y2020	17
Y2021	39

5.3 EL CONJUNTO DE PROBLEMAS ANALIZADOS CADA AÑO

Todos los años hay 9 problemas de dificultad similar que deben ser resueltos por todos los grupos.

Para aproximarnos al concepto subjetivo de "dificultad del problema" vamos a analizar el número de sesiones fallidas que necesita cada alumno para resolverlos por primera vez con

respecto al número total de sesiones de ese problema (tasa de fallo) y la duración de este periodo en horas.

5.3.1 *Dificultad del problema: la tasa de fallo*

La apertura de un problema se corresponde con una sesión de trabajo, la cual puede terminar como fallo (fail) si no se consigue resolver el problema, o éxito (solved) en caso de que se haya resuelto el problema. Así pues, se definirá la tasa de fallo como el cociente entre el número total de sesiones fallidas entre el número total de sesiones de un mismo problema. El boxplot de las tasas de fallo por problema puede verse en la Figura 21.

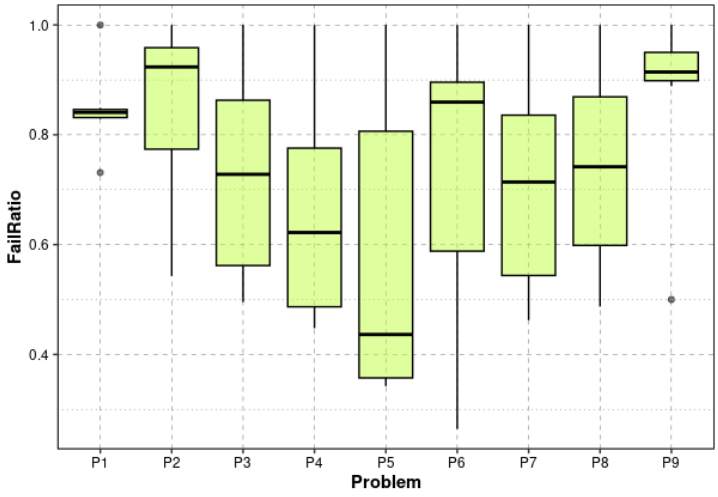


Figura 21: Boxplot de la tasa de fallo (Fail ratio) por problema.

Tras realizar el test ANOVA de un factor (resultados en la Tabla 3), cuya hipótesis nula establece que la tasa de fallo media de los nueve problemas considerados es la misma, se detecta que las distribuciones de probabilidad de la tasa de fallo son estadísticamente iguales en los distintos problemas ($p = 0{,}1733 > 0{,}05$)[1].

Cuadro 3: Resultados del test ANOVA de un solo factor (tasa de fallo).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ndsp[[nVariable]]	8	0.51	0.06	1.52	0.1733
Residuals	54	2.26	0.04		

1 Nótese que hemos establecido un nivel de significancia de $\alpha = 0{,}05$.

Además, se ha realizado un test de Tukey por pares de problemas (Tabla 20). En él se observa que casi todos los pares pueden considerarse estadísticamente iguales (p adj $>$ 0,2 en todos ellos) salvo quizá, el par P9-P5 (p adj $=$ 0,2). La Figura 22 muestra los intervalos de confianza de todas las diferencias entre las distintas parejas de años.

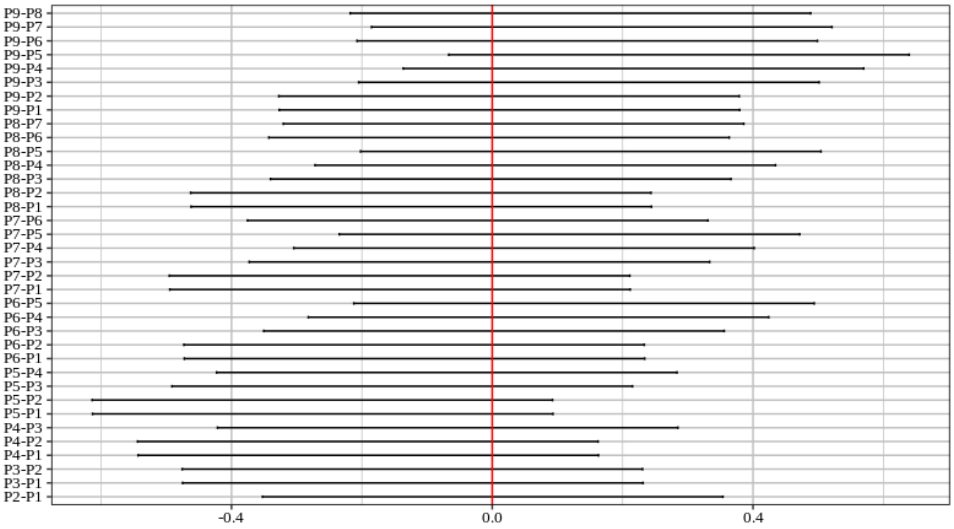


Figura 22: Intervalos de confianza de la tasas de fallo de los problemas.

5.3.2 *Dificultad del problema: tiempo necesario en resolverlo*

Es el número de horas que transcurren desde que el problema se abre por primera vez hasta que es resuelto por primera vez.

Falta imagen.

De nuevo los tests detectan comportamientos diferentes (ANOVA p=9.27e-5, KW p=8.8e-8).

Falta tabla.

Por pares.

Falta tabla.

Intervalos de confianza.

Falta imagen.

Por lo tanto, se puede ver, dadas las evidencias aportadas que la resolución de cada problema exige respuestas claramente diferentes por parte del alumnado.

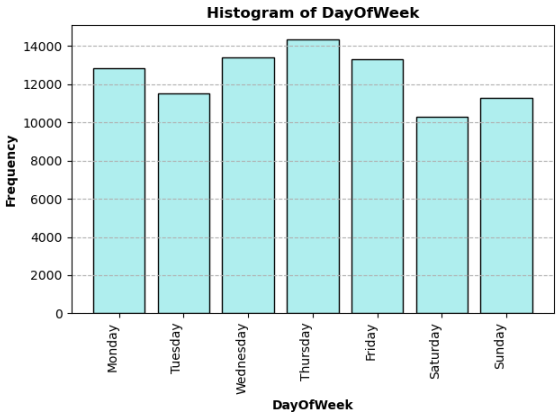
5.4 ACTIVIDAD REGISTRADA

El número de registros y de sesiones de trabajo de cada uno de los años analizados se muestran en la Tabla 4. Como podemos ver, aunque el curso académico 2021 registra más actividad que los demás, no es el que presenta un mayor número de sesiones.

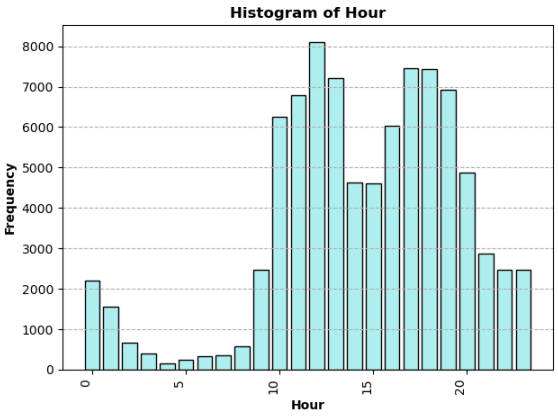
Cuadro 4: Número de registros y sesiones almacenados en el servidor por años.

Year	Activity Records	Sessions
Y2015	12088	4489
Y2016	12525	4538
Y2017	9088	3661
Y2018	5705	2811
Y2019	14475	5156
Y2020	21188	3900
Y2021	11961	6113

Al ser un servicio 24 horas los 7 días de la semana, los alumnos interactúan con el laboratorio remoto en cualquier día de la semana tal y como puede verse en la Figura 23a y a cualquier hora del día (Figura 23b).



(a) Histograma de los días de la semana.



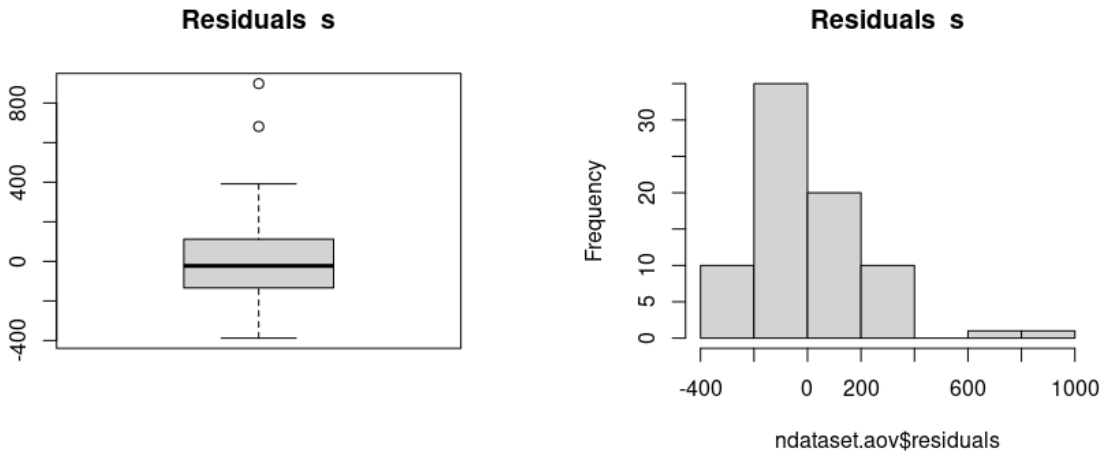
(b) Histograma de las horas del día.

Figura 23: Actividad registrada en el servidor remoto.

El número y tipo de las sesiones de trabajo de cada uno de los grupos puede contemplarse en la Tabla 19.

5.4.1 *Análisis de la normalidad de la distribución del número de sesiones*

En las Figuras 24a y 24b podemos ver el boxplot de los residuos y el histograma de los mismos.



(a) Boxplot de los residuos del número de sesiones.

(b) Histograma de los residuos del número de sesiones.

Figura 24: Distribución de los residuos del número de sesiones.

A continuación, en las Figuras 25 y 26, podemos observar que la distribución del número de sesiones no es perfectamente normal pero es casi-normal si eliminaremos algunos outsiders.

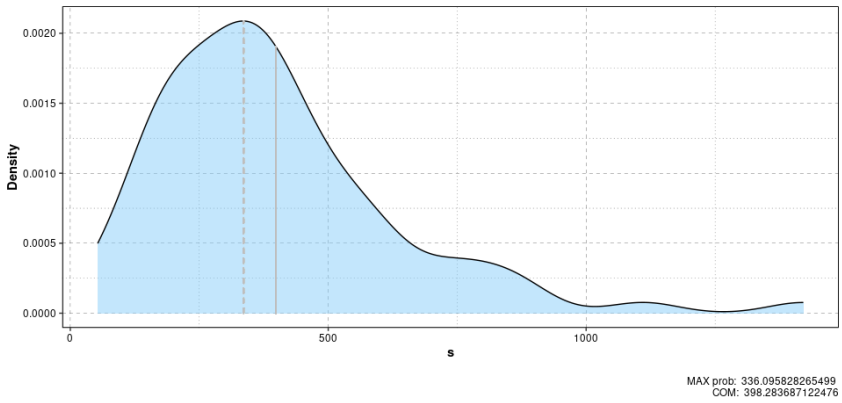


Figura 25: Función de densidad de probabilidad del número de sesiones.

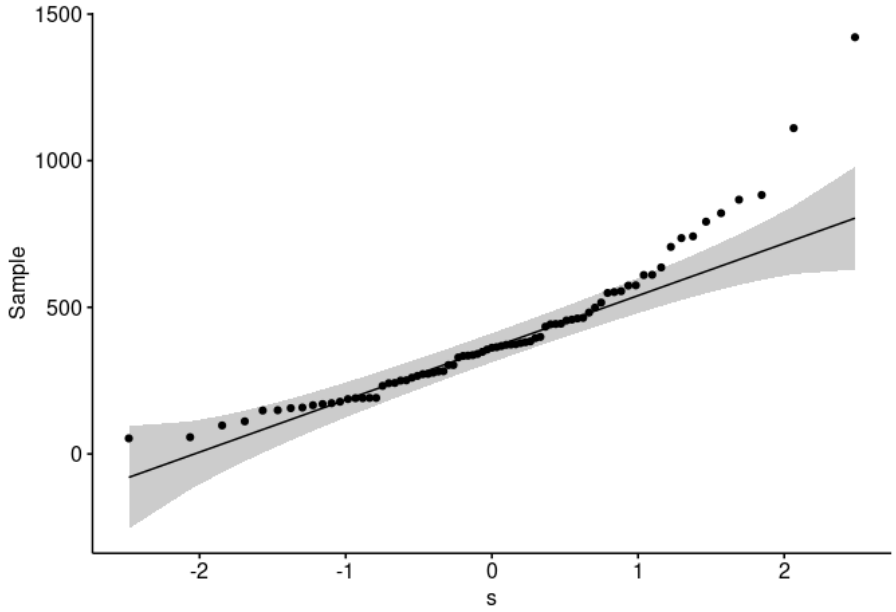


Figura 26: Gráfico Q-Q del número de sesiones.

Además, podemos ver en la Figura 28 que hay algunos outliers (867, 883, 1421 y 1111) considerando la distribución del número total de sesiones por grupo de alumnos. Segmentando por años, obtenemos los boxplots que se muestran en la Figura 27. Así pues, eliminaremos aquellos registros que sean outliers en todos los años. Tras realizar la acción anterior, obtenemos la distribución del número de sesiones que se muestra en la Figura 29.

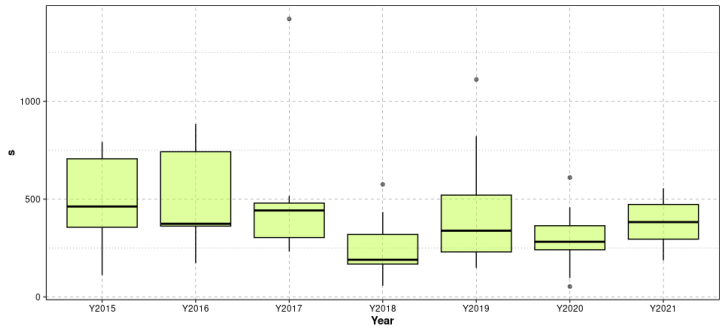


Figura 27: Boxplot del número de sesiones por año inicialmente.

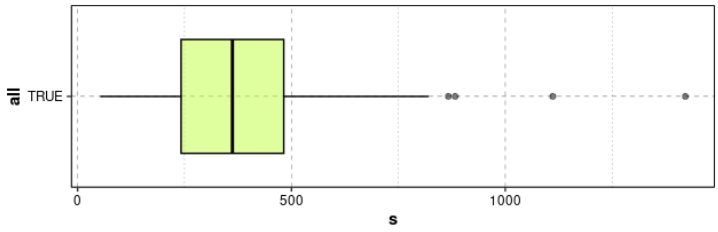


Figura 28: Distribución del número de sesiones inicial.

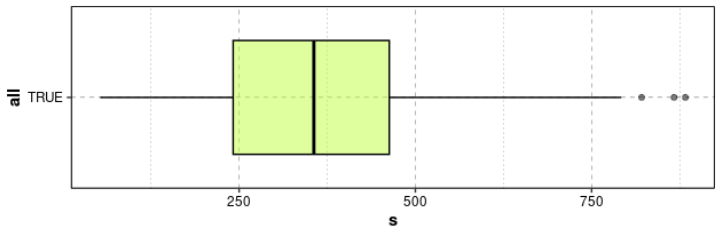


Figura 29: Distribución del número de sesiones tras la eliminación de algunos outliers.

Examinamos ahora los bloques significativos entre ellos agrupando los datos en ocho particiones mediante el algoritmo de las K-medias, tal y como se muestra en la Figura 30. Los resultados obtenidos pueden verse en las Figuras 31a y 31b.

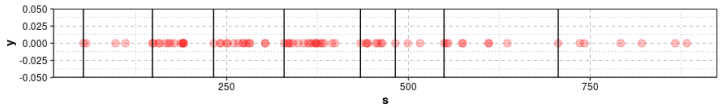
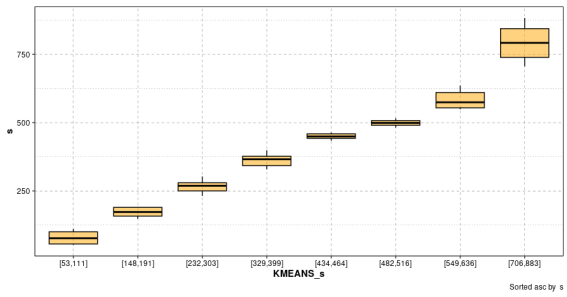
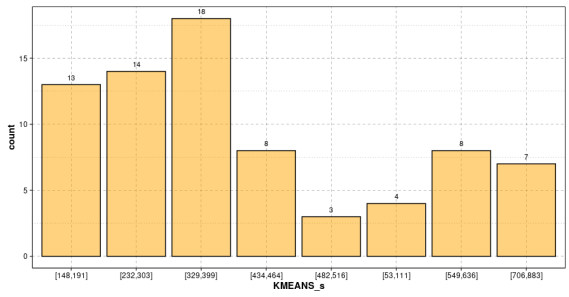


Figura 30: Particiones obtenidas con $K = 8$.



(a) Boxplot de cada una de las particiones.



(b) Número de grupos por partición.

Figura 31: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 8$.

Nótese que como hemos una obtenido una precisión del 0,9795579 % $>$ 95 %, no eliminaremos más outliers.

Así pues, tras la eliminación de los outliers correspondientes tanto al número de sesiones como al número de problemas resueltos como veremos en la subsección 5.4.4 (podemos ver la nueva función de densidad en la Figura 32) se procede a aplicar el test de normalidad de Shapiro-Wilk. Como se obtiene $p - value = 0{,}003307 < 0{,}05$, podemos decir que estadísticamente no sigue una distribución normal. No obstante, teniendo en cuenta que el tamaño de la muestra es relativamente pequeño (tenemos un total de 77 grupos de prácticas), podemos considerar que se trata de una distribución normal.

