



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 7 de agosto de 2023

DECLARACIÓN DE ORIGINALIDAD

Dña. María Isabel Ruiz Martínez

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2022-2023, es original, entendida esta, en el sentido de que no ha utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 7 de agosto de 2023

Fdo: María Isabel Ruiz Martínez

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 7 de agosto 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 7 de agosto de 2023

Fdo: Luis Castillo Vidal

A GRADECIMIENTOS

A mi familia, por su apoyo incondicional y, en especial, a mi madre, por sus cafés cargados de cariño y consejos, y a mi padre, por ser una inagotable fuente de inspiración.

A mis amigos, por todas las experiencias compartidas y las que nos quedan por compartir.

A Luis, por haberme iniciado en la minería de procesos y haberme instruido en una infinidad de herramientas, siempre con entusiasmo y dedicación.

A todas las personas que, de una forma u otra, me han transmitido alguna enseñanza a lo largo de mi vida.

¡Gracias por todo!

RESUMEN

El estudio del proceso de aprendizaje que realizan los alumnos cuando se les plantean una serie de tareas es fundamental puesto que éste podría facilitar la asimilación de nuevos conocimientos y de hacer accesible aquello que se enseña al estudiantado.

El objetivo principal de este trabajo es, precisamente, identificar los patrones de comportamiento de aquellos estudiantes en riesgo de obtener un rendimiento más bajo de lo esperado con el fin de permitir una temprana intervención del personal docente, evitando así el fracaso de los mismos en la adquisición de nuevos conocimientos.

Así pues, en este trabajo fin de grado se usarán datos obtenidos en un laboratorio virtual para, a través de técnicas de Minería de Procesos, extraer una serie de grafos, representados mediante matrices, que reflejan el comportamiento de los alumnos en la plataforma tal y como se demostrará en este estudio.

Una parte importante del mismo ha sido el desarrollo de una herramienta de Minería de Procesos de creación propia, *Graph Miner*, que se encarga de traducir los registros del labotorio virtual en las matrices que representan el comportamiento de los alumnos. Como veremos, ésta suplirá los inconvenientes del programa de Minería de Procesos ya existente, *Disco Günther & Rozinat (2012)*.

Adicionalmente, se usarán técnicas de aprendizaje automático supervisado no sólo para prever grupos en riesgo sino también para predecir en qué intervalo de notas se encuentra la calificación de los diferentes grupos de alumnos con evidencias estadísticas. De hecho, se pueden realizar tales predicciones con una alta fiabilidad en estadios tempranos del desarrollo de la práctica.

Por último, destacar que, para realizar las clasificaciones descritas en el párrafo anterior, se han utilizado tanto medidas clásicas del rendimiento de los alumnos como medidas de complejidad basadas únicamente en la topología del grafo representado a través de la matriz característica de cada grupo, demostrando éstas últimas ser igual de útiles que las primeras.

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

SUMMARY

The study of the learning process of students when they are given a series of tasks is fundamental since it could facilitate the assimilation of new knowledge and make what is being taught accessible to the students.

The main objective of this work is, precisely, to identify the behavioural patterns of those students at risk of underachieving in order to allow early intervention by teaching staff, preventing their failure in the acquisition of new knowledge.

Thus, in this project, data obtained in a virtual laboratory will be used to extract, through Process Mining techniques, a series of graphs, represented by matrices, which reflect the behaviour of the students on the platform as it will be demonstrated in this study.

An important part of this thesis has been the development of my own Process Mining tool, *Graph Miner*, which is responsible for translating the virtual lab logs into the matrices that represent the behaviour of the learners. As we will see, it will make up for the drawbacks of the existing Process Mining programme, *Disco*. [Günther & Rozinat \(2012\)](#).

In addition, supervised machine learning techniques will be used not only to predict groups at risk but also to predict in which range of grades the different groups of students are with statistical evidence. In fact, such predictions can be made with a high reliability at early stages of the development of students' practical work.

Finally, it should be noted that, in order to carry out the classifications described in the previous paragraph, both classical measures of student performance and measures of complexity based solely on the topology of the graph represented by the characteristic matrix of each group have been used to make the rankings described in the previous paragraph, the latter proving to be just as useful as the former.

KEY WORDS: learning analytics, learning styles, process mining, student success, graph theory, directed acyclic graph

ÍNDICE GENERAL

I. MOTIVACIONES	15
Motivación	16
0.1. Introducción	16
0.2. Motivación	16
0.3. Objetivos	17
0.4. Estructura del Trabajo Fin de Grado	17
II. ESTADO DEL ARTE	20
1. PLANTEAMIENTO DEL PROBLEMA	21
Planteamiento del problema	21
1.1. Funcionamiento del laboratorio virtual	23
1.2. Descripción del escenario inicial	26
2. INTRODUCCIÓN A LA MINERÍA DE PROCESOS Y TRABAJOS PREVIOS	29
Introducción a la Minería de Procesos y trabajos previos	29
3. MINERÍA DE PROCESOS	33
Minería de Procesos	33
3.1. Extracción de los procesos con DISCO	33
4. IMPLEMENTACIÓN DE LA HERRAMIENTA DE MINERÍA DE PROCESOS <i>graph miner</i>	38
Implementación de la herramienta de Minería de Procesos <i>Graph Miner</i>	38
4.1. Uso de grafos como representación de los grupos	38
4.2. La matriz característica de un grupo	40
4.3. Resultados obtenidos	41
5. TEORÍA DE GRAFOS	49
Teoría de grafos	49
5.1. Grafos	49
5.2. Medidas de complejidad de propósito general	62
5.2.1. Densidad	63
5.2.2. Grado medio	63
5.2.3. Longitud del camino característico	63
5.2.4. Diámetro del grafo	63
5.2.5. Conectividad	64
5.2.6. Betweenness	64

5.2.7. Dag	64
5.2.8. WDag	64
5.2.9. St	64
5.2.10. Peso del comportamiento	67
5.2.11. Eficacia	67
5.2.12. Balance	67
III. ANÁLISIS DESCRIPTIVO	69
6. LOS REGISTROS EXISTENTES	70
Los registros existentes	70
6.1. Número de grupos cada año	70
6.2. El periodo de tiempo analizado cada año	71
6.3. El conjunto de problemas analizados cada año	71
6.3.1. Dificultad del problema: la tasa de fallo	72
6.3.2. Dificultad del problema: tiempo necesario en resolverlo	75
6.4. Actividad registrada	78
6.4.1. Análisis de la normalidad de la distribución del número de sesiones .	79
6.4.2. Sesiones por cada problema	83
6.4.3. Sesiones cada año	84
6.4.4. Análisis de la distribución del número de problemas resueltos . . .	86
7. HIPÓTESIS DE ESTUDIO	88
Hipótesis de estudio	88
7.1. Métricas de calidad y correlaciones entre ellas	88
7.1.1. Medidas a posteriori del resultado de la práctica	89
7.1.2. Medidas continuas durante la práctica	89
8. RENDIMIENTO OBSERVADO DE LOS ALUMNOS	91
Rendimiento observado de los alumnos	91
8.1. Calificaciones obtenidas (Grade)	91
8.2. Número total de problemas resueltos (p)	95
8.3. Finalizar la práctica (ft)	97
8.4. Número de sesiones realizadas (s)	98
8.5. Abrir un problema por primera vez (ot)	98
8.6. Tasa de fallo (fr)	102
8.7. Tiempo empleado en la resolución de un problema por primera vez (rt) . . .	106
8.8. Exploración de nuevas vías y mejoras (ps)	109
8.9. Resolver un problema por primera vez (st)	109
8.10. Siguiendo el plan del profesor (sq)	113