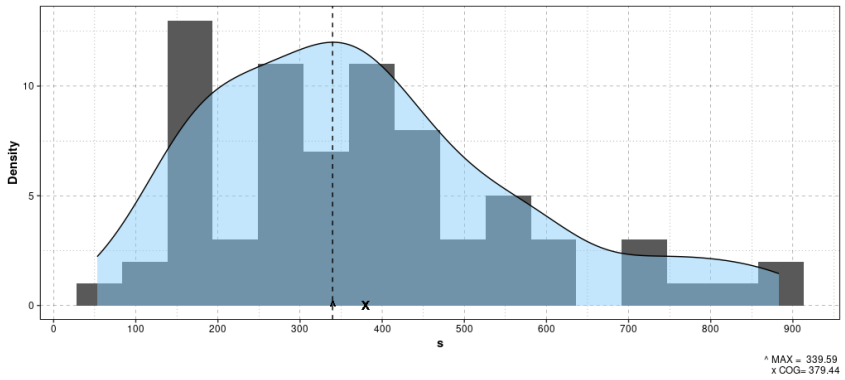


Figura 32: Función de densidad de probabilidad del número de sesiones tras eliminar algunos outliers.

5.4.2 *Sesiones por cada problema*

En la Figura 33 podemos ver el boxplot del número de sesiones por problema. Como podemos ver, el problema P1 es mucho más frecuentado que el resto. No obstante, esto se debe la principal diferencia entre el número de sesiones abiertas del problema P1 y restantes se debe a que los alumnos utilizan el primer problema como base de todos los experimentos y para testear las comunicaciones con el servidor. Así pues, el problema P1 es frecuentemente utilizado, no ya sólo al comienzo, sino durante toda la práctica.

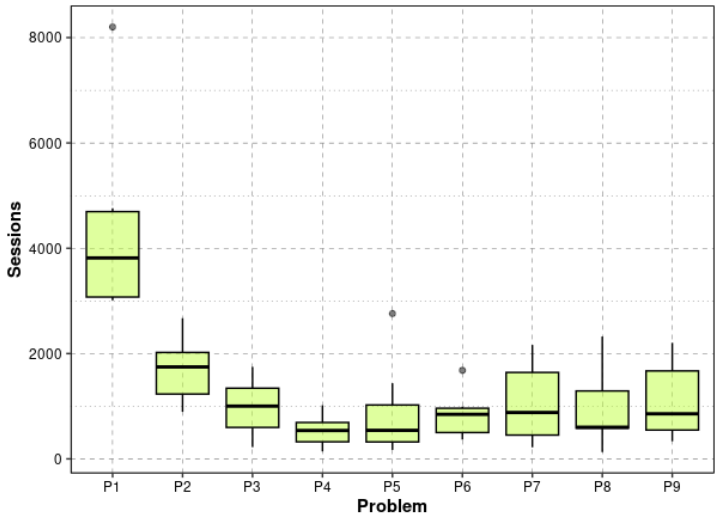


Figura 33: Boxplot del número de sesiones por problema.

5.4.3 *Sesiones cada año*

Como podemos ver en la Figura 34, las sesiones de trabajo abiertas en el servidor año tras año, parecen seguir la misma distribución de probabilidad.

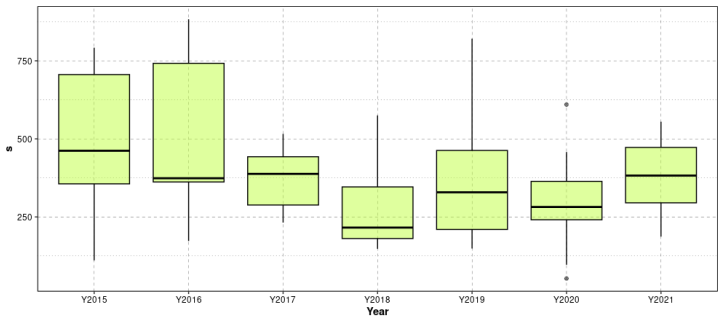


Figura 34: Boxplot del número de sesiones por año tras la eliminación de algunos outliers.

Un resumen de los resultados obtenidos al realizar el test ANOVA se muestra en la Tabla 5. La hipótesis nula establece que el número de sesiones medio de los siete cursos académicos estudiados es el mismo. Así pues, estableciendo un nivel de significancia de 0,05, como tenemos que $p = 0{,}0412 < 0{,}05$, las diferencias entre las medias podrían ser estadísticamente significativas. Aplicando el test de Kruskal-Wallis, obtenemos un $p - value$ igual a $0{,}08798 > 0{,}05$.

Cuadro 5: Resultados del test ANOVA de un solo factor (número de sesiones).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ndsp[[nVariable]]	6	460434.32	76739.05	2.34	0.0412
Residuals	67	2197451.96	32797.79		

Además, se ha realizado un test de Tukey por pares de años (Tabla 6). En él se observa que todos los pares pueden considerarse estadísticamente iguales (p adj $>$ 0,1 en todos ellos). Así pues, podemos concluir que el número de sesiones abiertas en el servidor sigue la misma distribución de probabilidad año tras año.

La Figura 35 muestra los intervalos de confianza de todas las diferencias entre las distintas parejas de años. Así pues, consideraremos que el número de sesiones de cada grupo por año es equivalente (las variaciones son debidas al azar). Esto es importante porque indica que el comportamiento más basico de los alumnos, que viene dado por cuantas veces se conectan, es el mismo año tras año.

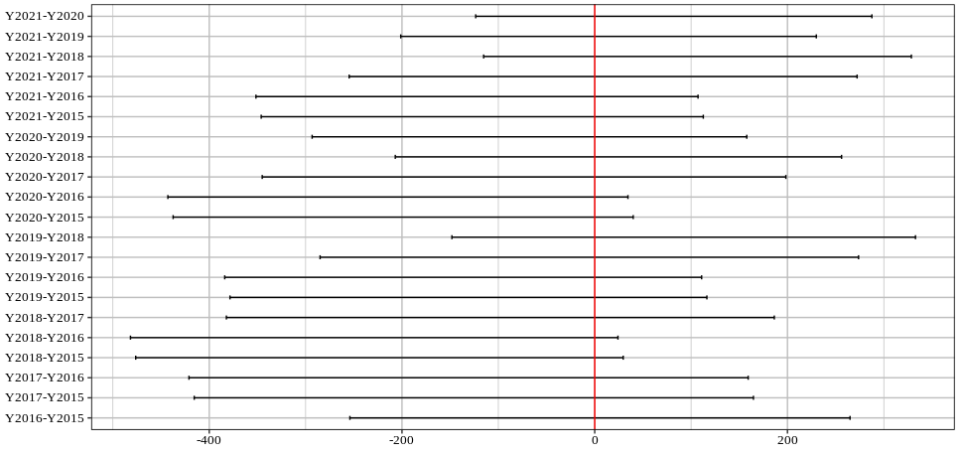


Figura 35: Intervalos de confianza del número de sesiones por años.

5.4.4 *Análisis de la distribución del número de problemas resueltos*

Además, a partir de los registros almacenados en el servidor se calcularán el número de problemas resueltos por cada grupo de prácticas. Como se puede intuir, se tratará de una variable discreta. En la Figura 36 podemos ver que el número de problemas resueltos oscila entre 6 y 9.

Cuadro 6: Test HSD de Tukey (Honestly-significance-difference) del número de sesiones por años.

	diff	lwr	upr	p adj
Y2016-Y2015	5.44	-254.06	264.95	1.00
Y2017-Y2015	-125.44	-415.58	164.69	0.84
Y2018-Y2015	-223.38	-476.31	29.56	0.12
Y2019-Y2015	-131.05	-378.48	116.38	0.68
Y2020-Y2015	-198.78	-437.49	39.93	0.16
Y2021-Y2015	-116.72	-346.09	112.66	0.72
Y2017-Y2016	-130.89	-421.03	159.25	0.81
Y2018-Y2016	-228.82	-481.76	24.11	0.10
Y2019-Y2016	-136.49	-383.92	110.93	0.63
Y2020-Y2016	-204.22	-442.93	34.49	0.14
Y2021-Y2016	-122.16	-351.53	107.21	0.67
Y2018-Y2017	-97.93	-382.21	186.34	0.94
Y2019-Y2017	-5.61	-284.99	273.78	1.00
Y2020-Y2017	-73.33	-345.03	198.36	0.98
Y2021-Y2017	8.73	-254.80	272.26	1.00
Y2019-Y2018	92.33	-148.20	332.86	0.90
Y2020-Y2018	24.60	-206.95	256.15	1.00
Y2021-Y2018	106.66	-115.25	328.57	0.77
Y2020-Y2019	-67.73	-293.25	157.80	0.97
Y2021-Y2019	14.34	-201.28	229.95	1.00
Y2021-Y2020	82.06	-123.49	287.61	0.89

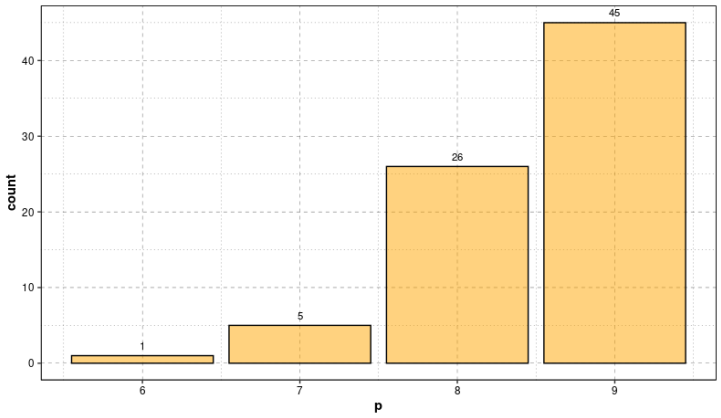


Figura 36: Distribución del número de problemas resueltos.

Además, podemos ver en la Figura 37 que hay un elemento extremo (6) considerando la distribución del número de problemas resueltos por grupo de alumnos. Segmentando por años, obtenemos los boxplots que se muestran en la Figura 38. Así pues, eliminaremos el outlier encontrado puesto que se trata de un valor extremo en todos los años incluidos en este estudio.

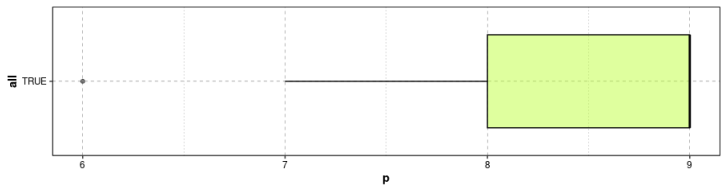


Figura 37: Distribución del número de problemas resueltos inicial.

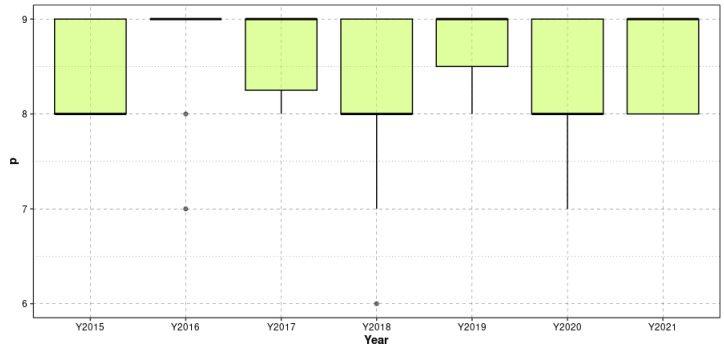


Figura 38: Boxplot del número de problemas resueltos por año.

6

HIPÓTESIS DE ESTUDIO

A pesar de que el estudio descriptivo anterior muestra unos datos muy variados, casi todos ellos son homogéneos año tras año. No obstante, el objetivo de este estudio es sentar las bases para conseguir una experiencia de aprendizaje óptima para todos los grupos de alumnos. Así pues, se va a poner énfasis en detectar a aquellos grupos que estén en riesgo de obtener un peor rendimiento o peores calificaciones. La detección temprana de éstos podría permitir al profesor actuar a tiempo para mejorar su proceso de aprendizaje. Para ello, se van a proponer una serie de métricas de calidad que se definirán sobre los registros de actividad de los alumnos con el objetivo de encontrar aquella que, con mayor certeza, identifique a los alumnos que peor están progresando.

6.1 MÉTRICAS DE CALIDAD Y CORRELACIONES ENTRE ELLAS

Se definirán dos grandes grupos de métricas. El primer grupo consistirá en una colección de métricas de los grupos que solamente podrán calcularse tras la finalización de la práctica. Por el contrario, las métricas del segundo grupo podrán calcularse durante la realización de la práctica y, por tanto, serán más interesantes porque podrán facilitar la detección precoz de los grupos en riesgo.

Gráficamente, las medidas se han representado en la Figura 39.

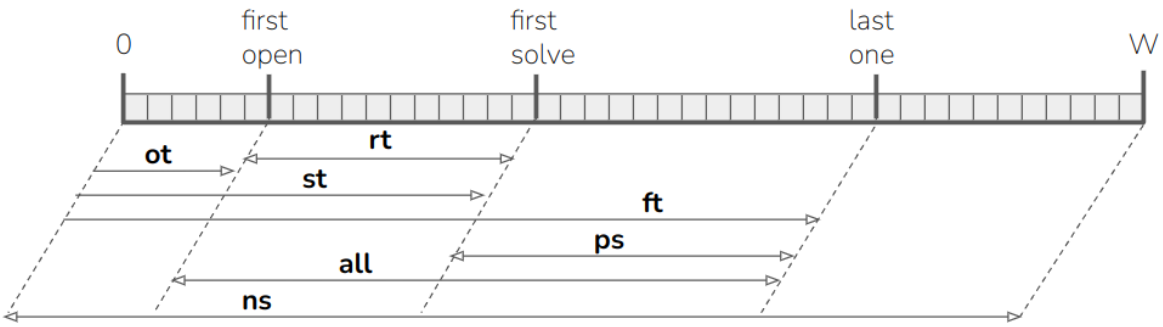


Figura 39: Representación gráfica de las medidas de rendimiento empleadas que se extraen directamente de los registros del servidor (no se incluyen las medidas derivadas del análisis espectral de grafos).

6.1.1 *Medidas a posteriori del resultado de la práctica*

- La calificación conseguida por el alumno (*Grade*). Obviamente, cuanto mayor sea ésta, mejor.
- Número de problemas resueltos u objetivos resueltos. Se denotará por p. Trivialmente, cuantos más objetivos haya resuelto un grupo, mejor. El número de problemas normalizado se denotará por np en los estudios que se realizarán a continuación.
- Punto de finalización de toda la práctica. En la Figura 39 se representa esta medida de rendimiento normalizada por ft. Cuanto antes, mejor (para disponer de más tiempo para repasar y corregir errores). Sin embargo, no es una métrica muy relevante.
- Tiempo consumido por el alumno durante las prácticas. Este es un valor trampa, pues puede significar algo positivo (el alumno ha tardado poco en resolver la práctica porque la domina), o negativo (porque no ha podido dedicarle más tiempo). En la Figura 39, el tiempo consumido normalizado se representa por *all*.
- Número de sesiones realizadas (s). Ya se ha hecho un estudio de esta medida en la sección 5.4. En la Figura 39, se representa el número de sesiones normalizado (ns).

6.1.2 *Medidas continuas durante la práctica*

- Como los comienzos son siempre costosos, se definirá una nueva métrica correspondiente al promedio de tiempo de la primera apertura de cada problema. Se representa por *ot* en la Figura 39.
- Número de fails consecutivos (*fail ratio* o *fr*) hasta resolver un problema dividido entre el numero de sesiones de ese problema. La tasa de fallo así como el tiempo dedicado a un mismo problema dependerá de la dificultad del mismo.

- *También se definirá una medida que cuantificará la posposición de tareas de los grupos. Es decir, pretende identificar a aquellos alumnos que, cuando intentan resolver un problema y no lo consiguen, saltan, curiosamente, a otros problemas más complejos (los cuales, obviamente, tampoco pueden resolver) perdiendo así un tiempo precioso.*
- Tiempo promedio empleado en resolver los problemas (representado por *rt*).
- Tiempo promedio empleado en los problemas tras su resolución. Se representa por *ps* en la Figura 39 y trata de evidenciar el interés de los alumnos en la materia. En general, será positivo que los grupos no sólo resuelvan los problemas sino que traten de encontrar mejores soluciones para los mismos como se podrá ver en el Capítulo 11.
- Tiempo de promedio de resolución de los problemas por primera vez. Se representará por *st* en la Figura 39.
- Se tendrá igualmente una medida que refleja si los grupos de prácticas siguen el orden esperado de las mismas. Se le denotará por *sq* en los análisis que se realizarán posteriormente.
- El coeficiente *DAG*, que cuantificará como de balanceado están los grafos que describen la actividad de un grupo en la plataforma. Se trata una medida de rendimiento basada en el análisis espectral de grafos que se verá más adelante.
- El coeficiente de *Laplace*.

7

RENDIMIENTO OBSERVADO DE LOS ALUMNOS

7.1 CALIFICACIONES OBTENIDAS (GRADE)

La primera medida de rendimiento que tendremos en cuenta son las calificaciones de los grupos de prácticas. Hay que tener en cuenta que estas calificaciones no son la evaluación final de cada grupo, sino la nota de la práctica cuya evolución se está analizando. Esta es la parte más subjetiva de la evaluación del rendimiento de cada alumno pues implica la participación del profesor y la toma en consideración de otros factores, además de los registrados en el servidor como puede ser la calidad de la memoria de la práctica. Las calificaciones muestran una distribución normal a lo largo de estos siete años de registros, ligeramente inclinada a la derecha porque las notas medias de esta asignatura suelen ser altas tal y como puede verse en la Figura 40.