IV. PLANIFICACIÓN DEL PROYECTO	115
9. ETAPAS DEL PROYECTO: DIVISIÓN EN OBJETIVOS	116
Etapas del proyecto: división en objetivos	116
10. ETAPAS DEL PROYECTO: DIVISIÓN EN SPRINTS Y SEGUIMIENTO DE LOS MISMOS	117
Etapas del proyecto: división en sprints	117
10.1. Análisis de cada sprint	118
10.1.1. Sprint 1 (Figura 86a)	119
10.1.2. Sprint 2 (Figura 86b)	119
10.1.3. Sprint 3 (Figura 86c)	119
10.1.4. Sprint 4 (Figura 86d)	119
10.1.5. Sprint 5 (Figura 86e)	120
10.1.6. Sprint 6 (Figura 86f)	120
10.1.7. Sprint 7 (Figura 87a)	122
10.1.8. Sprint 8 (Figura 87b)	122
10.1.9. Sprint 9 (Figura 87c)	122
10.1.10. Sprint 10 (Figura 87d)	122
10.1.11. Sprint 11 (Figura 87e)	122
10.1.12. Sprint 12 (Figura 87f)	122
10.2. Análisis global del proyecto	122
V. RESULTADOS OBTENIDOS	123
11. ANÁLISIS DE LAS CORRELACIONES ENTRE LAS DISTINTAS MÉTRICAS	124
Análisis de las correlaciones entre las distintas métricas	124
12. PERFILES DE ESTUDIANTES SEGÚN SU RENDIMIENTO	128
Perfiles de estudiantes según su rendimiento	128
12.1. Por clusters fijos de notas	128
12.2. Por clusters dinámicos de notas	129
12.3. Por clusters aproximados de rendimiento	132
12.4. Clustering mediante las propiedades espectrales de los grafos	135
12.4.1. Clustering mediante el coeficiente LOGLAP09	135
12.4.2. Clustering mediante el coeficiente DAG	137
12.4.3. ¿Quiénes son los grupos en riesgo?	137
13. CAPACIDAD DE LAS MÉTRICAS	139
Capacidad de las métricas	139
13.0.1. Estudio de la capacidad de separación de las medidas clásicas del rendimiento	140
13.0.2. Estudio de la capacidad de separación de las medidas de complejidad de propósito general	140
14. CLASIFICACIÓN DE LOS GRUPOS DE ALUMNOS SEGÚN SU RENDIMIENTO	142

Clasificación de los grupos de alumnos según su rendimiento	142
14.1. Clasificación empleando las métricas clásicas de rendimiento	143
14.2. Clasificación empleando las medidas de complejidad de propósito general . .	143
14.2.1. Clasificación empleando los niveles 3, 4 y 5	143
14.2.2. Clasificación empleando los niveles 8, 9 y 10	144
14.3. Clasificación empleando una combinación de todas las métricas	144
14.3.1. Clasificación empleando los niveles 3, 4 y 5	145
14.3.2. Clasificación empleando los niveles 8, 9 y 10	146
VI. CONCLUSIONES Y VÍAS FUTURAS	147
15. CONCLUSIONES	148
Conclusiones	148
16. VÍAS FUTURAS	149
Vías futuras	149
Apéndice A. TABLAS DESCRIPTIVAS	150

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 planificados 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.

Adicionalmente, este proyecto ha inspirado la presentación del artículo “*In heaven as on earth: The performance of students is as good as it is the digraph that describes their behavior*” Vidal et al. (2023) al XXV International Symposium on Computers in Education (SIIE).

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.

Detectar grupos con dificultades para superar sus tareas de laboratorio es primordial para que el profesor pueda ayudarles y, cuanto antes, mejor. En este caso, el uso de técnicas de minería de procesos, comúnmente asociadas a indicadores de rendimiento de los alumnos, para extraer su comportamiento de un laboratorio virtual y la utilización de aprendizaje automático supervisado para identificar estos comportamientos han demostrado ser las herramientas fundamentales para identificar el comportamiento de los alumnos y prever grupos en riesgo. Así pues, en este estudio se demuestra que los indicadores de rendimiento de los alumnos pueden ser muy útiles, tanto como su comportamiento, estrictamente desde un punto de vista topológico.

0.3 OBJETIVOS

El principal objetivo de este proyecto es 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.

0.4 ESTRUCTURA DEL TRABAJO FIN DE GRADO

Este trabajo fin de grado consta de seis partes, 16 capítulos y otros elementos como la portada, la declaración de originalidad, la autorización para su ubicación en la biblioteca de la escuela, la autorización para su defensa, sendos resúmenes tanto en español como en inglés (con sus respectivas palabras clave), la sección de agradecimientos, el índice general así como de una bibliografía y un apéndice.

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

- **Parte I:** Motivaciones.
 - *Motivación*: Se trata de un capítulo inicial en el que se expone una introducción en el contexto del proyecto junto con las motivaciones existentes, los objetivos que se pretenden conseguir con la realización del mismo y la estructuración de la memoria.
- **Parte II:** Estado del arte.
 - *Capítulo 1*: Planteamiento del problema. Describe el funcionamiento de la plataforma educativa que recoge los datos que se usarán en este trabajo fin de grado y

la clase de tareas que se le plantean al alumnado que hace uso de la misma. Igualmente, se presenta un ejemplo fictio que refleja el tipo de información recogida de la actividad de los alumnos.

- **Capítulo 2:** Introducción a la Minería de Procesos y trabajos previos. Introducción al concepto clave de minería de procesos en relación a los objetivos que se plantean en este proyecto. Adicionalmente, se realiza una breve presentación de los estudios en este área previos a este trabajo y se realiza una clasificación de los trabajos realizados en el campo de la Minería de Procesos.
- **Capítulo 3:** Minería de Procesos. Breve descripción de la suite de minería de procesos Disco [Günther & Rozinat \(2012\)](#) y de su algoritmo subyacente (*Fuzzy Miner* [Günther & Aalst \(2007\)](#)). Igualmente, se muestran algunos diagramas extraídos con dicha herramienta y se exponen sus principales limitaciones.
- **Capítulo 4:** Implementación de la herramienta de Minería de Procesos *Graph Miner*. Breve introducción a la manera en la que se representarán los grupos de prácticas en la herramienta de minería de procesos de creación propia *Graph Miner*: la matriz característica de un grupo. Finalmente, se muestran los distintos gráficos que pueden obtenerse a partir de esta nueva herramienta.
- **Capítulo 5:** Teoría de grafos. Definiciones básicas relativas a la teoría de grafos entre las que se encuentra el concepto de *Grafo Dirigido Acíclico*, concepto clave en el desarrollo de este trabajo. Además, se incluye la exposición y demostración del Teorema de Kirchhoff, generalización de la fórmula de Cayley que será de gran utilidad en el cálculo del número de árboles de expansión de grafos conexos. El capítulo finaliza con la presentación de algunas medidas de propósito general que se usarán para realizar una clasificación de los distintos grupos de prácticas.
- **Parte III:** Análisis descriptivo.
 - **Capítulo 6:** Los registros existentes. Incluye una descripción de los registros existentes en el servidor. Se muestra el número de grupos estudiados cada año, el periodo de tiempo analizado cada año y el conjunto de problemas analizados cada año, estudiando la dificultad de los mismos. Por último, se realiza un análisis de la actividad registrada en el servidor.
 - **Capítulo 7:** Hipótesis de estudio. Exposición de las principales métricas de calidad definidas sobre los registros de actividad de los alumnos que podrían permitir discernir qué grupos presentan una mayor dificultad para resolver los problemas de prácticas propuestos. Se realiza una subdivisión de las mismas en continuas y a posteriori.
 - **Capítulo 8:** Rendimiento observado de los alumnos. Descripción detallada y estudio estadístico de las medidas de rendimiento de los alumnos clásicas. Se estudiarán las variaciones de las distintas variables a lo largo de los años y la normalidad de las mismas. El objetivo de este capítulo es la familiarización del lector con algu-

nas de las métricas de calidad que serán determinantes en la predicción de grupos en riesgo.

- **Parte IV:** Planificación del proyecto.