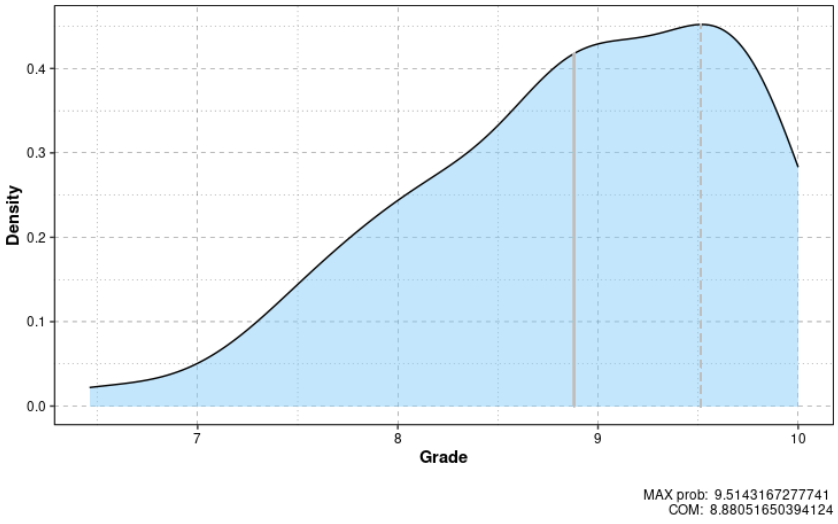
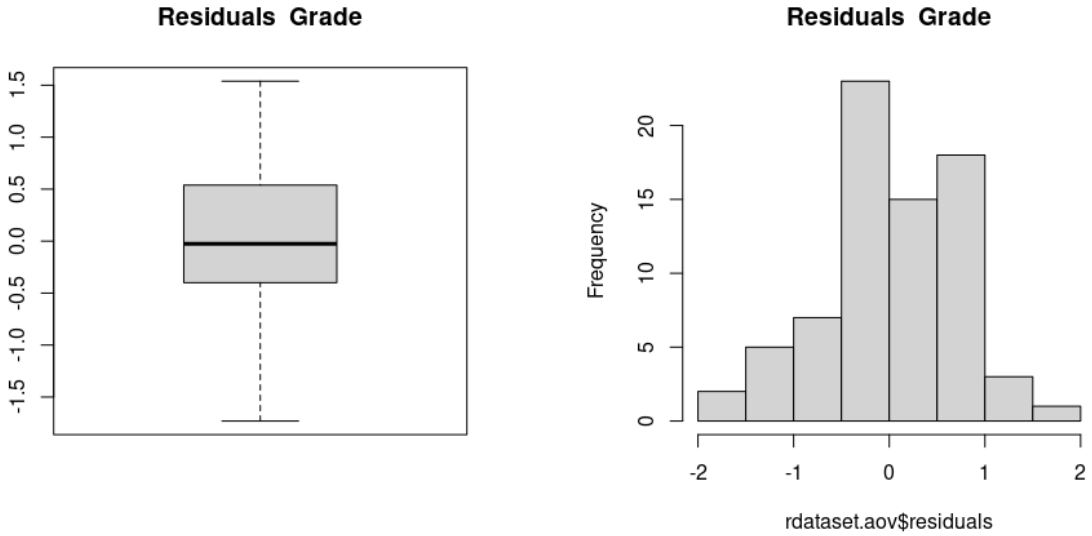


Figura 40: Función de densidad de probabilidad de las calificaciones obtenidas por los distintos grupos de prácticas.

Además, la media está bastante balanceada (Figuras 41a y 41b) y no se detecta la presencia de outliers (Figura 42).



(a) Boxplot de los residuos de las calificaciones.

(b) Histograma de los residuos de las calificaciones.

Figura 41: Distribución de los residuos de las calificaciones.

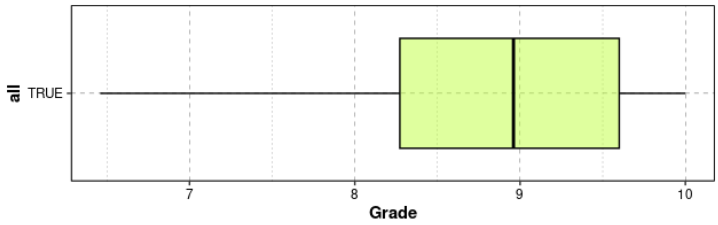


Figura 42: Distribución de las calificaciones obtenidas por los distintos grupos de alumnos inicial.

Más aún, la regresión es muy aceptable (Figura 43).

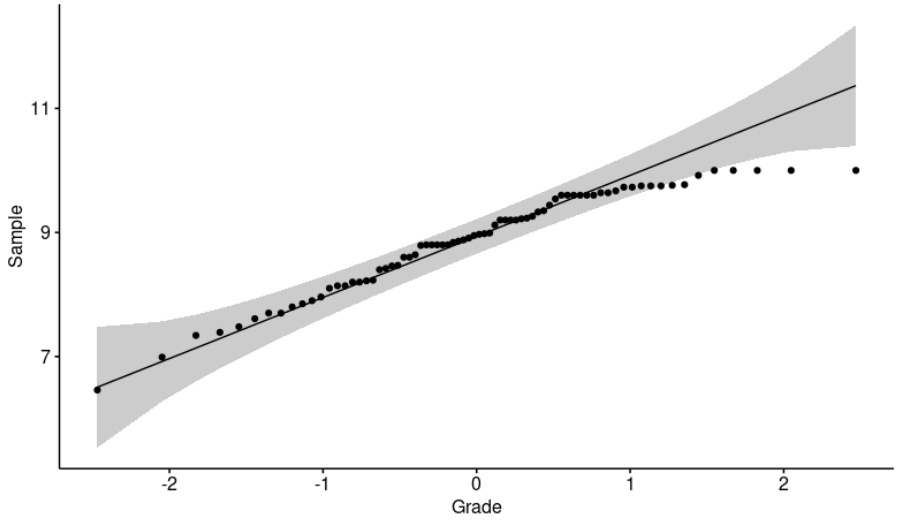


Figura 43: Gráfico Q-Q de las calificaciones.

A continuación se realizará un estudio por años de las calificaciones obtenidas. El boxplot de las calificaciones por años puede verse en la Figura 44. Como puede verse, los datos recogidos muestran una variación muy perceptible en las notas a lo largo de los años. De hecho, los test estadísticos ratifican que hay diferentes significativas entre ellas. Tras realizar el test ANOVA de un factor (resultados en la Tabla 7) obtenemos $p = 0{,}00143 < 0{,}05$. Además, tras realizar el test de Kruskal-Wallis se obtiene $p - value = 0{,}01534$. Un análisis posterior por pares de Tukey muestra las diferencias entre los años (resultados en la Tabla 8). Podemos ver que el año 2017 es el principal elemento de disrupción, pero por poco margen. Lo mismo indica el análisis de los intervalos de confianza de la Figura 45.

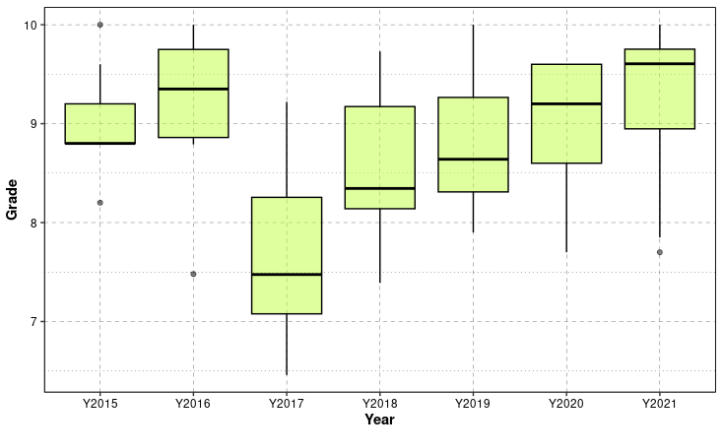


Figura 44: Boxplot de las calificaciones por año.

Cuadro 7: Resultados del test ANOVA de un solo factor (calificaciones).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rdataset[[Variable]]	6	13.05	2.18	4.11	0.0014
Residuals	67	35.47	0.53		

Cuadro 8: Test HSD de Tukey (Honestly-significance-difference) de las calificaciones por años.

	diff	lwr	upr	p adj
Y2016-Y2015	0.18	-0.87	1.22	1.00
Y2017-Y2015	-1.35	-2.52	-0.19	0.01
Y2018-Y2015	-0.45	-1.47	0.57	0.83
Y2019-Y2015	-0.24	-1.23	0.76	0.99
Y2020-Y2015	-0.10	-1.06	0.85	1.00
Y2021-Y2015	0.22	-0.70	1.14	0.99
Y2017-Y2016	-1.53	-2.70	-0.36	0.00
Y2018-Y2016	-0.63	-1.64	0.39	0.51
Y2019-Y2016	-0.41	-1.41	0.58	0.87
Y2020-Y2016	-0.28	-1.24	0.68	0.97
Y2021-Y2016	0.05	-0.88	0.97	1.00
Y2018-Y2017	0.90	-0.24	2.05	0.21
Y2019-Y2017	1.12	-0.00	2.24	0.05
Y2020-Y2017	1.25	0.16	2.34	0.01
Y2021-Y2017	1.57	0.52	2.63	0.00
Y2019-Y2018	0.21	-0.75	1.18	0.99
Y2020-Y2018	0.34	-0.59	1.28	0.92
Y2021-Y2018	0.67	-0.22	1.56	0.27
Y2020-Y2019	0.13	-0.78	1.04	1.00
Y2021-Y2019	0.46	-0.41	1.32	0.68
Y2021-Y2020	0.33	-0.50	1.15	0.89

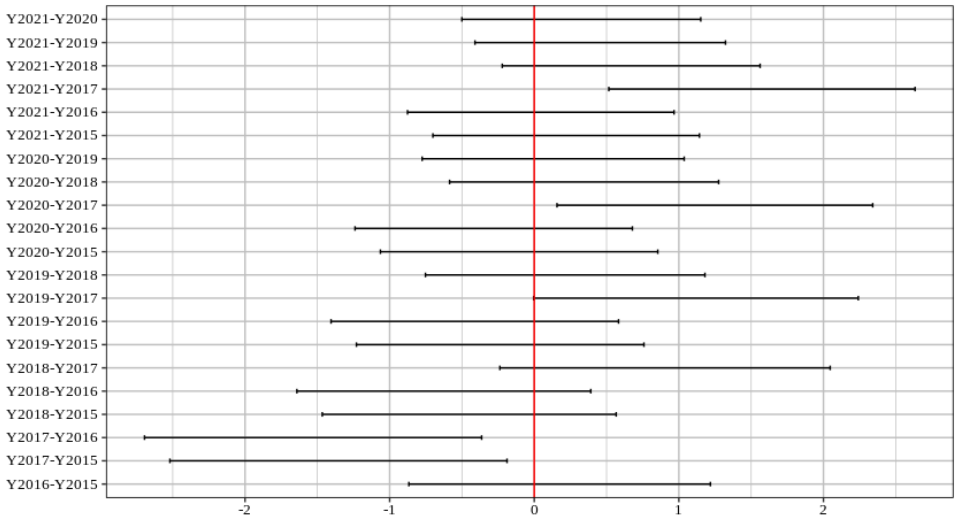


Figura 45: Intervalos de confianza de las calificaciones por años.

7.2 NÚMERO TOTAL DE PROBLEMAS RESUELTOS (P)

Podemos observar que ha habido variaciones perceptibles durante los distintos años, con un caso especial en 2018 en el que hubo muchos grupos que no resolvieron todos los problemas (Figura 46).

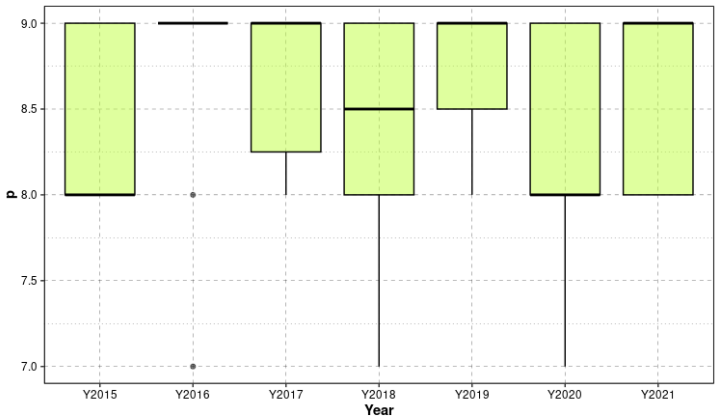


Figura 46: Boxplot del número de problemas resueltos por año.

Sin embargo, las diferencias entre las medias no son estadísticamente significativas (considerando un nivel de significancia de 0,05) tal y como puede verse en la Tabla 9. Además, si consideramos el test estadístico de Kruskal-Wallis llegamos a la misma conclusión ($p - value = 0{,}365 > 0{,}05$). Así pues, se concluye que el número de problemas resueltos por años son uniformes (cualquier variación es debida al azar).

Además, realizando un test de Tukey por pares de años (Tabla 10) se observa que todos los pares pueden considerarse estadísticamente iguales. La Figura 47 muestra los intervalos de confianza de todas las diferencias entre las distintas parejas de años.

Cuadro 9: Resultados del test ANOVA de un solo factor (número de problemas resueltos).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rdataset[[Variable]]	6	2.91	0.48	1.27	0.2835
Residuals	67	25.58	0.38		

Cuadro 10: Test HSD de Tukey (Honestly-significance-difference) del número de problemas resueltos por año.

	diff	lwr	upr	p adj
Y2016-Y2015	0.22	-0.66	1.11	0.99
Y2017-Y2015	0.22	-0.77	1.21	0.99
Y2018-Y2015	-0.04	-0.91	0.82	1.00
Y2019-Y2015	0.28	-0.56	1.13	0.95
Y2020-Y2015	-0.29	-1.11	0.52	0.93
Y2021-Y2015	0.18	-0.60	0.96	0.99
Y2017-Y2016	0.00	-0.99	0.99	1.00
Y2018-Y2016	-0.27	-1.13	0.60	0.96
Y2019-Y2016	0.06	-0.78	0.90	1.00
Y2020-Y2016	-0.51	-1.33	0.30	0.48
Y2021-Y2016	-0.04	-0.82	0.74	1.00
Y2018-Y2017	-0.27	-1.24	0.70	0.98
Y2019-Y2017	0.06	-0.89	1.01	1.00
Y2020-Y2017	-0.51	-1.44	0.41	0.63
Y2021-Y2017	-0.04	-0.94	0.86	1.00
Y2019-Y2018	0.33	-0.49	1.15	0.89
Y2020-Y2018	-0.25	-1.04	0.54	0.96
Y2021-Y2018	0.22	-0.53	0.98	0.97
Y2020-Y2019	-0.57	-1.34	0.20	0.28
Y2021-Y2019	-0.10	-0.84	0.63	1.00
Y2021-Y2020	0.47	-0.23	1.17	0.40

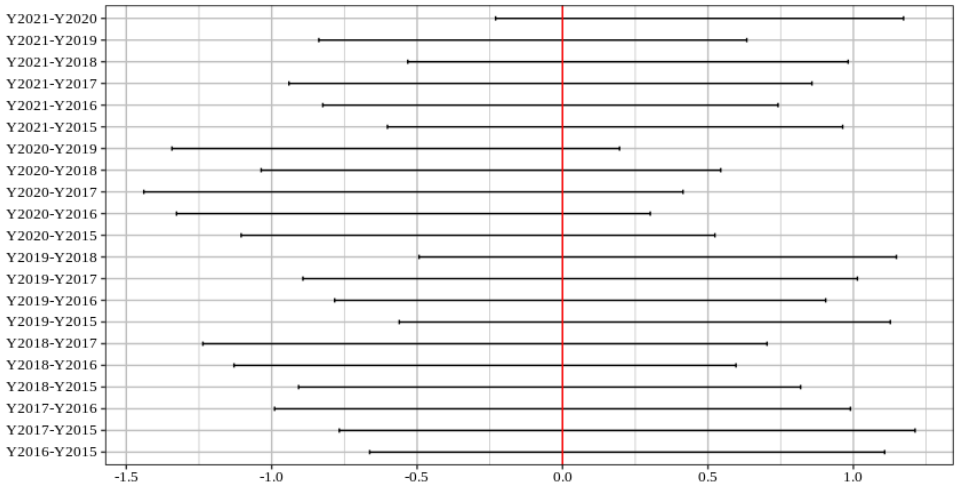
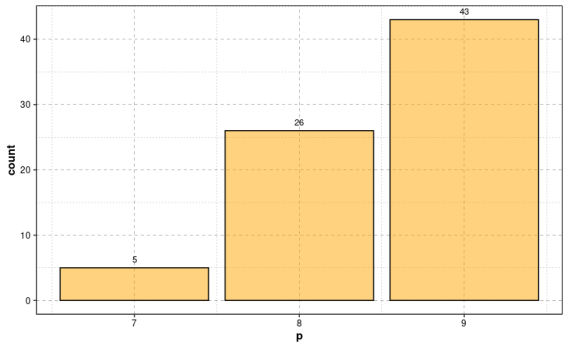
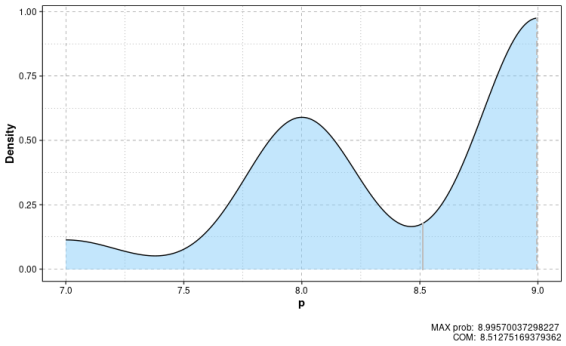


Figura 47: Intervalos de confianza del número de problemas resueltos por año.

Por último, podemos concluir que la variable número de problemas resueltos por año (p) no es muy normal en tanto que varía muy poco y es discreta (Figuras 48a y 48b).



(a) Número de grupos que han resuelto una determinada cantidad de problemas.



(b) Función de densidad de probabilidad del número de problemas resueltos.

Figura 48: Distribución del número de problemas resueltos por cada grupo de alumnos.

7.3 FINALIZAR LA PRÁCTICA (FT)

Como vemos en la Figura 49, el punto de finalización de la práctica normalizado tiende a estar cerca del final de la misma.

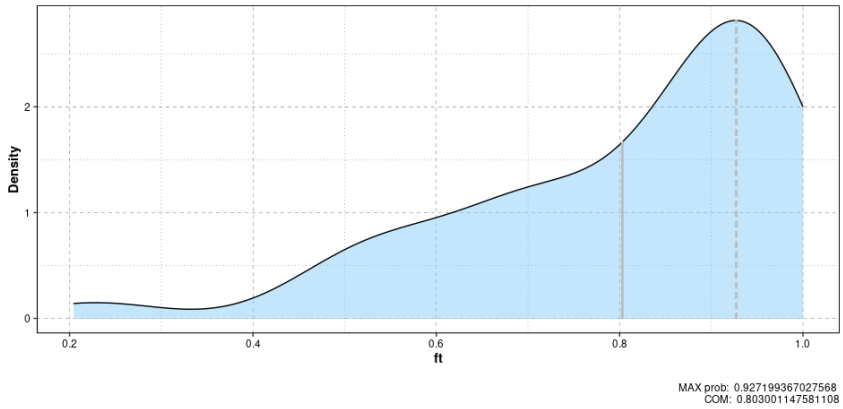


Figura 49: Función de densidad de probabilidad del momento en el que los distintos grupos de prácticas finalizan la práctica.

7.4 NÚMERO DE SESIONES REALIAZADAS (S)

El número de sesiones realizadas por los diferentes grupos de prácticas se ha analizado con anterioridad. En la subsección 5.4.1 vimos que seguía una distribución casi-normal. Además, vimos que el número de sesiones sigue la misma distribución de probabilidad en cada uno de los cursos académicos considerados (subsección 5.4.3).