- [Capítulo 9](#): Etapas del proyecto: división en objetivos. Se describen las iteraciones en las que se divide el proyecto siguiendo la metodología *Scrum*.
- [Capítulo 10](#): Etapas del proyecto: división en sprints y seguimiento de los mismos. Se muestra la organización temporal del proyecto y el seguimiento del mismo mediante *burndown charts*.

- **Parte V:** Resultados obtenidos.

- [Capítulo 11](#): Análisis de las correlaciones entre las distintas métricas. Análisis de la correlación existente entre las métricas clásicas para cuantificar el rendimiento de los alumnos y la nota obtenida por los mismos. Además, se estudiarán las correlaciones entre las medidas de complejidad (características topológicas de los grafos que representan el proceso de resolución de los problemas) y la nota alcanzada por los alumnos al final de la práctica.
- [Capítulo 12](#): Perfiles de estudiantes según su rendimiento. Aplicación de técnicas de clustering para agrupar a los alumnos en diferentes intervalos de calificaciones empleando diferentes métricas (medidas clásicas de rendimiento y topológicas).
- [Capítulo 13](#): Capacidad de las métricas. Estudio de las diferentes métricas respecto a su capacidad de separación entre alumnos en riesgo de obtener peores resultados y los que no presentarán dificultades para superar los problemas propuestos.
- [Capítulo 14](#): Clasificación de los grupos de alumnos según su rendimiento. Utilización del clasificador estadístico C5.0 para generar un conjunto de reglas que se emplearán para poder decidir si el rendimiento de una determinada agrupación de estudiantes es más bajo de lo que debería. Además, se utilizará otro clasificador del mismo tipo para predecir el intervalo de notas en el que se encontraría un determinado grupo de alumnos.

- **Parte VI:** Conclusiones y vías futuras.

- [Capítulo 14](#): Conclusiones. Repaso de los logros alcanzados a lo largo del proyecto y recapitulación de la principal motivación del desarrollo de este trabajo.
- [Capítulo 15](#): Vías futuras. Exposición de un conjunto de propuestas para ampliar el trabajo aquí realizado.

Parte II

ESTADO DEL ARTE

Planteamiento del problema y Minería de Procesos.

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.

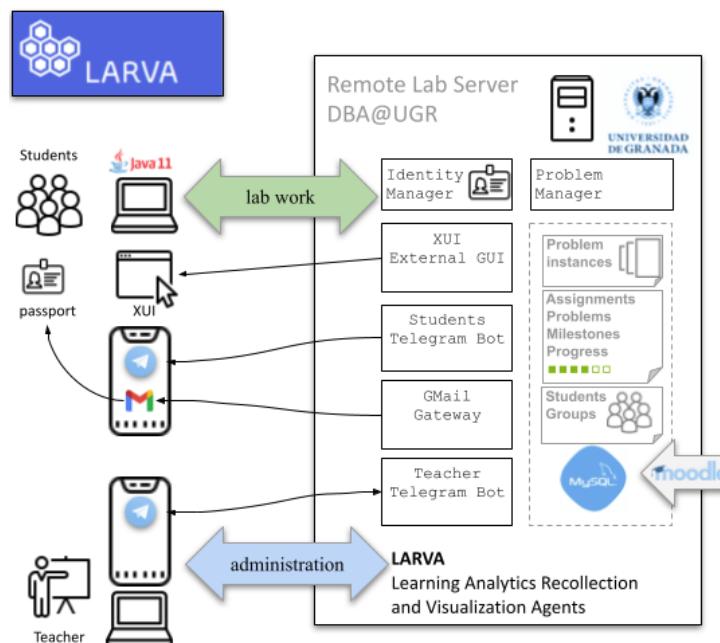


Figura 1: Arquitectura del Servidor Remoto. Por un lado, contiene el laboratorio virtual para sistemas multiagente distribuidos. Además, los alumnos también pueden consultar su progreso y el de sus compañeros a través de un Bot de Telegram. Por otro lado, el profesor también puede conocer el número de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

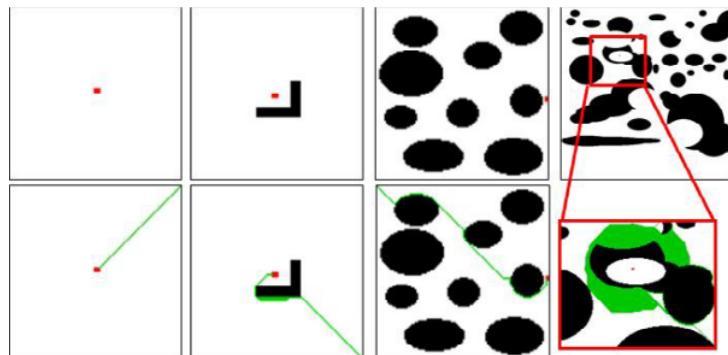


Figura 2: Alguno de los mapas que el alumnado debe resolver. Los agentes de los grupos de estudiantes deben acceder a uno de esos mundos y deben alcanzar los objetivos (coloreados en rojo) navengando a través del mundo y evitando los obstáculos (coloreados en negro). Alguno de los mundos no son resolubles 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 5×5 (observar Figura 3b).
- Un **sensor escáner** que actúa como *detector del objetivo* e indica al agente la distancia al mismo medida desde cada una las celdas de su entorno 5×5 (observar Figura 3c).

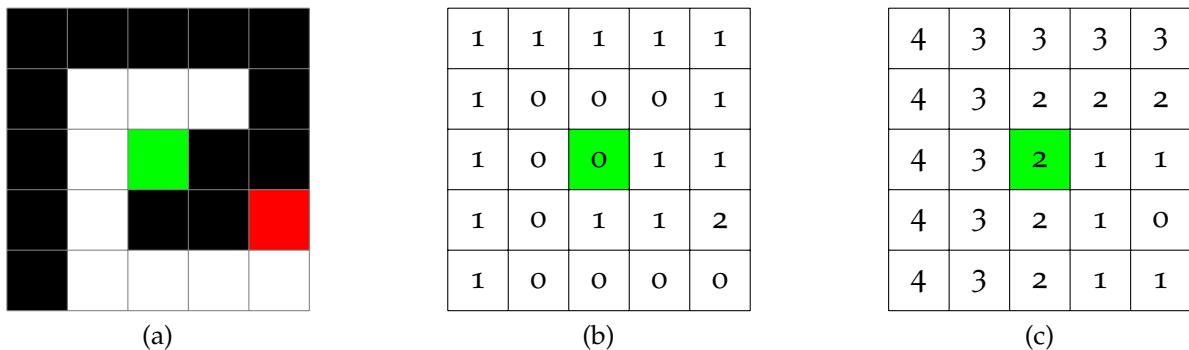


Figura 3: Un agente (representado por una celda verde en el centro de cada figura) tiene una percepción local de su entorno: solamente percibe el entorno 5×5 de celdas colindantes. El Radar 3b muestra dicho entorno 5×5 que rodea al agente e informa de si una celda está vacía (valor 0), de si hay un obstáculo (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 sale del mundo virtual.
- REFUEL. El agente recarga completamente su batería. A los agentes se les permite recargar su batería tantas veces como deseen.

1.1 FUNCIONAMIENTO DEL LABORATORIO VIRTUAL

Supongamos que una persona tiene que completar una determinada tarea con nueve pasos o milestones diferentes, numerados del 1 al 9, de los cuales los pasos 3, 6 y 9 tienen una recompensa. Esta persona podría intentar pasar por cada una de las subtareas tantas veces como considere oportuno para conseguir todas las recompensas, llevando a cabo un registro de su actividad. Por ejemplo, la Figura 4 muestra uno de estos registros. Ignorando el paso 1, que sólo se utiliza para marcar el inicio del registro, esta persona ha realizado 35 pasos, repitiendo el paso 2 siete veces, el paso 3 seis veces, el paso 4 dos veces, el paso 5 seis veces, el paso 6 cinco veces, el paso 7 cuatro veces, el paso 8 cuatro veces y el paso 9 sólo una vez. Este comportamiento puede representarse con un grafo dirigido ponderado (cíclico) donde los nodos representan los pasos dados y las aristas se ponderan con la frecuencia detectada en el registro (Figura 4), de modo que el número de veces que se ejecuta un paso viene dado

por la suma de los pesos de sus aristas entrantes, lo cual denominaremos factor de grado de entrada (*in-degree*) de aquí en adelante.