7.5 NÚMERO DE INTENTOS PARA RESOLVER CADA PROBLEMA (SESSIONSBEFORE)

Número de veces que se abre un mismo problema sin resolver hasta que se resuelve por primera vez. Este valor está directamente relacionado con la tasa de fallo analizada anteriormente. Podemos ver que esta medida de rendimiento sigue una distribución casi normal, ligeramente inclinada a la izquierda tal y como puede verse en la Figura 50.

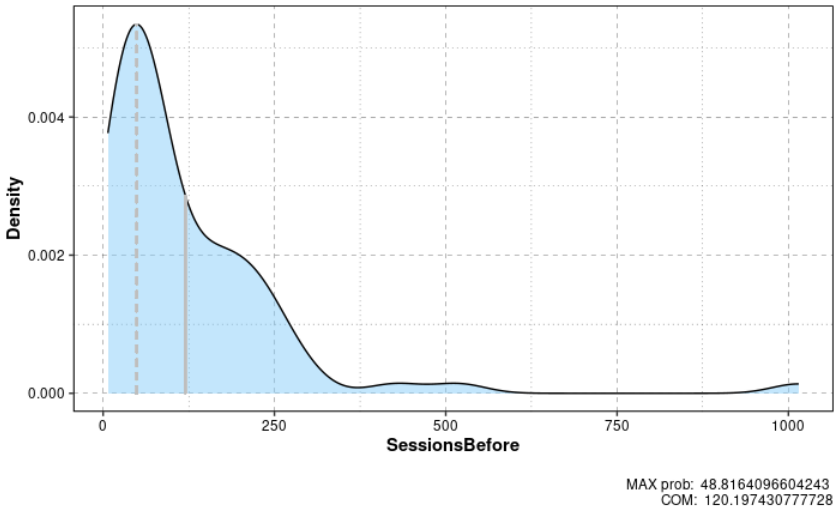
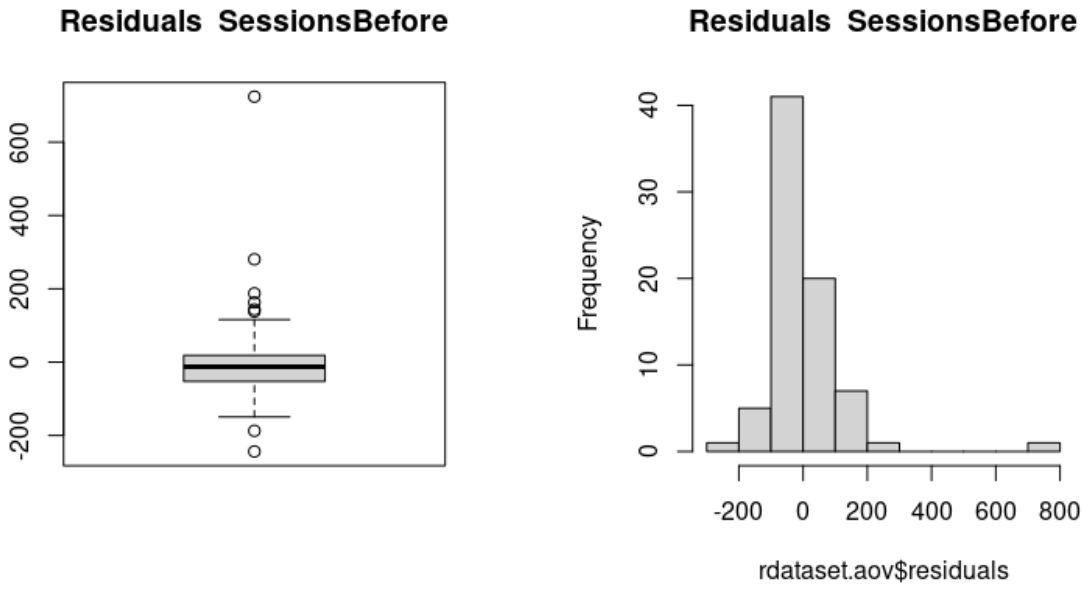


Figura 50: Función de densidad de probabilidad del número de intentos necesarios para resolver un problema por los distintos grupos de prácticas.

Sin embargo, podemos ver algunos grupos que emplean muy pocos intentos para resolver un problema o que, por el contrario, emplean más intentos que los demás (Figuras 51a y 51b).



(a) Boxplot de los residuos del número de intentos.

(b) Histograma de los residuos del número de intentos.

Figura 51: Distribución de los residuos del número de intentos.

No obstante, podemos considerar que la regresión es aceptable (Figura 52).

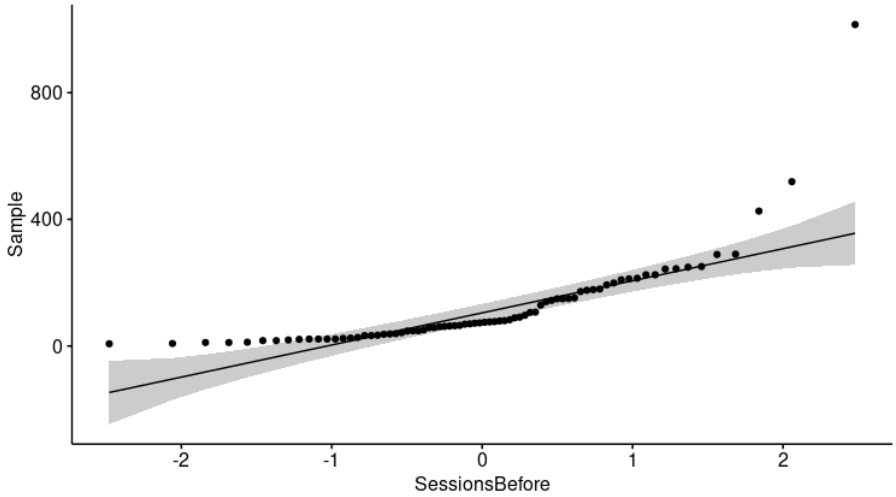


Figura 52: Gráfico Q-Q del número de intentos.

A continuación, se realizará un estudio por años del número de intentos antes de resolver un problema realizados. Como podemos observar en la Figura 53, hay una variación muy perceptible en el número de intentos realizados a lo largo de los años. La realización del test ANOVA de un factor (resultados en la Tabla 11) ratifica que hay diferencias significativas ($p = 0{,}0001 < 0{,}05$). Posteriormente, se ha realizado el test de Tukey por pares y se ha visto que hay diferencias entre las distribuciones de los años 2015 y 2017 son las que introducen diferencias (Tabla 12). Esto mismo puede observarse en la Figura 54.

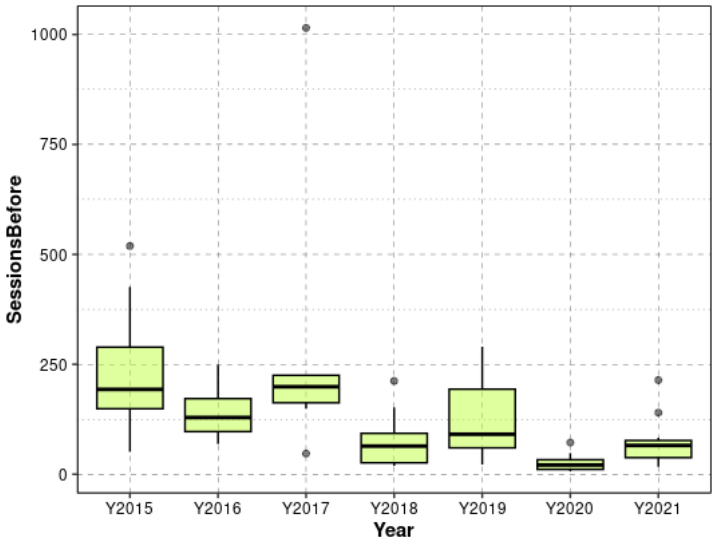


Figura 53: Boxplot del número de intentos para resolver un problema por año.

Cuadro 11: Resultados del test ANOVA de un solo factor (calificaciones).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rdataset[[Variable]]	6	516318.38	86053.06	5.83	0.0001
Residuals	69	1018845.04	14765.87		

Cuadro 12: Test HSD de Tukey (Honestly-significance-difference) del número de intentos realizados para resolver un problema por años.

	diff	lwr	upr	p adj
Y2016-Y2015	-95.44	-269.41	78.52	0.64
Y2017-Y2015	52.30	-133.68	238.28	0.98
Y2018-Y2015	-164.56	-330.43	1.32	0.05
Y2019-Y2015	-111.37	-277.25	54.50	0.40
Y2020-Y2015	-213.56	-373.58	-53.53	0.00
Y2021-Y2015	-168.24	-322.01	-14.48	0.02
Y2017-Y2016	147.75	-38.23	333.73	0.21
Y2018-Y2016	-69.11	-234.98	96.76	0.87
Y2019-Y2016	-15.93	-181.80	149.94	1.00
Y2020-Y2016	-118.11	-278.14	41.92	0.29
Y2021-Y2016	-72.80	-226.57	80.97	0.78
Y2018-Y2017	-216.86	-395.29	-38.43	0.01
Y2019-Y2017	-163.68	-342.10	14.75	0.09
Y2020-Y2017	-265.86	-438.87	-92.85	0.00
Y2021-Y2017	-220.54	-387.78	-53.31	0.00
Y2019-Y2018	53.18	-104.18	210.54	0.95
Y2020-Y2018	-49.00	-200.19	102.19	0.96
Y2021-Y2018	-3.69	-148.23	140.86	1.00
Y2020-Y2019	-102.18	-253.37	49.00	0.39
Y2021-Y2019	-56.87	-201.41	87.68	0.89
Y2021-Y2020	45.31	-92.49	183.11	0.95

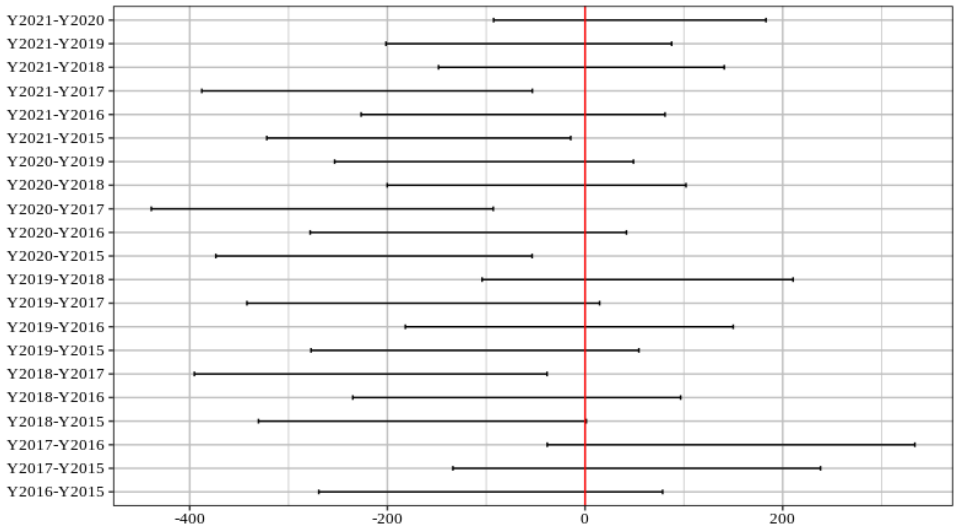


Figura 54: Intervalos de confianza del número de intentos antes de resolver un problema por años.

7.6 SESIONES PERDIDAS DURANTE UN PROBLEMA

Con frecuencia ocurre que los alumnos, cuando intentan resolver un problema y no lo consiguen, saltan, curiosamente, a sesiones en otros problemas más complejos, los cuales, obviamente, tampoco pueden resolver.

Falta gráfica.

Este hábito es más frecuente de lo que parece y mantiene mucha variación en cada problema, sobre todo es más fuerte al comienzo de las prácticas, pero es homogéneo año tras año (ANOVA p=0.746, KW p=0.9) y se puede decir que está presente en la mayoría de los grupos. Aunque la mediana sea 0, casi todos los grupos (80 %) exhiben este comportamiento en algún momento.

Referenciar tabla ya existente.

Falta gráfica.

7.7 ABRIR UN PROBLEMA POR PRIMERA VEZ (OT)

Momento exacto en el que se consigue abrir cada problema por primera vez en el servidor, normalizado para poder compararlo (normalizado porque la duración de la práctica no es la misma todos los años). Como podemos ver en su función de densidad (Figura 55), se aproxima a una distribución normal, un poco ladeada hacia la derecha. Esto puede deberse a

que los distintos equipos tienen a abrir los problemas cuando se va aproximando la fecha de entrega de la práctica.

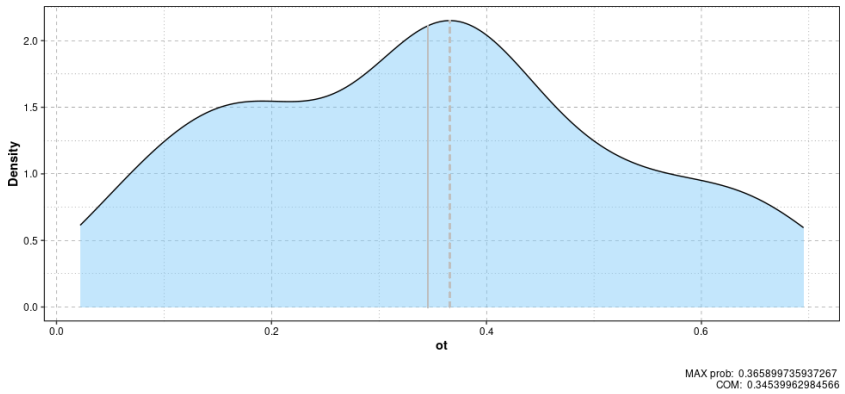
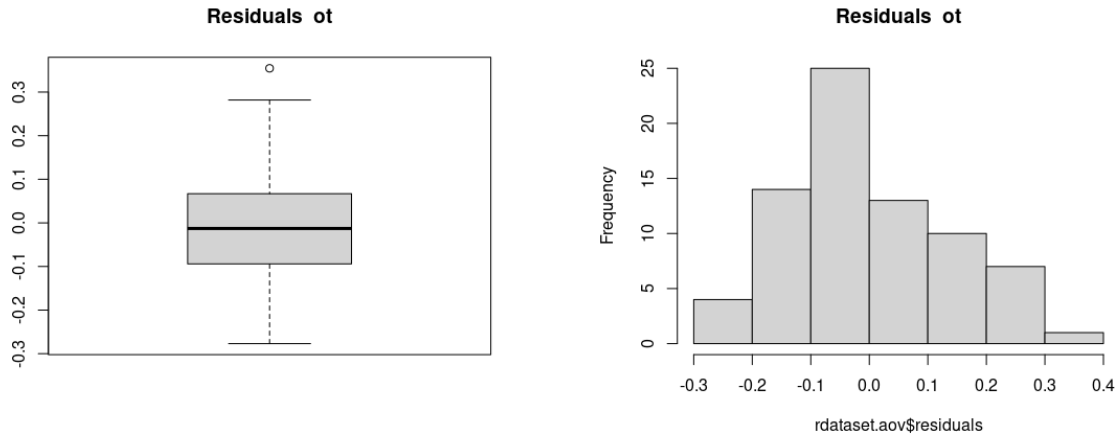


Figura 55: Función de densidad de probabilidad del momento en el que los distintos grupos de prácticas abren por primera vez un problema.

En las Figuras 56a y 56b pueden verse el boxplot de los residuos y el histograma de los mismos respectivamente.



(a) Boxplot de los residuos del número de intentos.

(b) Histograma de los residuos del momento en el que se abren los problemas por primera vez.

Figura 56: Distribución de los residuos del momento en el que se abren los problemas por primera vez.

Si realizamos una segmentación por años, podemos ver que la media de esta medida de rendimiento puede variar según el año que estemos considerando (Figura 57). El Test ANOVA de un sólo factor ha confirma lo observado (se ha obtenido $p = 4{,}05e - 06 < 0{,}05$ como puede

verse en la Tabla 13). Igualmente, el test de Kruskal-Wallis coincide con lo anteriormente visto ($p - value = 4{,}633e - 05 < 0{,}05$).

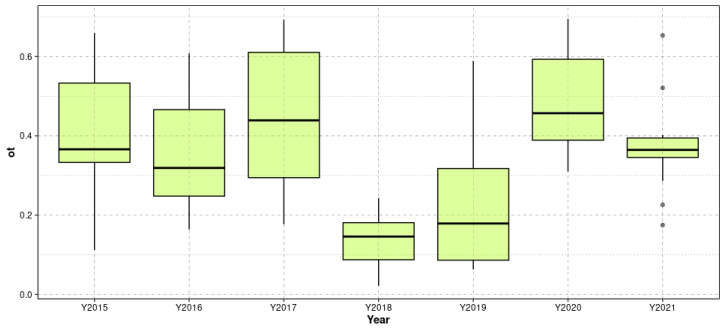


Figura 57: Boxplot del número del momento en el que se abre un problema por primera vez por años.

Cuadro 13: Resultados del test ANOVA de un solo factor (momento en el que se abre un problema por primera vez).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rdataset[[Variable]]	6	0.93	0.15	7.43	4.05e-06
Residuals	67	1.40	0.02		

Tras realizar el Test de Tukey por pares se ha visto que claramente hay años con distribuciones diferentes de esta variable (Tabla 14 y Figura 58).

Cuadro 14: Test HSD de Tukey (Honestly-significance-difference) del momento exacto en el que se abre un problema por primera por años.

	diff	lwr	upr	p adj
Y2016-Y2015	-0.02	-0.23	0.18	1.00
Y2017-Y2015	0.05	-0.18	0.29	0.99
Y2018-Y2015	-0.26	-0.46	-0.06	0.00
Y2019-Y2015	-0.15	-0.35	0.04	0.22
Y2020-Y2015	0.09	-0.10	0.28	0.74
Y2021-Y2015	-0.02	-0.20	0.17	1.00
Y2017-Y2016	0.08	-0.15	0.31	0.94
Y2018-Y2016	-0.23	-0.44	-0.03	0.01
Y2019-Y2016	-0.13	-0.33	0.07	0.42
Y2020-Y2016	0.12	-0.07	0.31	0.50
Y2021-Y2016	0.01	-0.18	0.19	1.00
Y2018-Y2017	-0.31	-0.54	-0.09	0.00
Y2019-Y2017	-0.21	-0.43	0.01	0.08
Y2020-Y2017	0.04	-0.18	0.26	1.00
Y2021-Y2017	-0.07	-0.28	0.14	0.94
Y2019-Y2018	0.10	-0.09	0.29	0.66
Y2020-Y2018	0.35	0.17	0.54	0.00
Y2021-Y2018	0.24	0.06	0.42	0.00
Y2020-Y2019	0.25	0.07	0.43	0.00
Y2021-Y2019	0.14	-0.04	0.31	0.21
Y2021-Y2020	-0.11	-0.28	0.05	0.38

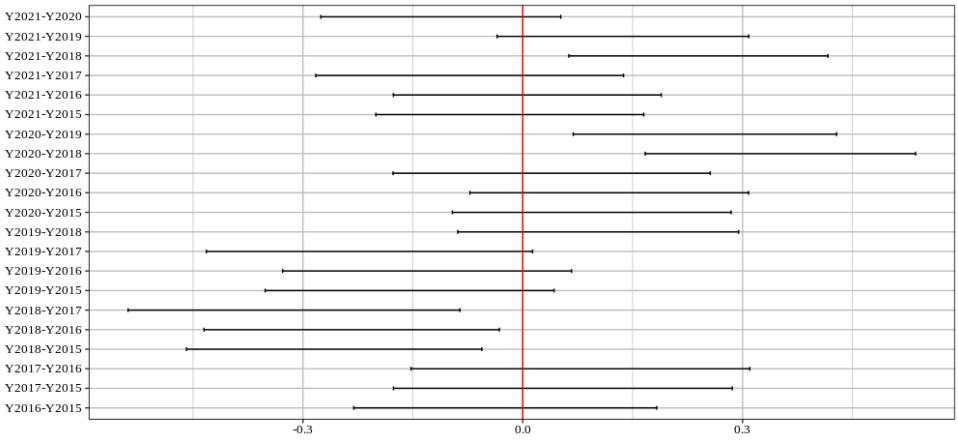


Figura 58: Intervalos de confianza del momento en el que se abre un problema por años.

7.8 TASA DE FALLO (FR)

Se ha estudiado con anterioridad en la sección .

7.9 TIEMPO EMPELADO EN LA RESOLUCIÓN DE UN PROBLEMA POR PRIMERA VEZ (RT)

Como podemos ver en la Figura 59, los grupos suelen emplear poco tiempo en resolver un problema por primera vez.

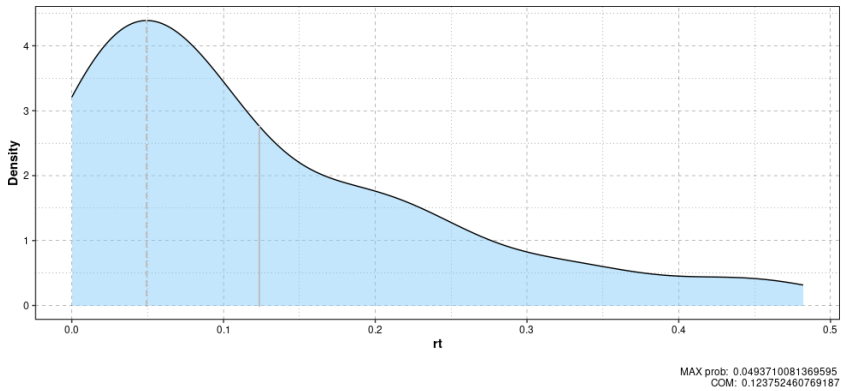


Figura 59: Función de densidad de probabilidad del tiempo que los distintos grupos de prácticas dedican a la resolución de problemas por primera vez.

7.10 EXPLORACIÓN DE NUEVAS VÍAS Y MEJORAS (PS)

Curiosamente, algunos grupos siguen trabajando en un problema que ya han resuelto (Figura 60).

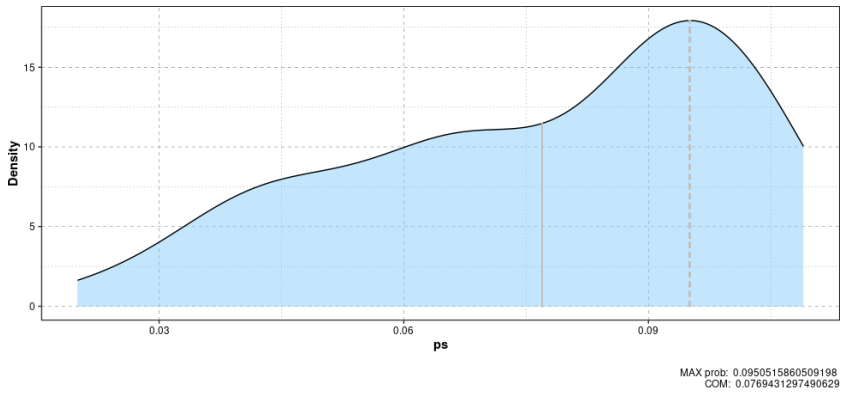


Figura 60: Función de densidad de probabilidad del tiempo que los distintos grupos de prácticas emplean en un problema tras su resolución.

7.11 RESOLVER UN PROBLEMA POR PRIMERA VEZ (ST)

Momento exacto en el que se consigue resolver cada problema por primera vez, normalizado para poder compararlo (porque la duración de la práctica no es la misma todos los años). Como podemos ver en la Figura 61, sigue una distribución casi normal, ladeada hacia la derecha ya que los distintos grupos tienden a resolver los problemas por primera vez al final de la práctica.

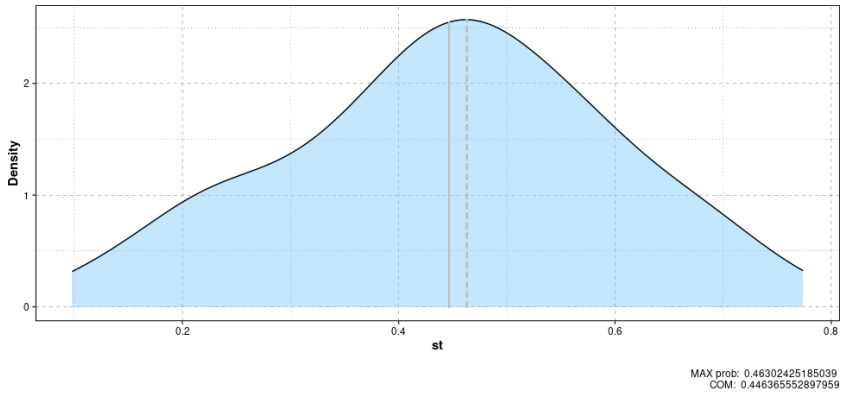
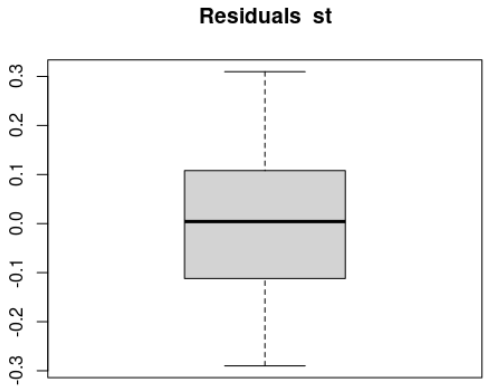
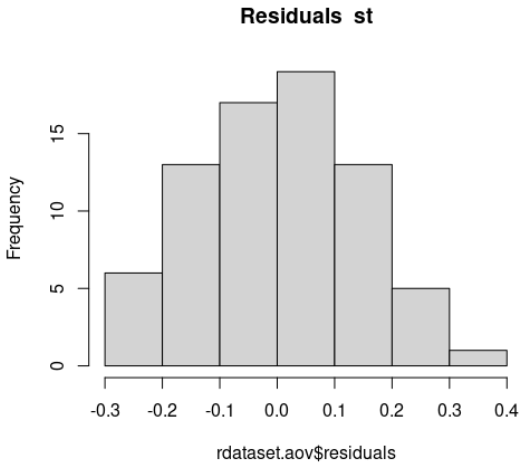


Figura 61: Función de densidad de probabilidad del momento en el que los distintos grupos de prácticas resuelven por primera vez un problema.

Sin embargo, podemos observar la presencia de algunos outliers (Figuras 62a y 62b).



(a) Boxplot de los residuos del número de intentos.



(b) Histograma de los residuos del momento en el que se resuelven los problemas por primera vez.

Figura 62: Distribución de los residuos del momento en el que se resuelven los problemas por primera vez.

Si se realiza una segmentación por años, intuimos que esta medida de rendimiento sigue la misma distribución de probabilidad todos los cursos académicos estudiados (Figura 63). Esto se confirma tras la realización del Test ANOVA de un solo factor (Tabla 15) en el que obtenemos $p = 0{,}1387 > 0{,}05$. Realizando el Test de Kruskal-Wallis llegamos a la misma conclusión ($p - value = 0{,}2627$). Además, el Test de Tukey muestra que no hay diferencias entre los años considerados (Tabla 16).

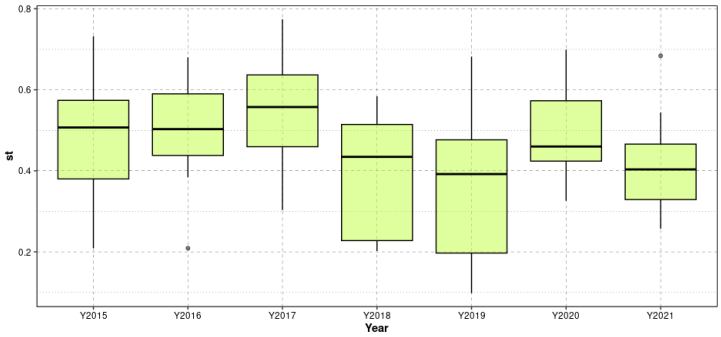


Figura 63: Boxplot del momento en el que se resuelve un problema por primera vez por años.

Cuadro 15: Resultados del test ANOVA de un solo factor (momento en el que se resuelve un problema por primera vez).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rdataset[[Variable]]	6	0.22	0.04	1.68	0.1387
Residuals	67	1.46	0.02		

Cuadro 16: Test HSD de Tukey (Honestly-significance-difference) del momento en el que se resuelve un problema por primera vez, por años.

	diff	lwr	upr	p adj
Y2016-Y2015	0.02	-0.19	0.23	1.00
Y2017-Y2015	0.07	-0.17	0.31	0.97
Y2018-Y2015	-0.09	-0.29	0.12	0.86
Y2019-Y2015	-0.11	-0.31	0.10	0.69
Y2020-Y2015	0.01	-0.19	0.20	1.00
Y2021-Y2015	-0.06	-0.25	0.13	0.95
Y2017-Y2016	0.05	-0.19	0.28	1.00
Y2018-Y2016	-0.11	-0.31	0.10	0.69
Y2019-Y2016	-0.13	-0.33	0.08	0.48
Y2020-Y2016	-0.01	-0.21	0.18	1.00
Y2021-Y2016	-0.08	-0.27	0.10	0.82
Y2018-Y2017	-0.16	-0.39	0.08	0.40
Y2019-Y2017	-0.17	-0.40	0.05	0.25
Y2020-Y2017	-0.06	-0.28	0.16	0.98
Y2021-Y2017	-0.13	-0.35	0.08	0.51
Y2019-Y2018	-0.02	-0.21	0.18	1.00
Y2020-Y2018	0.09	-0.09	0.28	0.73
Y2021-Y2018	0.02	-0.16	0.21	1.00
Y2020-Y2019	0.11	-0.07	0.30	0.51
Y2021-Y2019	0.04	-0.13	0.22	0.99
Y2021-Y2020	-0.07	-0.24	0.10	0.86

Por último, el análisis de los intervalos de confianza se muestra en la Figura 64.

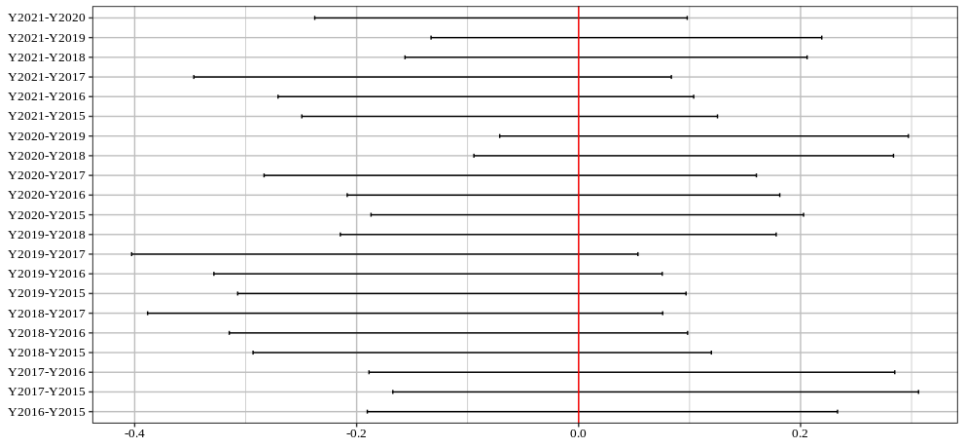


Figura 64: Intervalos de confianza del momento en el que se resuelve un problema por primera vez, por años.

7.12 SIGUIENDO EL PLAN DEL PROFESOR (SQ)

Se incorpora una medida de similaridad (*sq*) que cuantifica cómo se parece el patrón encontrado con respecto al patrón esperado. En un principio, el profesor espera que los problemas se resuelvan en orden de dificultad creciente (`P1 P2 P3 P4 P5 P6 P7 P8 P9`). No obstante, tras el análisis detallado de los registros almacenados, encontramos que muchos grupos optan por resolver los problemas siguiendo un orden diferente al propuesto (por ejemplo, el orden `P1 P3 P4 P5 P7 P2 P6 P8 P9`).

En la Figura 65 podemos ver la función de densidad de probabilidad de la medida de rendimiento que estamos considerando.

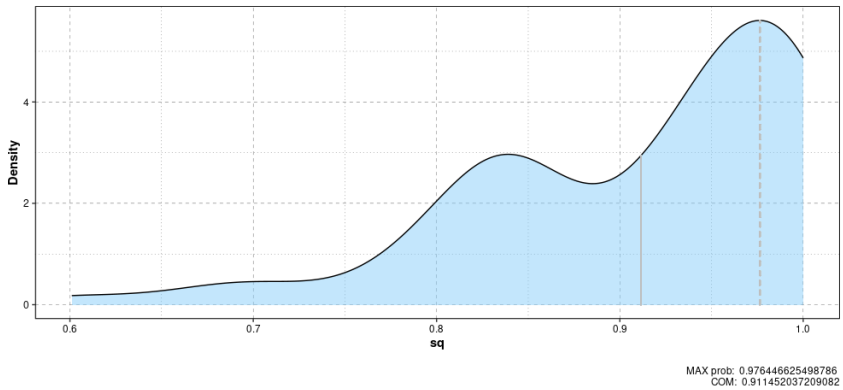


Figura 65: Función de densidad de probabilidad del seguimiento del orden de resolución previsto por el profesor que realizan los diferentes grupos de prácticas.

8

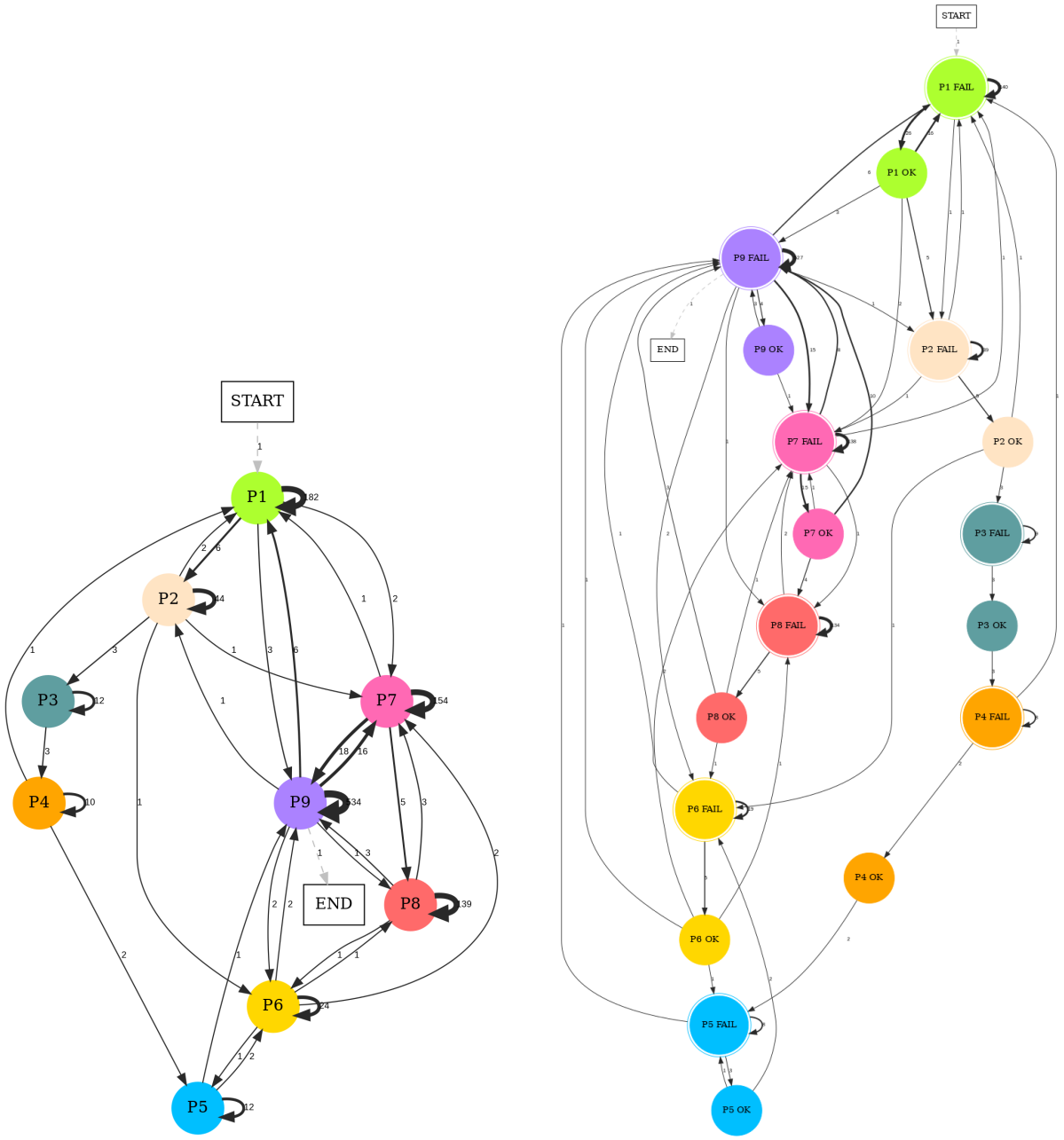
CARACTERÍSTICAS TOPOLÓGICAS DE LOS GRAFOS DE PROCESOS

Teniendo en mente que la finalidad última de este trabajo es encontrar indicadores del progreso de los alumnos y la detección precoz de aquellos grupos con problemas para que el profesor pueda proporcionarles ayuda y orientación, se han obtenido medidas de rendimiento sobre los procesos minados. Se tratarán de funciones *off-the-shelf* genéricas sobre grafos, lo que permitiría aplicar los resulados aquí obtenidos en otras plataformas educativas.