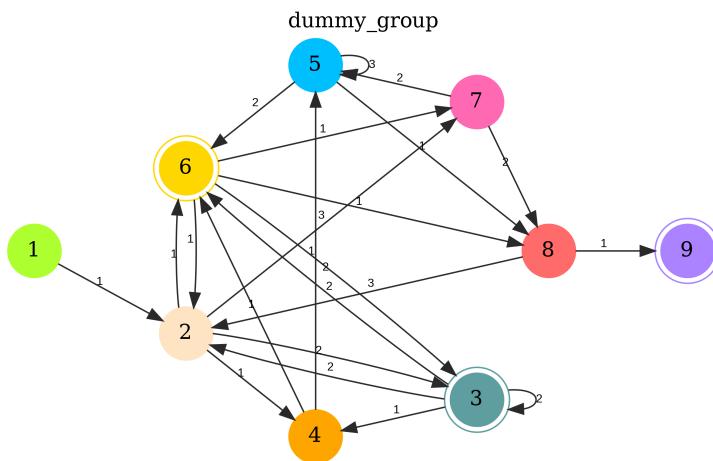


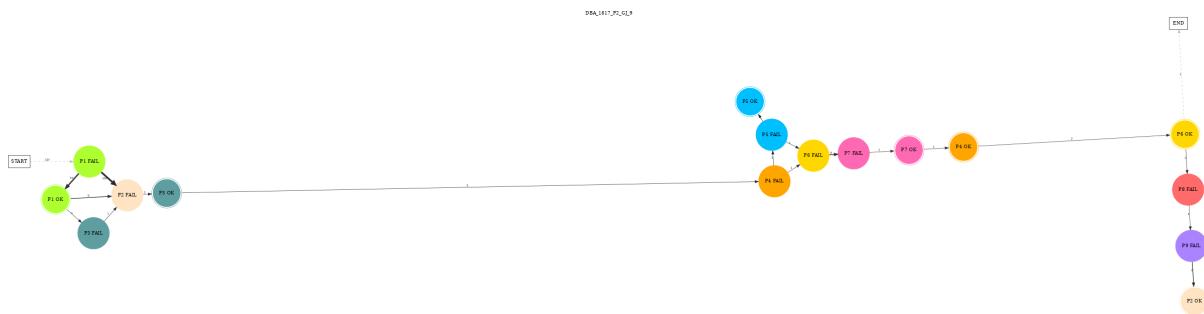
Figura 4: La secuencia correspondiente al grafo es: 1 2 (3) 2 4 5 5 (6) (3) (3) 4 (6) 7 5 5 5 8 2 (6) (3) 2 7 5 (6) 8 2 (3) (3) (6) 2 7 8 2 7 8 (9). Así pues, el gráfico muestra el comportamiento básico de un determinado grupo en el que se repiten 9 pasos, tres de los cuales, representados entre paréntesis, tienen una recompensa.

Este registro lineal de actividades refleja exactamente cómo se almacena la realización de tareas de laboratorio en LARVA [Vidal & Sánchez-Carrascosa \(2022\)](#), [Vidal & Sánchez-Carrascosa \(2023\)](#), un laboratorio virtual [Vidal \(2016\)](#) diseñado con un único propósito: permitir a los estudiantes alcanzar su mayor éxito. Para ello, LARVA mantiene un registro completo de la actividad de los alumnos, por lo que su evaluación se basa no sólo en los objetivos alcanzados, sino también en su progreso. Además, cuenta con un sistema de retroalimentación multimodal [Vidal & Sánchez-Carrascosa \(2022\)](#) para mantener a los estudiantes informados sobre su progreso, en tiempo real, gracias a mensajes de chat de Telegram enviados directamente a sus teléfonos móviles. Puede demostrarse que cuanto antes reciban el feedback sobre su actividad y cuanto más rica sea esta retroalimentación, mayor será la autorregulación y la eficacia de la experiencia de aprendizaje [Keller \(1968\)](#).

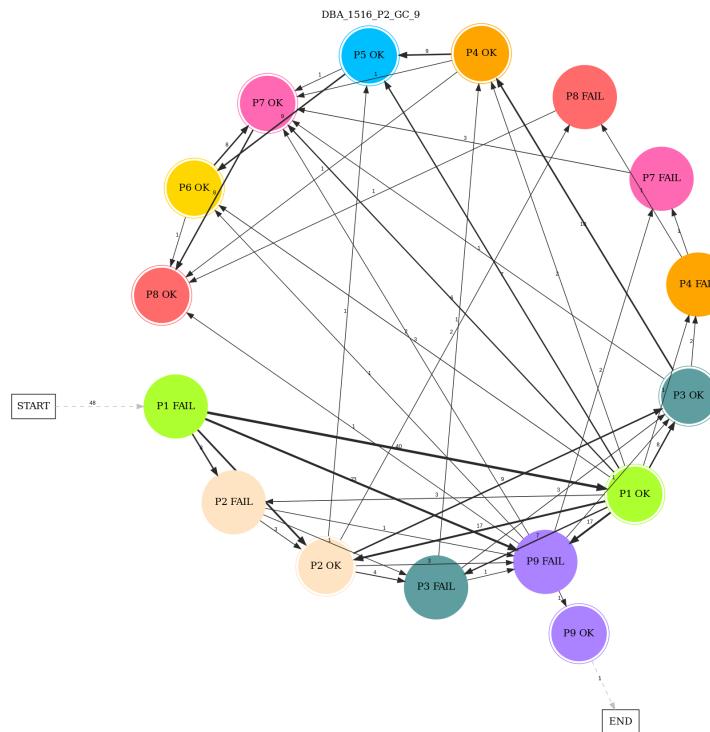
Revelar este comportamiento también puede ser útil para el profesor para detectar, cuanto antes, posibles dificultades de los alumnos para completar sus tareas y permitir una intervención clave por parte del profesor. Pero, ¿cómo detectar estas desviaciones del rendimiento esperado y con qué antelación podrían detectarse? Además, ¿sería posible ignorar cualquier detalle sobre los indicadores de progreso habituales asociados al rendimiento, como el momento de los éxitos y fracasos, la perseverancia, etc., para no depender demasiado en la “eficiencia clásica”, y centrarse en las características topológicas del comportamiento de los alumnos? Para principiantes, en la Figura 5 se muestran dos comportamientos extremos. Por un lado, en la Figura 5a, después de 372 sesiones, la mayoría de ellas quemadas en los 3-4 preliminares pasos, acaba con muy pocas sesiones en los últimos problemas, una especie de

piloto automático, y resuelve 7 de cada 9 problemas. Por otro lado, en la Figura 5b, después de 297 sesiones (¿podría considerarse como un menor esfuerzo?) muestra una exploración bastante exhaustiva de las alternativas y termina con 9 de 9 problemas resueltos. Este documento responde con éxito a todas estas preguntas a partir de un sólido análisis basado en evidencias de los registros de los últimos siete años de este laboratorio virtual. Las siguientes secciones están dedicadas a discutir trabajos similares en la literatura, a presentar el escenario y la hipótesis principal y, a continuación, a extraer las principales conclusiones tras un análisis exhaustivo de los datos registrados. Todos los conjuntos de datos mencionados en este documento y todos los artefactos de software, completamente escritos en R, están abiertos y disponibles en GitHub¹.

¹ <https://github.com/maribeloo/Analysis-of-processes>



(a) Grafo acíclico dirigido que captura el comportamiento de un grupo con 372 sesiones de trabajo.



(b) Grafo acíclico dirigido que captura el comportamiento de un grupo con 297 sesiones de trabajo.

Figura 5: Dos comportamientos reales diferentes de grupos enfrentándose a las mismas tareas de laboratorio.

1.2 DESCRIPCIÓN DEL ESCENARIO INICIAL

Como se ha mencionado anteriormente, este estudio se ha llevado a cabo a partir de los datos de una asignatura obligatoria de 4º curso de la rama Ingeniería del Software en la Universidad de Granada registrados a través de un laboratorio virtual Vidal (2016). Esta asignatura sigue la estructura típica del Sistema Personalizado de Instrucción Keller (1968) donde el progreso de los estudiantes se monitorea continuamente gracias a un sistema de hitos y logros continuos, proporcionando tanta retroalimentación como sea posible y tan pronto como sea posible. Así,

durante el curso, los alumnos se organizan en grupos (de 4-5 miembros cada uno) y cada grupo debe resolver una serie de tareas o problemas. De ahora en adelante denotaremos al conjunto de problemas por $P = \{p_i\}$ y al número de problemas a resolver por $n = |P|$.

Por lo general, el número de problemas es $n = 9$ y podrían considerarse como un conjunto de misiones a realizar en escenarios similares. Adicionalmente, cada problema ha sido cuidadosamente elaborado por el profesor para que requiera un nivel de dificultad creciente por parte de los estudiantes de tal forma que la estrategia para resolver p_i puede no ser lo suficientemente buena resolver p_{i+1} pero la estrategia para resolver p_{i+1} también debe ser válida en p_i . Estos problemas están abiertos durante un determinado periodo de tiempo durante el cual los estudiantes pueden abrir los problemas tantas veces como necesiten, ya sea para intentar resolver el problema, para mejorar sus soluciones en estos problemas o para probar nuevas estrategias. Además de esto, cada problema p_i se ha diseñado como una secuencia de 5 hitos consecutivos, también en nivel creciente de dificultad:

$$p_i = \{p_i^1, p_i^2, p_i^3, p_i^4, p_i^5\} \quad (1)$$

siendo p_i^1 el suceso de sólo abrir el problema i y p_i^5 la obtención de una solución válida para p_i . Estos hitos deben ser alcanzados progresivamente por los alumnos, por lo que todo acaba por empujando a los alumnos un poco hacia adelante en sus capacidades Keller (1968). Por lo tanto, a medida que los estudiantes progresan en el laboratorio, dejan un registro de sus logros, que se llamará *comportamiento*, B^g , de un determinado grupo g , y está compuesto por una secuencia de sesiones.

$$B^g = \{{}^s p_i^k, \dots\} \quad (2)$$

Cada *sesión* del laboratorio virtual ${}^s p_i^k$ se etiqueta como el mayor hito k alcanzado en cada problema p_i y un número natural secuencial s que es un registro del tiempo actual del sistema. Además, se utiliza una marca A para señalar el inicio del de laboratorio y un indicador Z para señalar el cierre del mismo.

$$B^g = \{{}^0 A\} \cup \{{}^s p_i^k, \dots\} \cup \{Z\} \quad (3)$$

Cuadro 1: El comportamiento de un grupo ficticio con 20 sesiones de trabajo en el laboratorio virtual.

${}^0 A$	${}^1 p_1^1$	${}^2 p_1^3$	${}^3 p_1^5$	${}^4 p_1^5$	${}^5 p_1^5$	${}^6 p_1^4$	${}^7 p_2^3$	${}^8 p_2^4$	${}^9 p_1^5$	${}^{10} p_1^5$	Z
${}^{11} p_2^5$	${}^{12} p_1^1$	${}^{13} p_2^5$	${}^{14} p_3^2$	${}^{15} p_3^4$	${}^{16} p_1^2$	${}^{17} p_3^3$	${}^{18} p_2^3$	${}^{19} p_3^4$	${}^{20} p_3^5$		

Por ejemplo, en un escenario con tres problemas, la secuencia mostrada en la Tabla 1 podría ser un comportamiento factible B , que ya se introdujo en la Sección 1.1, y ahora se explica bajo una nueva perspectiva:

"Los estudiantes se han conectado 20 veces al laboratorio virtual. En la primera sesión se abre el problema p_1 sin más éxito y en la segunda sesión se deja el problema p_1 a medias. No obstante, en la tercera sesión se resuelve completamente p_1 . Los estudiantes ponen en práctica nuevas estrategias y resuelven p_1 dos veces más. A continuación, introducen un nuevo cambio que casi resuelve p_1 , acaba fallando y deja p_2 justo en la mitad. Después, tras algunos cambios, resuelven de nuevo p_1 en la sesión 9 y p_2 en la sesión 11. Después, cambiaron algo en la implementación que resultó ser un fracaso en p_1 y acaban teniendo éxito en p_3 justo al final".

La Figura 32 y Tablas 4 y 9 muestran información descriptiva de los datos registrados cada año sobre el comportamiento de 77 grupos (en total casi 400 alumnos). Como se puede ver, hay registros de 7 años consecutivos y un total de 30672 sesiones de trabajo en el laboratorio virtual. A continuación, el reto es cómo extraer un gráfico como los que se muestran en la Sección 1.1 y cómo utilizarlos para detectar, lo antes posible, los grupos con dificultades para progresar.

2

INTRODUCCIÓN A LA MINERÍA DE PROCESOS Y TRABAJOS PREVIOS

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 cada 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). Así pues, la minería de procesos para la educación es como los rayos X para la medicina: hacen visible lo invisible, en aras de un conocimiento mucho más preciso de la situación y de la planificación de nuevas intervenciones. Pero, ¿hasta qué punto es pertinente atenerse a un gráfico extraído de registros de eventos en lugar de, por ejemplo, una representación tabular equivalente? Este estudio se concibió bajo esta perspectiva y analiza el comportamiento de los alumnos no como en la forma clásica de producir restricciones temporales, de rendimiento o de cualquier semántica asociada a su rendimiento o a sus notas. Por el contrario, este trabajo analiza justo la topología pura y desnuda del comportamiento y proporciona pruebas sólidas de que estas topologías están, cómo no, también correlacionadas con el rendimiento, en particular, para la detección de grupos de estudiantes de perfil bajo.

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 numerosas aplicaciones en el área sanitaria, en el ámbito empresarial, institucional etc.

Como se ha mencionado anteriormente, el uso de técnicas de Minería de Procesos está suscitando un enorme interés y hay cantidades ingentes de artículos y trabajos relacionados que podrían referenciarse, especialmente cuando los profesores crean laboratorios virtuales Elmoazen et al. (2023) o servicios *online* como Coursera Mukala et al. (2015). También hay algunas revisiones excelentes de la Minería de Procesos en el contexto educativo como dos Santos Garcia et al. (2019), donde se recalca que la educación es la cuarta área más importante en la que se están aplicando técnicas de Minería de Procesos para detectar estilos de aprendizaje (o, en inglés, *learning styles*), o incluso mejor, en Bogarín et al. (2018) donde los autores también ponen el énfasis en el análisis de grafos como estructura subyacente del comportamiento (concepto de *graph mining*). Además, en dicho artículo se presentará lo que se denomina *intention mining* que hace referencia a la minería de procesos aplicada con el objetivo de anticipar el comportamiento. Los mismos autores en Bogarín Vega et al. (2018) también apuntan a una cuestión muy interesante e intimamente relacionada con este trabajo: cómo utilizar técnicas de Minería de Procesos para detectar a los que más necesitan la ayuda del profesor, simplemente observando el patrón de interacción de los alumnos con el LMS (*Learning Management System*). Esta relación del alumno con los recursos disponibles, normalmente en línea, también se explora en Mukala et al. (2015), en el que el uso de la Minería de Procesos no es realmente descubrimiento, sino el cumplimiento, para detectar si los alumnos se comportan como se espera de ellos o en la forma en que utilizan los recursos disponibles Juhaňák et al. (2019). En Sedrakyan et al. (2016) los autores analizan el potencial del feedback que el PM (*Process Mining*) puede dar a los alumnos, muy en la línea de (Keller, 1968) para fomentar la autorregulación de los alumnos. En todos ellos las principales métricas utilizadas para orientar el estudio son las “clásicas” relacionadas con los recursos, las personas o el material de aprendizaje a través de un enfoque cuantitativo para puntuar el comportamiento de los estudiantes. Este trabajo fin de grado adoptará un enfoque diferente e intenta analizar puramente la topología del grafo dirigido que describe el comportamiento de un determinado grupo de prácticas.