Así pues, se han definido dos métricas distintas: el Laplaciano (*Laplacian*), que usará el análisis espectral de grafos, y la heurística *DAG*, que trata de determinar cómo de balanceados están los nodos de un grafo dirigido acíclico.

8.0.1 *El Laplaciano (*Laplacian*)*

En primer lugar, empezamos analizando el *Learning Path* de los grupos para tener una idea de los problemas por los que ha ido navegando durante toda la práctica. En la Figura Figura 66a podemos ver el recorrido que hizo el grupo `DBA1920P2GG`. No obstante, para el cálculo de este coeficiente se usará un grafo de mayor complejidad, en el que se subdividen los estados en `Pi OK` o `Pi FAIL` dependiendo del milestone alcanzado (`OK` indica que se ha resuelto problema). En la Figura 66b se muestra este nuevo grupo para el grupo de prácticas que estamos considerando.



(a) Grafo que muestra la exploración de los problemas que ha realizado.

(b) Grafo que muestra la exploración de los problemas, considerando si un problema ha sido resuelto (`OK`) o no (`FAIL`).

Figura 66: Leaning Path del grupo de prácticas `DBA1516P2GG`.

En la Figura 67 podemos ver que el *Laplaciano*, denotado por *LOGLAP09*, no sigue una distribución perfectamente normal.

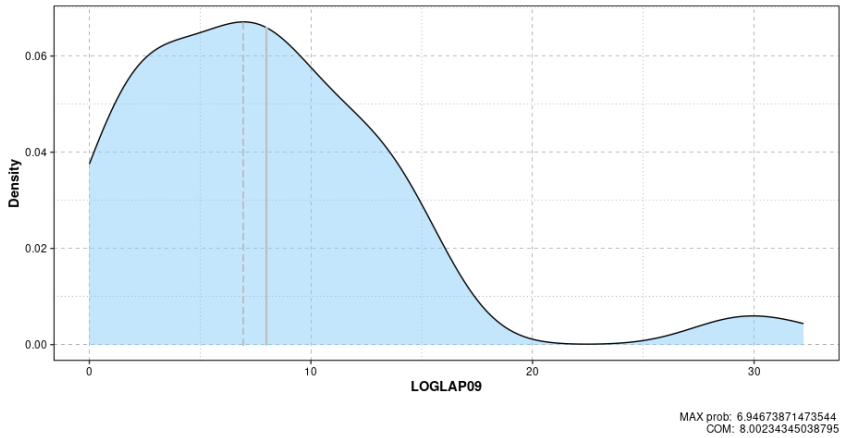


Figura 67: Función de densidad de probabilidad del *Laplaciano* obtenido por los distintos grupos de prácticas.

8.0.2 *El coeficiente* DAG

El coeficiente DAG se calcula a partir de la matriz de adyacencia de un grafo dirigido acíclico. Así pues, es la suma de dos componentes, ponderadas por 0,2 y 0,8 respectivamente.

La primera de las componentes del coeficiente se calcula a partir de todos los caminos del nodo inicial (`START`) al nodo final (`END`) de un grafo dirigido acíclico dado. Así pues, se calcula la media de los problemas abiertos en este tipo de caminos. Por ejemplo, dado el grado de la Figura, un posible camino es `{START,PROBLEM1-20,PROBLEM1-40,PROBLEM1-60,END}`, en el que se abre un problema.

Esta primera componente intenta captar el comportamiento de algunos grupos que van saltando de problema en problema en la misma sesión, valorándolo de manera negativa.

Por el contrario, la segunda componente no tiene ningún contenido semántico. Ésta simplemente consiste en cálcular la desviación estándar de las componentes de la matriz de adyacencia no nulas y dividirla por la media de dichas entradas. En resumen, ésta segunda componente trata de ver cómo de balanceados están los nodos de un grafo.

El pseudocódigo del cálculo de este coeficiente puede verse en el Algoritmo 1.

Algorithm 1 Función encargada del cálculo del coeficiente DAG.

```
1: function GET_COEFFICIENT
2:     coefficient ← 0
3:     problems ← ∅                                  ▷ Tiene el mismo tamaño que paths
4:     for i = 0 to paths.size() do
5:         problems[i].add(num_problems(paths[i]))
6:     end for
7:     if paths.size() == 0 then
8:         coefficient ← 3
9:     else
10:        mean ← calculate_mean(problems)
11:        mean ← (mean − 1)/(9 − 1)                 ▷ Normalización
12:        coefficient ← coefficient + 0,2 · mean
13:        nonzero ← ∅
14:        mean_frequency ← 0
15:        n ← 0
16:        for row in frequency do
17:            for element in row do
18:                if element > 0 then
19:                    nonzero.add(element)
20:                    mean_frequency ← mean_frequency + element
21:                    n ← n + 1
22:                end if
23:            end for
24:        end for
25:        mean_frequency ← mean_frequency/n
26:        sum_squares ← 0
27:        for i = 0 to n do
28:            sum_squares ← sum_squares + (nonzero[i] − mean_frequency)^2
29:        end for
30:        variance ← sum_squares/(n − 1)
31:        std_dev ← sqrt(variance)
32:        coefficient ← coefficient + 0,8 · std_dev/mean_frequency
33:    end if
34:    return coefficient
35: end function
```

Parte IV

PLANIFICACIÓN DEL PROYECTO

Estructuración en objetivos, división en sprints y seguimiento.

9

ETAPAS DEL PROYECTO: DIVISIÓN EN OBJETIVOS

El proyecto se ha realizado siguiendo la metodología *Scrum* Durante la fase inicial de planificación del proyecto se realizó una subdivisión del mismo en iteraciones:

- Realización de un estudio multianual y segmentado por calificaciones y primeros resultados de homogeneidad de las muestras transversal por años mediante la realición de análisis ANOVA.
- Extracción de procesos ocultos en los datasets utilizando el programa DISCO y programación y mejora del proceso de extracción. El resultado de esta fase serán una serie de grafos representando a cada uno de los grupos de prácticas considerados donde los arcos implican una relación de dependencia temporal. Estos grafos se representarán a partir de matrices de adyacencia cuyos vértices podrán representar problemas de prácticas, pares problema de prácticas y milestone alcanzado o pares problema de prácticas y estado (`FAIL` si no se ha resuelto el problema y `OK` en caso contrario).
- Análisis de los procesos por distintas categorías: por años (resultando ser estadísticamente iguales) y por calificación final del grupo (resultando en la existencia de diferencias). Así pues, se pretenderá caracterizar el comportamiento de los grupos.
- Análisis del comportamiento de un mismo grupo a lo largo del tiempo. Se realizará un estudio con el fin de determinar si el comportamiento de un grupo varía durante el desarrollo de las prácticas.

10

ETAPAS DEL PROYECTO: DIVISIÓN EN SPRINTS Y SEGUIMIENTO DE LOS MISMOS

El proyecto se ha dividido en diez sprints de dos semanas cada uno. Además, durante el transcurso de cada sprint se ha realizado un seguimiento del trabajo realizado en el mismo, aportando burdown charts de cada uno de ellos.

10.1 análisis de cada sprint

10.1.1 *Sprint 1*

10.1.2 *Sprint 2*

10.1.3 *Sprint 3*

10.1.4 *Sprint 4*

10.1.5 *Sprint 5*

10.1.6 *Sprint 6*

10.1.7 *Sprint 7*

10.1.8 *Sprint 8*

10.1.9 *Sprint 9*

10.1.10 *Sprint 10*

10.2 análisis

Parte V

RESULTADOS OBTENIDOS

Análisis de los resultados obtenidos.

11

ANÁLISIS DE LAS CORRELACIONES ENTRE LAS DISTINTAS MÉTRICAS

En primer, las correlaciones que más nos interesan son las de las distintas métricas con la variable *Grade*. En la Figura 68 podemos ver que no todas las métricas correlan con la misma.

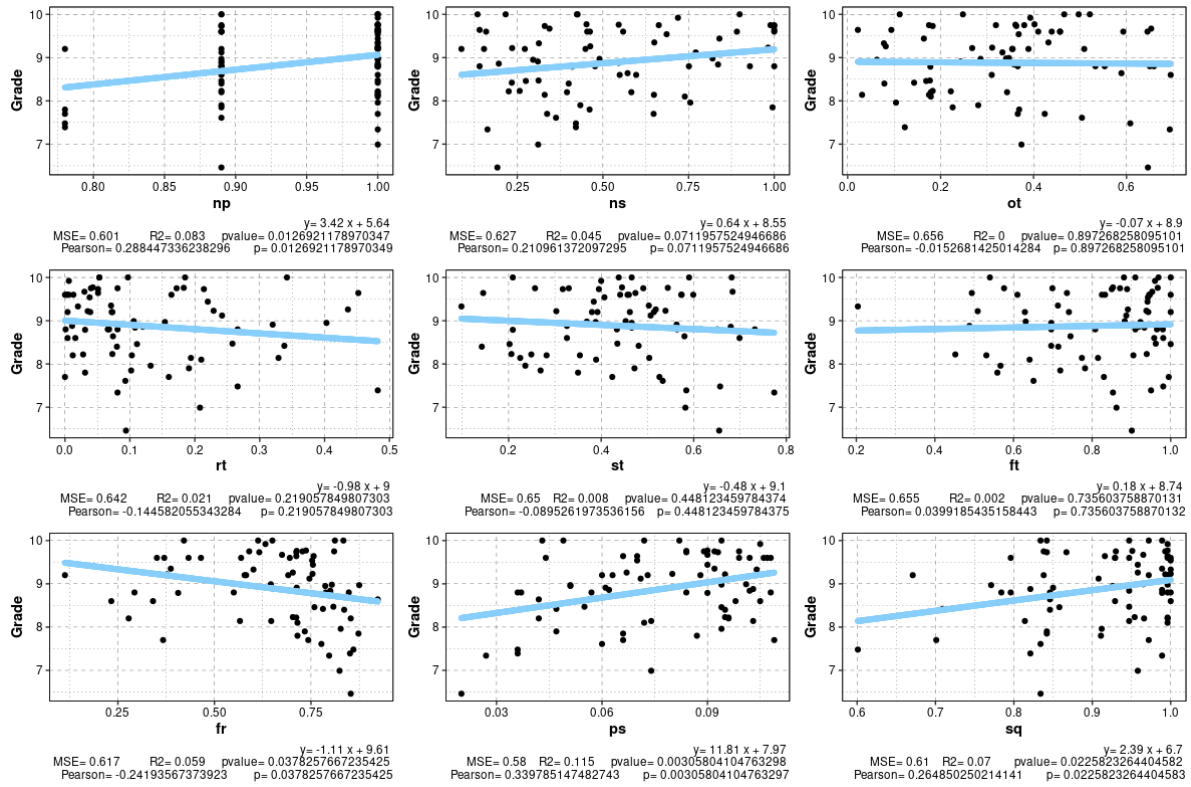


Figura 68: Correlaciones existentes entre las distintas métricas y la variable *Grade*.

Así pues, vemos que las variables *np* ($p = 0{,}0127 < 0{,}05$), *fr* ($p = 0{,}0378 < 0{,}05$), *ps* ($p = 0{,}0031 < 0{,}05$) y *sq* ($p = 0{,}0226 < 0{,}05$) correlan con la calificación obtenida y que la variable *ns* podría correlar con la variable *Grade* aunque con un grado de certeza menor que el resto ($p = 0{,}0712 < 0{,}1$).

También estudiaremos las correlaciones entre las medidas basadas en el análisis espectral de grafos y la variable *Grade*. Empezaremos estudiando la correlación entre la medida *LOGLAP09* anteriormente presentada y la calificación obtenida por los distintos grupos de prácticas.

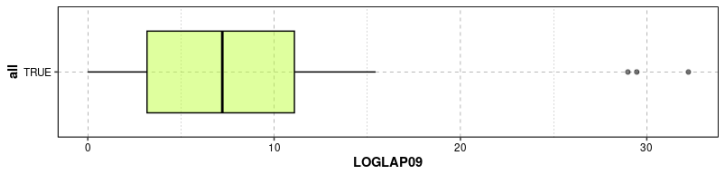


Figura 69: Distribución de los coeficientes *LOGLAP09* obtenidos por los distintos grupos de alumnos inicial.

En primer lugar, observamos la presencia de outliers en la distribución de la msima (Figura 69). Tras la eliminación de los mismos, podemos ver en la Figura 70 que la medida de rendimiento espectral *LOGLAP09* no correla con la variable *Grade* ($p = 0{,}5698 > 0{,}05$).

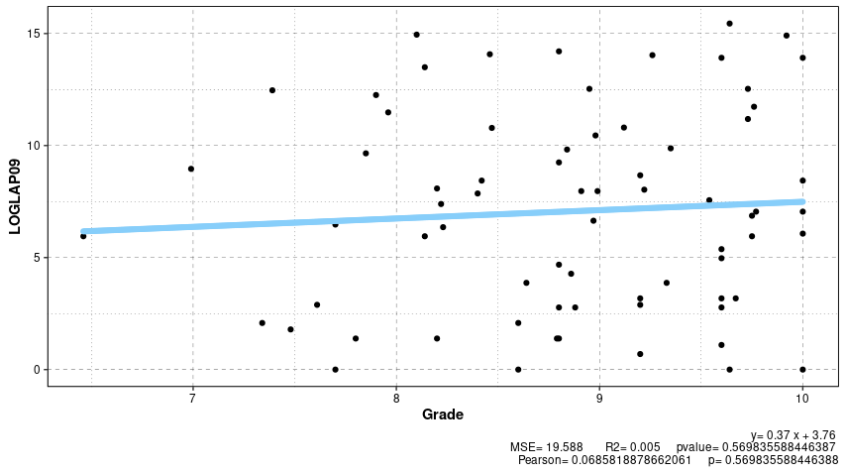


Figura 70: Correlación existente entre la métrica *LOGLAP09* y la variable *Grade*.

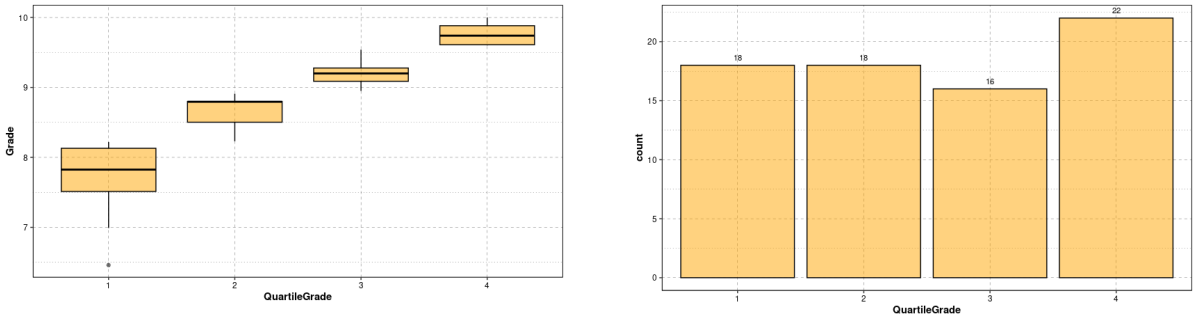
12

PERFILES DE ESTUDIANTES SEGÚN SU RENDIMIENTO

Debido a la dificultad de predecir de manera exacta la calificación obtenida por los alumnos a partir de las medidas de rendimiento descritas anteriormente, se agruparán las notas en clusters significativos y trataremos de predecir en qué cluster se encuentra la nota de un determinado grupo de alumnos.

12.1 POR CLUSTERS FIJOS DE NOTAS

En primer lugar, escogeremos como separación los cuartiles de las calificaciones. Así pues, podemos ver la distribución de los cuartiles en la Figura 71a, donde los límites inferiores de cada una de las cajas son 6,99, 8,23, 8,95 y 9,60 respectivamente. También puede verse en la Figura 71b el número de grupos que hay en cada cuartil.



(a) Boxplot de las calificaciones por cuartil.

(b) Número de grupos por cuartil.

Figura 71: Resultados obtenidos tras aplicar el algoritmo agrupar las caficaciones por cuartiles.

Sin embargo, en la Figura 71a y en la Figura 72, donde se representan cómo de frecuentes son cada una de las calificaciones obtenidas, notamos la presencia de un outlier.

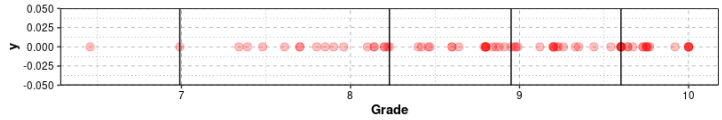


Figura 72: Calificaciones obtenidas por los distintos grupos. El límite de los cuartiles se ha indicado con líneas verticales negras.

En la Figura 73 vemos las funciones de densidad por cuartil. Prestaremos especial atención a los grupos del cluster `Q1` (el de las peores calificaciones al que le añadimos el outlier) puesto que son los que peor rendimiento han mostrado.

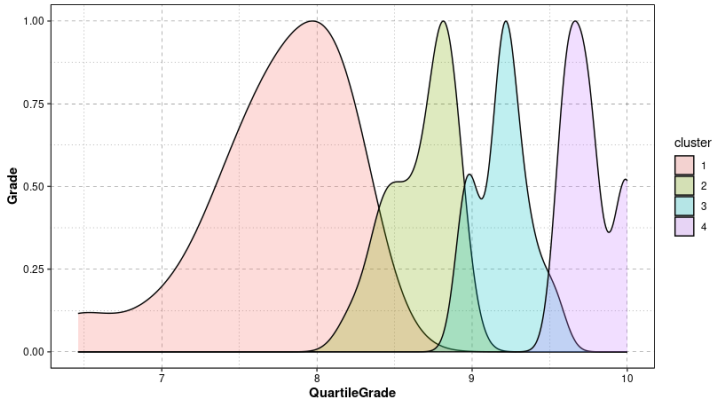


Figura 73: Funciones de densidad de las calificaciones obtenidas por cluster.

12.2 POR CLUSTERS DINÁMICOS DE NOTAS

Se agruparán los datos usando el algoritmo de las K-medias sobre la variable *Grade*. Para decidir el número de clusters en el que agruparemos los datos, se usarán métodos gráficos. Como podemos ver en las Figuras 74 y 75, el número óptimo de particiones podría ser 3 o 5. Para decidir entre un número de clusters u otro se realizarán los dos agrupamientos y nos quedaremos con el de menor error.