Figura 6: Los nodos de los grafos pueden representar recursos o herramientas, personas o instituciones e incluso unidades temáticas. Asimismo, las relaciones entre nodos también llevan asociadas un significado (energía o frecuencia, proximidad o distancia entre ambos o precedencia u orden entre ambos).

Los trabajos anteriormente citados así como los que se realizarán en este trabajo fin de grado relacionados con la Minería de Procesos son susceptibles a ser clasificados. Así pues, dado un grafo $G = (V, E)$ donde V y E son sus respectivos conjuntos de vértices y aristas (Definición 1) siendo una parte del mismo la que se muestra en la Figura 6, se puede realizar una segmentación de los proyectos de minería de procesos atendiendo a qué representan sus nodos, qué significado tiene la relación entre éstos y el momento temporal al que nos estemos refiriendo. Así pues, atendiendo a todas estas características se ha realizado la clasificación de las técnicas de Minería de Procesos en el ámbito educativo que se muestra en la Tabla 2.

Cuadro 2: Clasificación de las técnicas de Minería de Procesos aplicadas a la educación. Como podemos ver, pueden estar orientadas al pasado, al presente y al futuro.

Nodos	Aristas	Energía (Conteo)	Cercanía (Clustering)	Orden (Precedencia)																		
Recursos (vídeos, texto) y Herramientas		<table border="1"> <tr> <td>Pasado</td> <td>Coursera. Validación del curso. Uso de recursos (como podían ser los videos).</td> </tr> <tr> <td>Presente</td> <td>Patrones de uso. Seguimiento. Capacidad.</td> </tr> <tr> <td>Futuro</td> <td>Cuellos de botella (predicción de las necesidades del hardware). Posibles problemas. Recomendaciones.</td> </tr> </table>	Pasado	Coursera. Validación del curso. Uso de recursos (como podían ser los videos).	Presente	Patrones de uso. Seguimiento. Capacidad.	Futuro	Cuellos de botella (predicción de las necesidades del hardware). Posibles problemas. Recomendaciones.	<table border="1"> <tr> <td>Pasado</td> <td></td> </tr> <tr> <td>Presente</td> <td></td> </tr> <tr> <td>Futuro</td> <td></td> </tr> </table>	Pasado		Presente		Futuro		<table border="1"> <tr> <td>Pasado</td> <td></td> </tr> <tr> <td>Presente</td> <td></td> </tr> <tr> <td>Futuro</td> <td></td> </tr> </table>	Pasado		Presente		Futuro	
Pasado	Coursera. Validación del curso. Uso de recursos (como podían ser los videos).																					
Presente	Patrones de uso. Seguimiento. Capacidad.																					
Futuro	Cuellos de botella (predicción de las necesidades del hardware). Posibles problemas. Recomendaciones.																					
Pasado																						
Presente																						
Futuro																						
Pasado																						
Presente																						
Futuro																						
Personas e Instituciones		<table border="1"> <tr> <td>Pasado</td> <td>Trabajo en grupo bien estructurado. Análisis de los alumnos con mejor/peor rendimiento.</td> </tr> <tr> <td>Presente</td> <td>Medidas intragrupos e intergrupos. Redes sociales. Evaluación de los alumnos.</td> </tr> <tr> <td>Futuro</td> <td>Liderazgo. Dominancia. Personas influyentes. Predicciones del rendimiento de los alumnos.</td> </tr> </table>	Pasado	Trabajo en grupo bien estructurado. Análisis de los alumnos con mejor/peor rendimiento.	Presente	Medidas intragrupos e intergrupos. Redes sociales. Evaluación de los alumnos.	Futuro	Liderazgo. Dominancia. Personas influyentes. Predicciones del rendimiento de los alumnos.	<table border="1"> <tr> <td>Pasado</td> <td>Trabajo en equipo.</td> </tr> <tr> <td>Presente</td> <td></td> </tr> <tr> <td>Futuro</td> <td>Redes sociales: personas influyentes y colaboradores.</td> </tr> </table>	Pasado	Trabajo en equipo.	Presente		Futuro	Redes sociales: personas influyentes y colaboradores.	<table border="1"> <tr> <td>Pasado</td> <td></td> </tr> <tr> <td>Presente</td> <td></td> </tr> <tr> <td>Futuro</td> <td></td> </tr> </table>	Pasado		Presente		Futuro	
Pasado	Trabajo en grupo bien estructurado. Análisis de los alumnos con mejor/peor rendimiento.																					
Presente	Medidas intragrupos e intergrupos. Redes sociales. Evaluación de los alumnos.																					
Futuro	Liderazgo. Dominancia. Personas influyentes. Predicciones del rendimiento de los alumnos.																					
Pasado	Trabajo en equipo.																					
Presente																						
Futuro	Redes sociales: personas influyentes y colaboradores.																					
Pasado																						
Presente																						
Futuro																						
Temáticas (unidades temáticas de un curso), Productos o bienes		<table border="1"> <tr> <td>Pasado</td> <td>Validación del curso. Concienciación del profesorado.</td> </tr> <tr> <td>Presente</td> <td>Relaciones intergrupos. Redes sociales.</td> </tr> <tr> <td>Futuro</td> <td></td> </tr> </table>	Pasado	Validación del curso. Concienciación del profesorado.	Presente	Relaciones intergrupos. Redes sociales.	Futuro		<table border="1"> <tr> <td>Pasado</td> <td></td> </tr> <tr> <td>Presente</td> <td></td> </tr> <tr> <td>Futuro</td> <td></td> </tr> </table>	Pasado		Presente		Futuro		<table border="1"> <tr> <td>Pasado</td> <td>Similaridad del coseno.</td> </tr> <tr> <td>Presente</td> <td></td> </tr> <tr> <td>Futuro</td> <td>Planificación del curso.</td> </tr> </table>	Pasado	Similaridad del coseno.	Presente		Futuro	Planificación del curso.
Pasado	Validación del curso. Concienciación del profesorado.																					
Presente	Relaciones intergrupos. Redes sociales.																					
Futuro																						
Pasado																						
Presente																						
Futuro																						
Pasado	Similaridad del coseno.																					
Presente																						
Futuro	Planificación del curso.																					

3

MINERÍA DE PROCESOS

Como se ha comentado antes, cualquier posible evaluación de los alumnos basada meramente en los objetivos alcanzados es sólo una visión estática, sin muchas pistas sobre la dinámica que les ha llevado hasta ese punto. No sabemos muy bien cómo han llegado los alumnos hasta ahí y esta información es la clave para identificar a los grupos con dificultades que podrían necesitar la ayuda del profesor. Por lo tanto, vamos a emplear técnicas de Minería de Procesos [van der Aalst \(2016\)](#) para extraer las diferentes secuencias de eventos correlacionados del conjunto de datos original.

3.1 EXTRACCIÓN DE LOS PROCESOS CON DISCO

Para la extracción de los procesos ocultos se empleará la muy conocida suite de minería de procesos Disco [Günther & Rozinat \(2012\)](#). Disco es una herramienta que permite crear mapas visuales a partir de los registros en cuestión de minutos.

Fuzzy Miner [Günther & Aalst \(2007\)](#) es un algoritmo de minería de procesos muy flexible y sólido que está detrás de Disco. En un proceso típico minado con Disco, cada actividad del proceso se etiqueta como en la Tabla 1 e incluye las frecuencias tanto de las actividades como de las transiciones entre actividades. Por lo tanto, la frecuencia de cada actividad representa el número de veces que esta actividad aparece en B^g . Por otro lado, la frecuencia de cada transición entre las actividades x e y representa el número de veces que la actividad x aparece inmediatamente antes de la actividad y en B^g .

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 única 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 milestone y el identificador del caso son las sesiones. Las Figuras 7 y 8 muestran los diagramas obtenidos en cada uno de los casos.

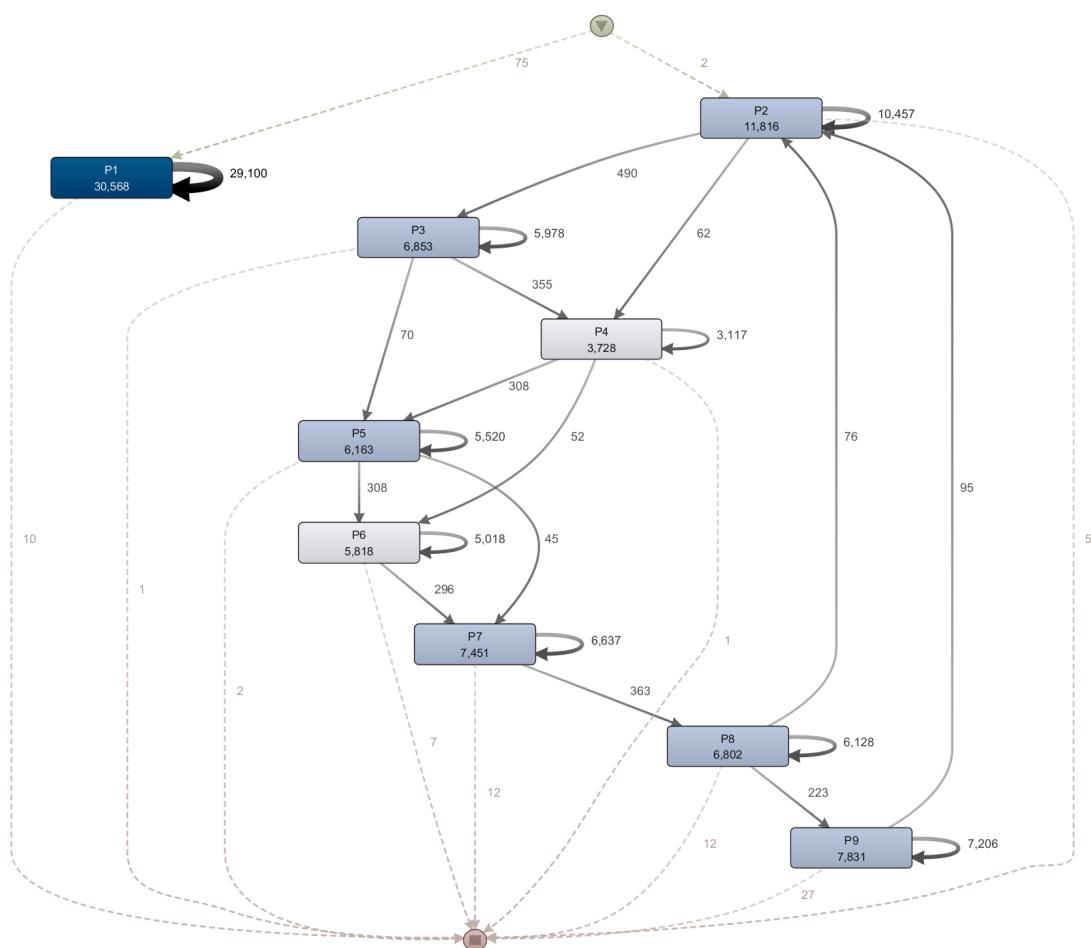


Figura 7: Análisis de procesos del dataset (Activity problema y CaseId grupo). Contiene el 100 % de las actividades y el 80 % de los caminos.

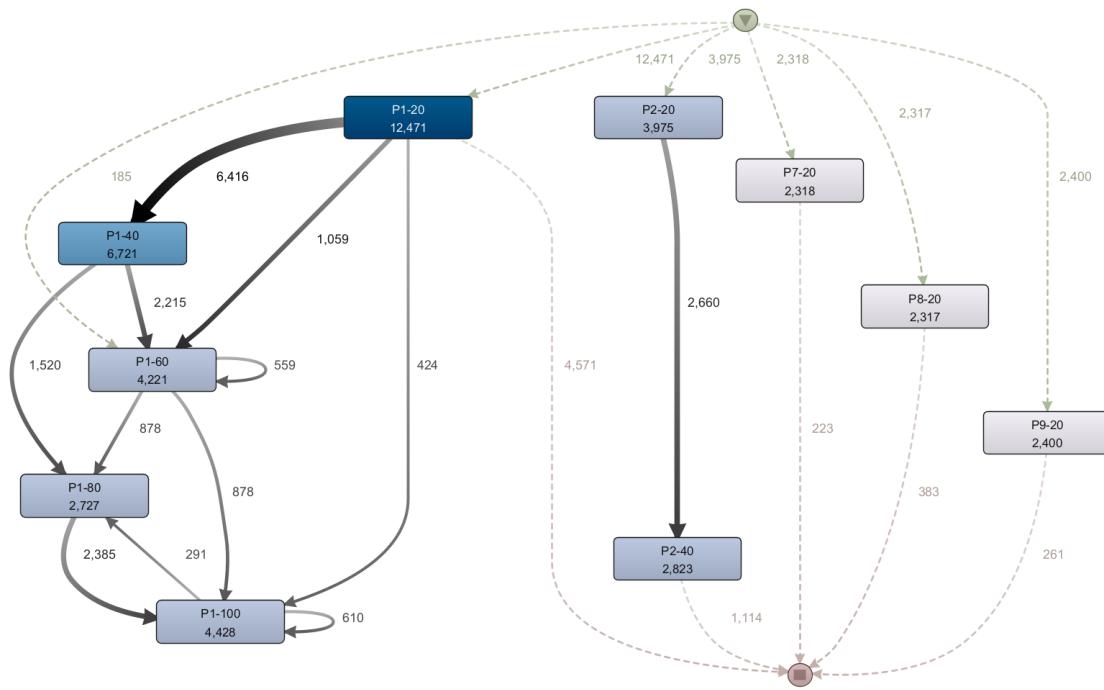


Figura 8: Análisis de procesos del dataset (Activity problema-milestone y CaseId sesión). Contiene el 20 % de las actividades y el 20 % de los caminos.

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 un grupo está en riesgo de obtener un rendimiento peor del esperado. 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 9 y 10.

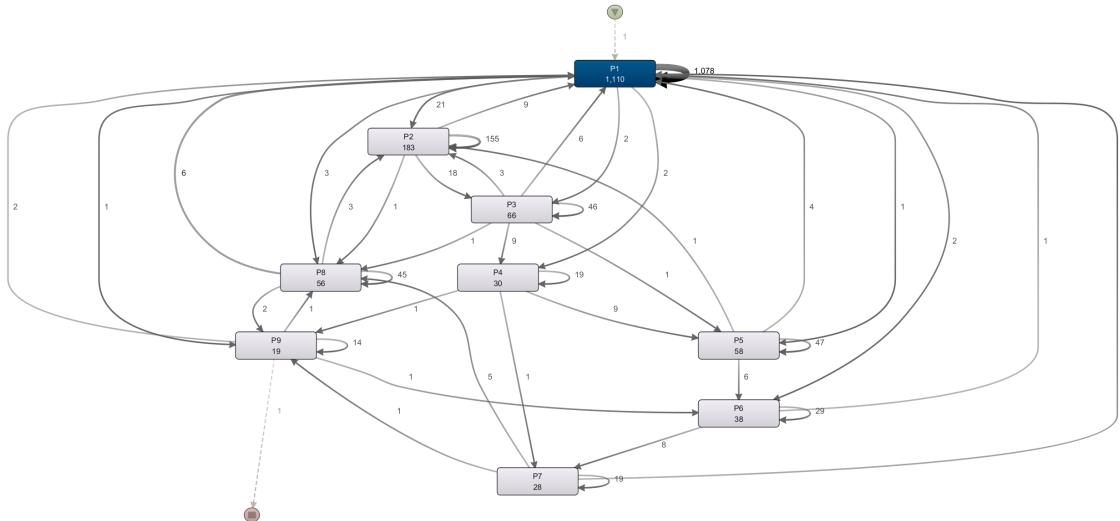


Figura 9: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema y CaseId grupo). Contiene el 100 % de las actividades y el 100 % de los caminos.