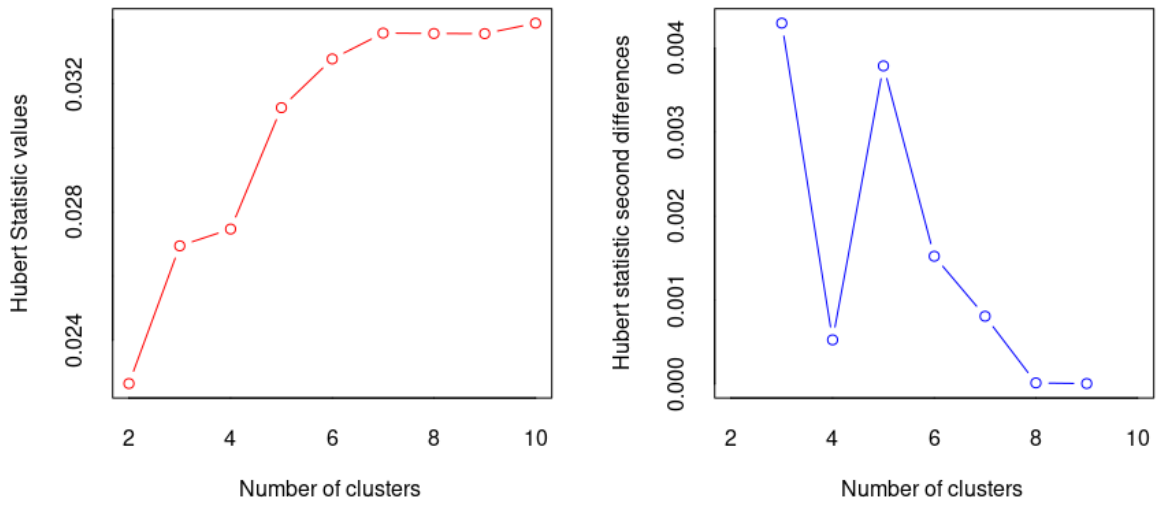


Figura 74: Valores estadísticos de Hubert.

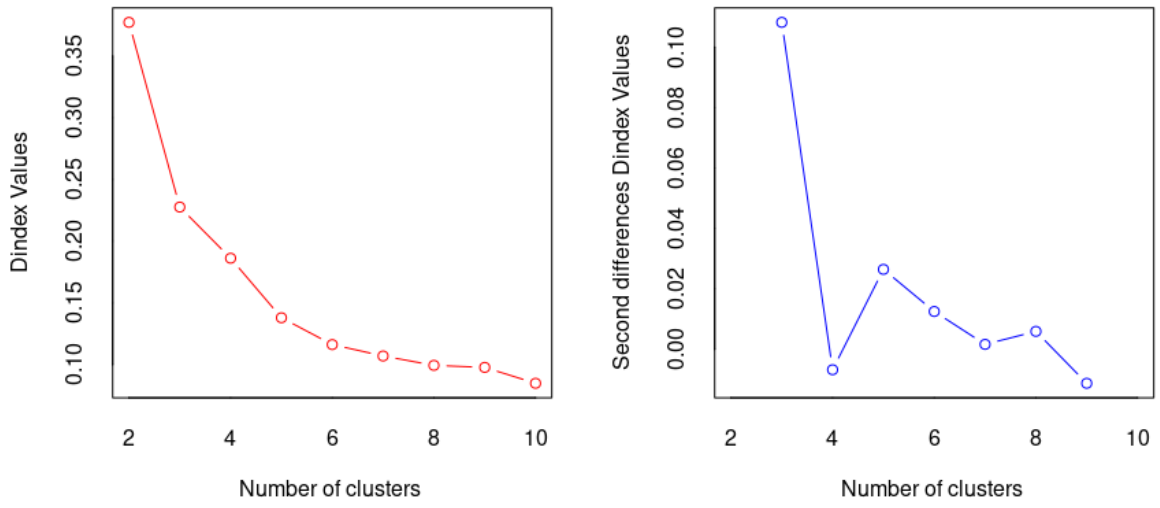


Figura 75: Valores de Dindex.

Aplicando el algoritmo de las K-medias para $K = 2$ (Figura 76) y para $K = 3$ (Figura 77), vemos que hay una gran diferencia la precisión de uno y otro (para $K = 2$ se tiene `accuracy` = 0,6978412 mientras que para $K = 3$ se tendrá `accuracy` = 0,8585982). Como podemos ver en la Figura 76, seguimos teniendo un outlier.

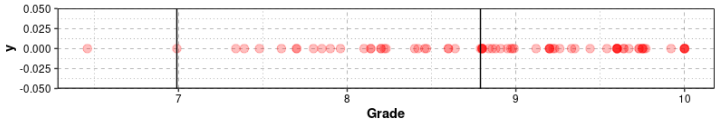


Figura 76: Particiones obtenidas con $K = 2$.

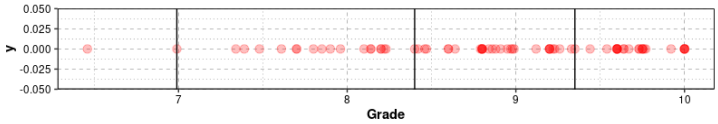
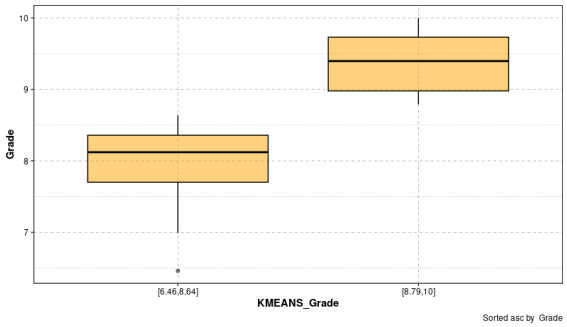
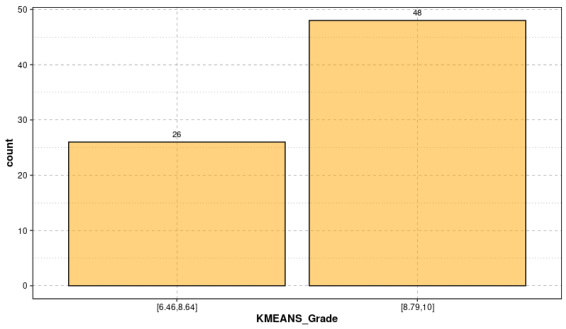


Figura 77: Particiones obtenidas con $K = 3$.

La distribución de la variable *Grade* dentro de cada partición puede verse en las Figuras 78a y 79a mientras que el número de grupo que hay en las particiones puede verse en las Figuras 78b y 79b.

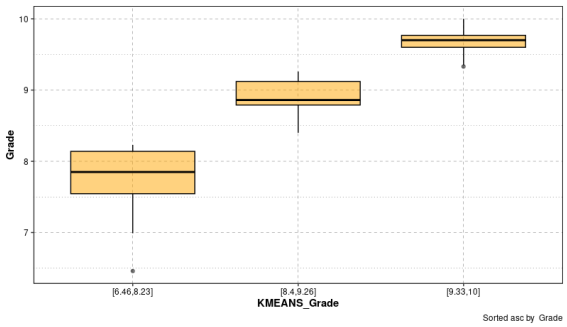


(a) Boxplot de cada una de las particiones.

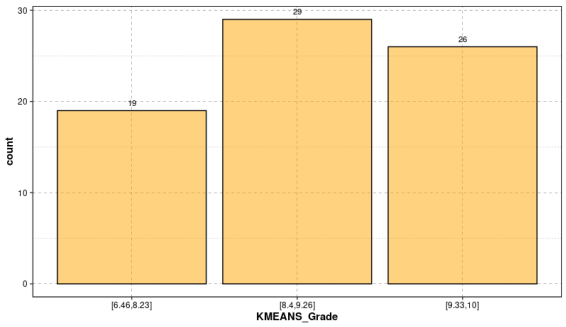


(b) Número de grupos por partición.

Figura 78: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 2$.



(a) Boxplot de cada una de las particiones.



(b) Número de grupos por partición.

Figura 79: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 3$.

Por último, para cinco particiones (Figura 80) se tendrá `accuracy` = 0,9474439. Es decir, tenemos más precisión con cinco particiones y ya no tenemos outliers. Nos centramos en estudiar los grupos de los dos primeros clusters (aquellos grupos con una nota inferior a 7,34).

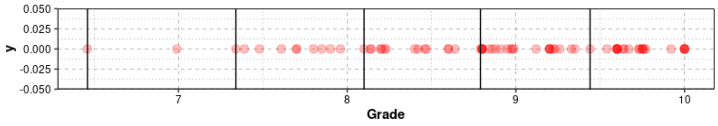
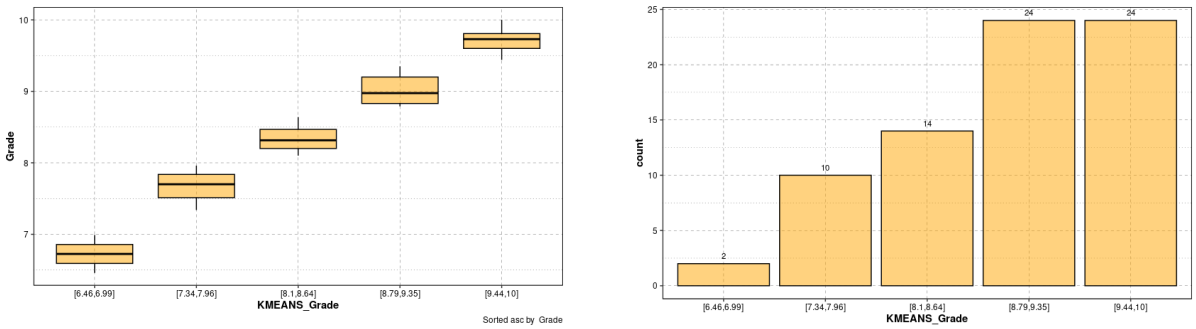


Figura 80: Particiones obtenidas con $K = 5$.



(a) Boxplot de cada una de las particiones.

(b) Número de grupos por partición.

Figura 81: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 5$.

12.3 POR CLUSTERS APROXIMADOS DE RENDIMIENTO

De las medidas de rendimiento estudiadas en el Capítulo 7, nos quedaremos con aquellas que correlan con la variable *Grade* (*np*, *fr*, *ps*, *sq* y *ns*). Así pues, se definirá una nueva métrica como la suma de las medidas de rendimiento *p*, *fr*, *ps*, *sq* y *s*. En la Figura 82 vemos que ésta no correla con la variable *Grade*.

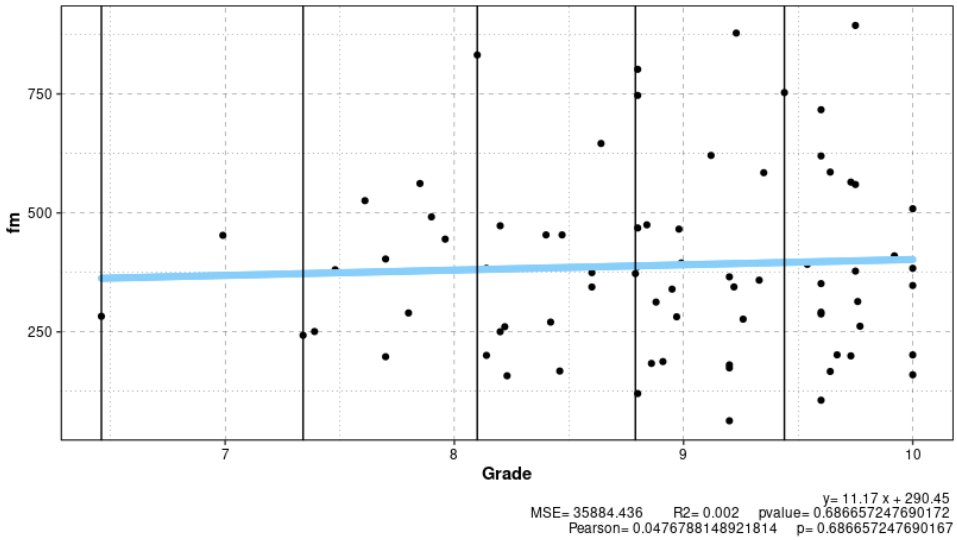


Figura 82: Regresión lineal para aproximar la relación de dependencia entre la variable *fm* y la variable *Grade*.

A continuación, se agruparán los datos usando el algoritmo de las K-medias sobre la variable *fm*. Para decidir el número de clusters en el que agruparemos los datos, se usarán métodos gráficos. Como podemos ver en las Figuras 83 y 84, el número óptimo de particiones podría ser 4 o 7. Para decidir entre un número de clusters u otro se realizarán los dos agrupamientos y nos quedaremos con el de menor error.

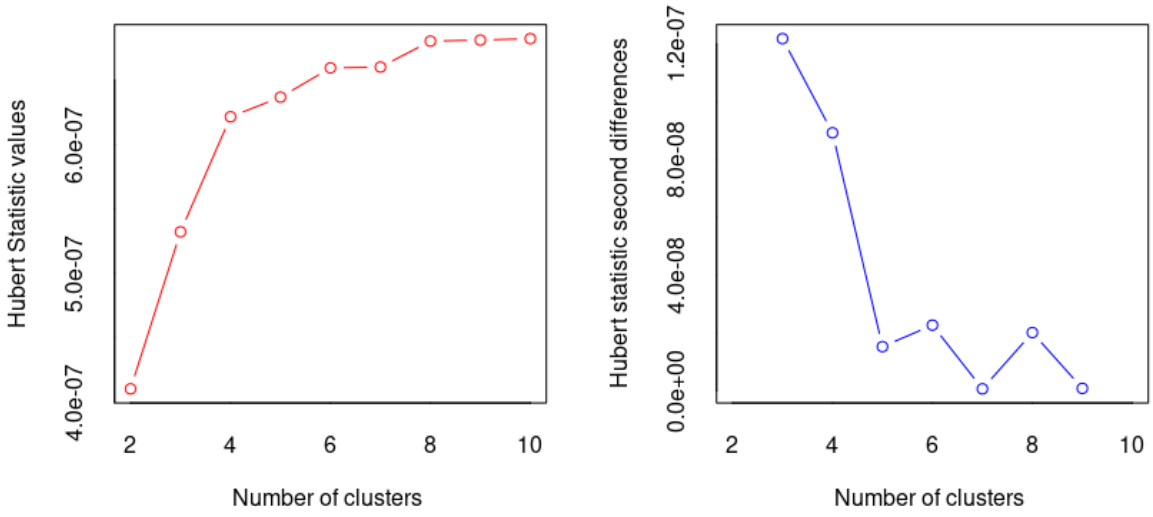


Figura 83: Valores estadísticos de Hubert.

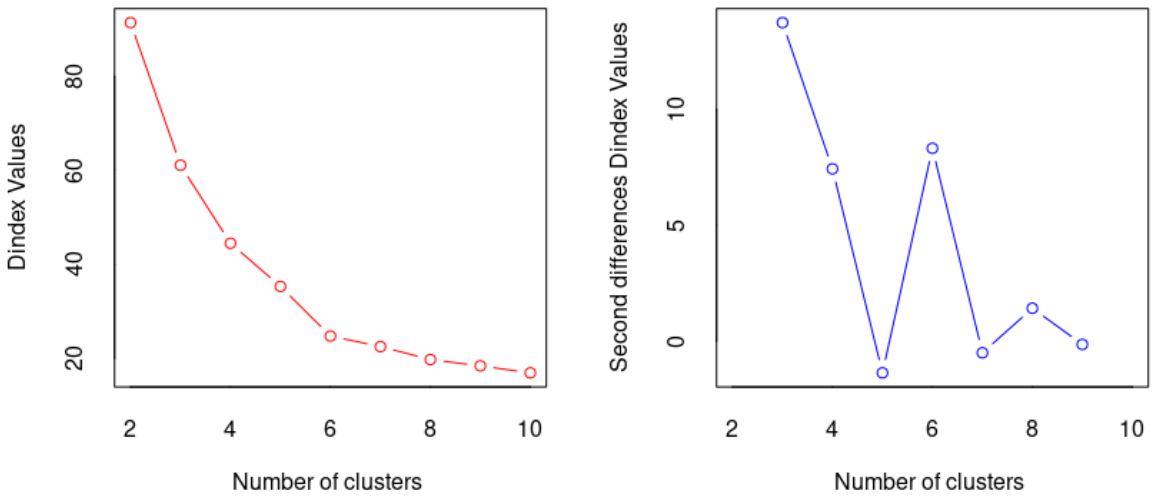


Figura 84: Valores de Dindex.

Aplicando el algoritmo de las K-medias para $K = 4$ (Figura 85) y para $K = 7$ (Figura 86), vemos que hay una gran diferencia la precisión de uno y otro (para $K = 4$ se tiene `accuracy` = 0,921137 mientras que para $K = 7$ se tendrá `accuracy` = 0,9772655). Además, en la Figura 85 podemos notar la presencia de outliers mientras que en la Figura 86 no.

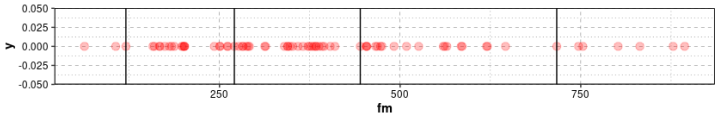


Figura 85: Particiones obtenidas con $K = 4$. Como podemos ver, se observa la presencia de outliers (260,804, 63,173, 106,379 y 261,823).

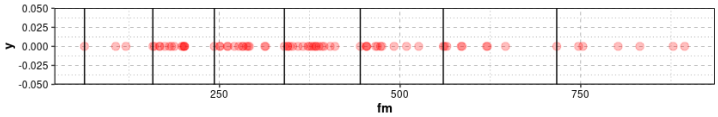
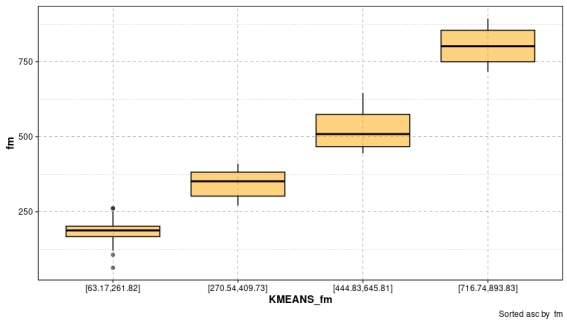
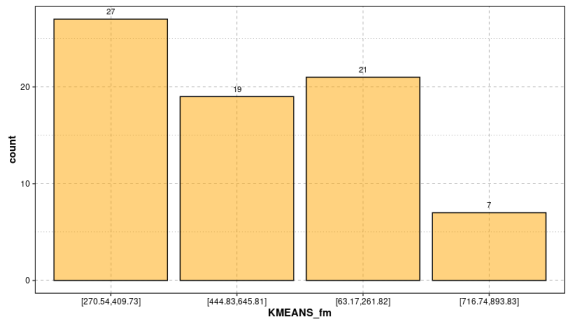


Figura 86: Particiones obtenidas con $K = 7$. No hay ningún outlier.

La distribución de la variable *Grade* dentro de cada partición puede verse en las Figuras 87a y 88a mientras que el número de grupo que hay en las particiones puede verse en las Figuras 87b y 88b.

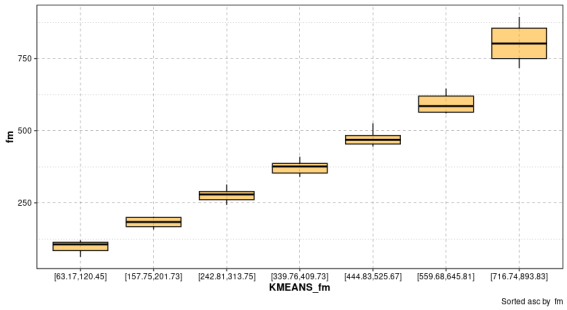


(a) Boxplot de cada una de las particiones.

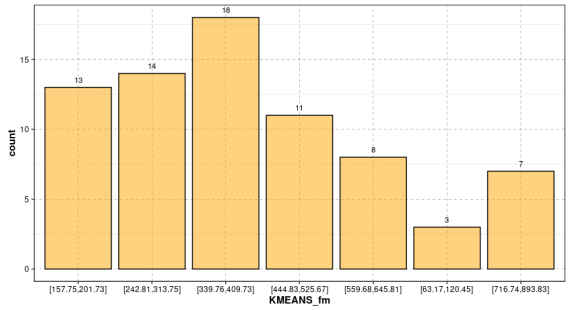


(b) Número de grupos por partición.

Figura 87: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 4$.



(a) Boxplot de cada una de las particiones.



(b) Número de grupos por partición.

Figura 88: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 7$.

12.4 CLUSTERING MEDIANTE LAS PROPIEDAS ESPECTRALES DE LOS GRAFOS

12.4.1 *Clustering mediante el coeficiente LOGLAP09*

Ahora, se ha decidido se asociar los datos usando el algoritmo de las K-medias sobre la variable *LOGLAP09*. Para decidir el número de clusters en el que agruparemos los datos, se usarán métodos gráficos. Como podemos ver en las Figuras 89 y 90, se ha decidido agrupar los datos en 5 clusters (Figura 91, `accuracy` = 0,9673813).

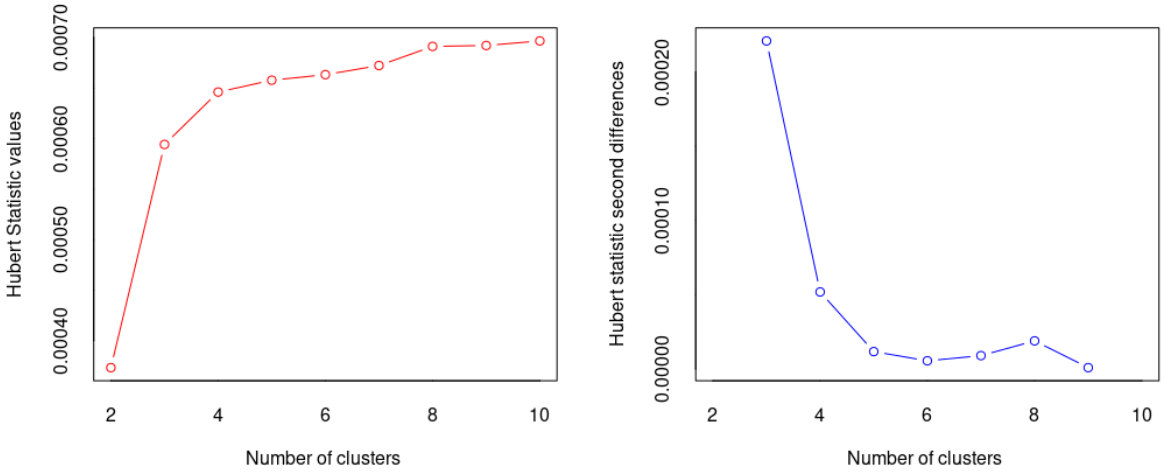


Figura 89: Valores estadísticos de Hubert.

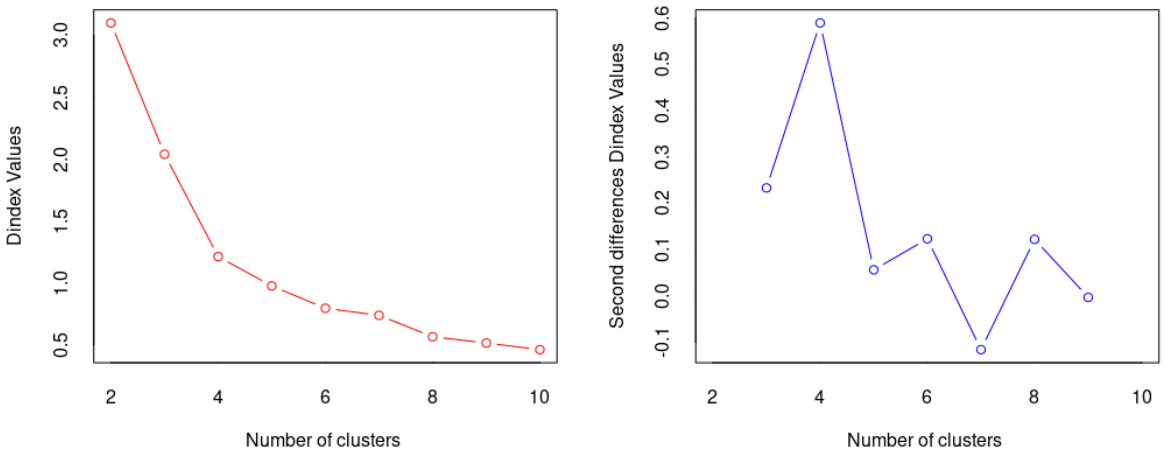


Figura 90: Valores de Dindex.

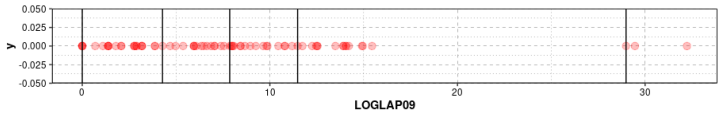
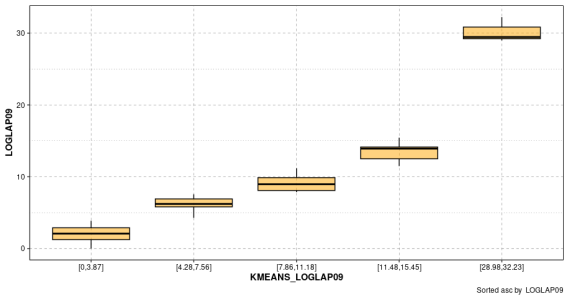
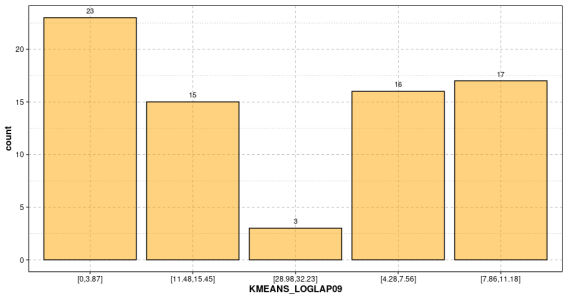


Figura 91: Particiones obtenidas con $K = 5$. No hay ningún outlier.

La distribución de la variable *Grade* dentro de cada partición puede verse en la Figura 92a mientras que el número de grupos que hay en las particiones puede verse en la Figura 92b.



(a) Boxplot de cada una de las particiones.



(b) Número de grupos por partición.

Figura 92: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 5$.

12.4.2 *Clustering mediante el coeficiente DAG*

13

CLASIFICACIÓN DE LOS GRUPOS DE ALUMNOS SEGÚN SU RENDIMIENTO

En primer lugar, clasificaremos los grupos en cuartiles utilizando las métricas que correlan con las calificaciones obtenidas y las medidas basdas en el análisis espectral de grafos *DAG* y *LOGLAP09*.

A

ALGUNAS TABLAS

Cuadro 17: Listado de los grupos por curso académico.

Y2015	**Y2016**	**Y2017**	**Y2018**
DBA 1516 P2 GA	DBA 1617 P2 GA	DBA 1718 P2 GA	DBA 1819 P2 GB
DBA 1516 P2 GB	DBA 1617 P2 GB	DBA 1718 P2 GB	DBA 1819 P2 GC
DBA 1516 P2 GC	DBA 1617 P2 GD	DBA 1718 P2 GC	DBA 1819 P2 GD
DBA 1516 P2 GD	DBA 1617 P2 GE	DBA 1718 P2 GD	DBA 1819 P2 GE
DBA 1516 P2 GE	DBA 1617 P2 GF	DBA 1718 P2 GE	DBA 1819 P2 GF
DBA 1516 P2 GF	DBA 1617 P2 GG	DBA 1718 P2 GG	DBA 1819 P2 GG
DBA 1516 P2 GG	DBA 1617 P2 GH	DBA 1718 P2 GH	DBA 1819 P2 GH
DBA 1516 P2 GH	DBA 1617 P2 GI		DBA 1819 P2 GI
DBA 1516 P2 GI	DBA 1617 P2 GJ		DBA 1819 P2 GJ
			DBA 1819 P2 GK
			DBA 1819 P2 GL

Cuadro 18: Listado de los grupos por curso académico.

Y2019	Y2020	Y2021
DBA 1920 P2 GB	DBA 2021 P2 GA	DBA 2122 P2 GA
DBA 1920 P2 GC	DBA 2021 P2 GB	DBA 2122 P2 GB
DBA 1920 P2 GD	DBA 2021 P2 GC	DBA 2122 P2 GC
DBA 1920 P2 GE	DBA 2021 P2 GD	DBA 2122 P2 GD
DBA 1920 P2 GF	DBA 2021 P2 GE	DBA 2122 P2 GE
DBA 1920 P2 GH	DBA 2021 P2 GF	DBA 2122 P2 GF
DBA 1920 P2 GI	DBA 2021 P2 GG	DBA 2122 P2 GG
DBA 1920 P2 GJ	DBA 2021 P2 GH	DBA 2122 P2 GH
DBA 1920 P2 GK	DBA 2021 P2 GI	DBA 2122 P2 GI
DBA 1920 P2 GL	DBA 2021 P2 GJ	DBA 2122 P2 GJ
DBA 1920 P2 GM	DBA 2021 P2 GK	DBA 2122 P2 GK
DBA 1920 P2 GN	DBA 2021 P2 GL	DBA 2122 P2 GL
	DBA 2021 P2 GM	DBA 2122 P2 GM
		DBA 2122 P2 GN
		DBA 2122 P2 GO
		DBA 2122 P2 GP

Year	Group	fail	solved	all
Y2015	DBA 1516 P2 GA	738	54	792
Y2015	DBA 1516 P2 GB	62	49	111
Y2015	DBA 1516 P2 GC	142	195	337
Y2015	DBA 1516 P2 GD	298	80	378
Y2015	DBA 1516 P2 GE	597	139	736
Y2015	DBA 1516 P2 GF	246	110	356
Y2015	DBA 1516 P2 GG	398	64	462
Y2015	DBA 1516 P2 GH	525	181	706
Y2015	DBA 1516 P2 GI	469	142	611
Y2016	DBA 1617 P2 GA	132	59	191
Y2016	DBA 1617 P2 GB	564	178	742
Y2016	DBA 1617 P2 GD	154	208	362
Y2016	DBA 1617 P2 GE	258	316	574
Y2016	DBA 1617 P2 GF	126	47	173
Y2016	DBA 1617 P2 GG	680	187	867

Y2016	DBA 1617 P2 GH	722	161	883
Y2016	DBA 1617 P2 GI	252	122	374
Y2016	DBA 1617 P2 GJ	333	39	372
Y2017	DBA 1718 P2 GA	186	46	232
Y2017	DBA 1718 P2 GB	1282	139	1421
Y2017	DBA 1718 P2 GC	369	73	442
Y2017	DBA 1718 P2 GD	369	74	443
Y2017	DBA 1718 P2 GE	468	48	516
Y2017	DBA 1718 P2 GG	156	178	334
Y2017	DBA 1718 P2 GH	235	38	273
Y2018	DBA 1819 P2 GB	148	0	148
Y2018	DBA 1819 P2 GC	178	0	178
Y2018	DBA 1819 P2 GD	190	0	190
Y2018	DBA 1819 P2 GE	158	0	158
Y2018	DBA 1819 P2 GF	190	0	190
Y2018	DBA 1819 P2 GG	266	0	266
Y2018	DBA 1819 P2 GH	434	0	434
Y2018	DBA 1819 P2 GI	242	0	242
Y2018	DBA 1819 P2 GJ	373	0	373
Y2018	DBA 1819 P2 GK	575	0	575
Y2018	DBA 1819 P2 GL	57	0	57
Y2019	DBA 1920 P2 GB	179	71	250
Y2019	DBA 1920 P2 GC	366	116	482
Y2019	DBA 1920 P2 GD	238	110	348
Y2019	DBA 1920 P2 GE	266	63	329
Y2019	DBA 1920 P2 GF	840	271	1111
Y2019	DBA 1920 P2 GH	206	54	260
Y2019	DBA 1920 P2 GI	119	37	156
Y2019	DBA 1920 P2 GJ	588	48	636
Y2019	DBA 1920 P2 GK	599	222	821
Y2019	DBA 1920 P2 GL	388	56	444
Y2019	DBA 1920 P2 GM	124	46	170
Y2019	DBA 1920 P2 GN	122	27	149

Y2020	DBA 2021 P2 GA	265	99	364
Y2020	DBA 2021 P2 GB	221	174	395
Y2020	DBA 2021 P2 GC	189	88	277
Y2020	DBA 2021 P2 GD	104	231	335
Y2020	DBA 2021 P2 GE	30	23	53
Y2020	DBA 2021 P2 GF	138	28	166
Y2020	DBA 2021 P2 GG	142	99	241
Y2020	DBA 2021 P2 GH	250	31	281
Y2020	DBA 2021 P2 GI	205	136	341
Y2020	DBA 2021 P2 GJ	376	82	458
Y2020	DBA 2021 P2 GK	177	105	282
Y2020	DBA 2021 P2 GL	517	93	610
Y2020	DBA 2021 P2 GM	60	37	97
Y2021	DBA 2122 P2 GA	336	48	384
Y2021	DBA 2122 P2 GB	471	28	499
Y2021	DBA 2122 P2 GC	516	39	555
Y2021	DBA 2122 P2 GD	347	34	381
Y2021	DBA 2122 P2 GE	168	23	191
Y2021	DBA 2122 P2 GF	418	37	455
Y2021	DBA 2122 P2 GG	331	37	368
Y2021	DBA 2122 P2 GH	258	45	303
Y2021	DBA 2122 P2 GI	490	59	549
Y2021	DBA 2122 P2 GJ	273	30	303
Y2021	DBA 2122 P2 GK	425	39	464
Y2021	DBA 2122 P2 GL	232	19	251
Y2021	DBA 2122 P2 GM	513	39	552
Y2021	DBA 2122 P2 GN	169	18	187
Y2021	DBA 2122 P2 GO	359	40	399
Y2021	DBA 2122 P2 GP	262	10	272

Cuadro 19: Número y tipo de las sesiones de trabajo.

Cuadro 20: Test HSD de Tukey (Honestly-significance-difference) de la tasa de fallo por problemas.

	diff	lwr	upr	p adj
P2-P1	0.00	-0.35	0.35	1.00
P3-P1	-0.12	-0.48	0.23	0.97
P4-P1	-0.19	-0.54	0.16	0.72
P5-P1	-0.26	-0.61	0.09	0.32
P6-P1	-0.12	-0.47	0.23	0.97
P7-P1	-0.14	-0.49	0.21	0.93
P8-P1	-0.11	-0.46	0.24	0.98
P9-P1	0.03	-0.33	0.38	1.00
P3-P2	-0.12	-0.48	0.23	0.97
P4-P2	-0.19	-0.54	0.16	0.72
P5-P2	-0.26	-0.61	0.09	0.31
P6-P2	-0.12	-0.47	0.23	0.97
P7-P2	-0.14	-0.50	0.21	0.93
P8-P2	-0.11	-0.46	0.24	0.98
P9-P2	0.03	-0.33	0.38	1.00
P4-P3	-0.07	-0.42	0.28	1.00
P5-P3	-0.14	-0.49	0.22	0.94
P6-P3	0.00	-0.35	0.36	1.00
P7-P3	-0.02	-0.37	0.33	1.00
P8-P3	0.01	-0.34	0.37	1.00
P9-P3	0.15	-0.20	0.50	0.91
P5-P4	-0.07	-0.42	0.28	1.00
P6-P4	0.07	-0.28	0.42	1.00
P7-P4	0.05	-0.30	0.40	1.00
P8-P4	0.08	-0.27	0.43	1.00
P9-P4	0.22	-0.14	0.57	0.56
P6-P5	0.14	-0.21	0.49	0.93
P7-P5	0.12	-0.23	0.47	0.97
P8-P5	0.15	-0.20	0.50	0.90
P9-P5	0.29	-0.07	0.64	0.20
P7-P6	-0.02	-0.38	0.33	1.00
P8-P6	0.01	-0.34	0.36	1.00
P9-P6	0.15	-0.21	0.50	0.92
P8-P7	0.03	-0.32	0.39	1.00
P9-P7	0.17	-0.19	0.52	0.83
P9-P8	0.14	-0.22	0.49	0.94

BIBLIOGRAFÍA

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *WIREs Data Mining Knowl Discov*, *8*, 12–30.

Mayorga, H. S. A., & García, N. R. (2015). Process mining: development, applications and critical factors. *Cuardenos de Administración*, *28*(50), 137–157.

Vidal, L. C. (2016). A virtual laboratory for multiagent systems: Joining efficacy, learning analytics and student satisfaction.