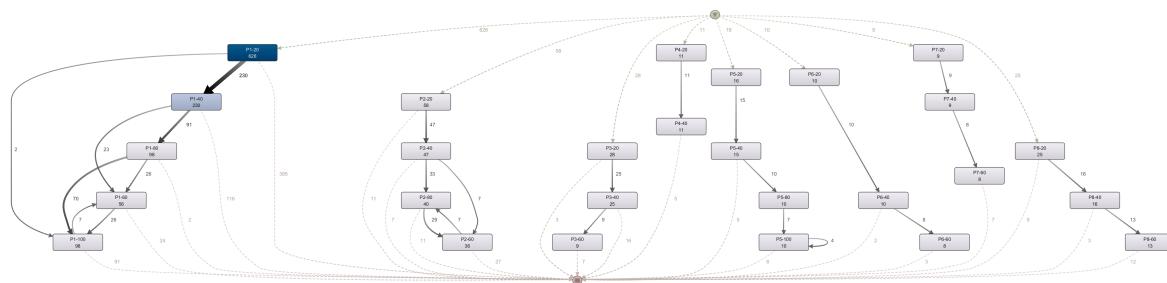


Figura 10: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema-milestone y CaseId sesión). Contiene el 60 % de las actividades y el 80 % de los caminos.

Aunque Disco tiene mucho éxito y se pueden personalizar métricas e interpretaciones del conjunto de datos fuente, no se ajusta completamente a nuestros requerimientos para desvelar la estructura del comportamiento. 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 cambiándolo 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 24 y 25), es inviable seguir usando el programa.

Así pues, en este trabajo fin de grado se ha implementado nuestra propia versión del programa, la herramienta de minería de procesos *Graph Miner* en R, personalizada y adaptada a las

necesidades del problema, basándose principalmente en la estructura principal de FuzzyMiner [Günther & Aalst \(2007\)](#) con algunas adiciones menores.

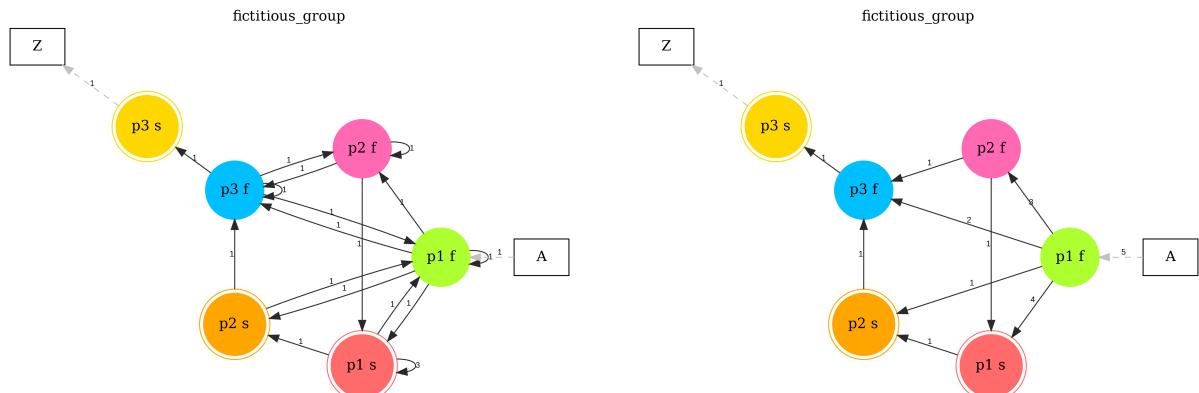
4

IMPLEMENTACIÓN DE LA HERRAMIENTA DE MINERÍA DE PROCESOS *GRAPH MINER*

4.1 USO DE GRAFOS COMO REPRESENTACIÓN DE LOS GRUPOS

Para dar una estructura de orden mínimo al comportamiento B^g , suponemos que es cierto lo siguiente: el resultado de la sesión i depende en su mayor parte de lo ocurrido en la sesión $i - 1$ o, como mucho, de cualquiera de las sesiones inmediatamente anteriores. Si aceptamos esta suposición, entonces B^g tiene una estructura de orden parcial en la que cada sesión aparece conectada a la anterior.

Además, podríamos etiquetar estas relaciones con un número natural que indica el número de veces que esta relación se ha dado en B^g . De este modo se, obtiene el gráfico de la Figura 11a, un grafo conexo (Definición 7) que describe el camino seguido por los estudiantes con un sentido claro de las transiciones de un estado a otro. Al igual que antes, también tenemos que la suma de las aristas entrantes a un nodo (o grado de entrada de dicho nodo) coincide con el número de veces que el nodo aparece en el registro. Nótese que como este estudio se centra en el éxito y el fracaso de los alumnos, el grafo de la Figura 11a es un grafo colapsado en el que sólo se distinguen dos tipos de sesiones: las que fracasan, es decir, las que alcanzan los hitos 1 a 4, y las que tienen éxito, es decir, las que alcanzan el hito final 5. Así pues, notaremos a las sesiones fallidas por ${}^n p_i^f$ y a las sesiones exitosas por ${}^n p_i^s$, con $n \in \mathbb{N}$, como en la Figura 11, que continúa el ejemplo presentado en la Sección 1.2. Además, la matriz de adyacencia del segundo grafo puede verse en la Ecuación 4.



(a) Grafo cíclico dirigido que captura las relaciones de precedencia entre las sesiones de la Tabla 1 (20 sesiones) obtenido colapsando los vértices de la misma en exitosas o fallidas. Los nodos con doble círculo representan sesiones en las que se ha resuelto un problema.

(b) El mismo grafo que el de la Figura 11a pero eliminando los ciclos y manteniendo el grado de entrada de cada uno de los vértices.

Figura 11: Grafos resultantes de la continuación del ejemplo de la Sección 1.2.

$$\mathcal{A}^g = \left(\begin{array}{c|ccccccccc} & A & p_1^f & p_1^s & p_2^f & p_2^s & p_3^f & p_3^s & Z \\ \hline A & \mathbf{0} & 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_1^f & 0 & \mathbf{0} & 4 & 3 & 1 & 2 & 0 & 0 \\ p_1^s & 0 & 0 & \mathbf{0} & 0 & 1 & 0 & 0 & 0 \\ p_2^f & 0 & 0 & 1 & \mathbf{0} & 0 & 1 & 0 & 0 \\ p_2^s & 0 & 0 & 0 & 0 & \mathbf{0} & 1 & 0 & 0 \\ p_3^f & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & 1 & 0 \\ p_3^s & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & 1 \\ Z & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} \end{array} \right) \quad (4)$$

Esta perspectiva nos permite ver B^g como un grafo dirigido $G = (V^g, E^g)$ que capta la dinámica de un grupo mientras se esfuerza por resolver los distintos problemas que se le han planteado. Por lo tanto, el conjunto de eventos (problemas y estado de los mismos más un evento de inicio y otro de finalización) que los estudiantes han realizado es el conjunto de vértices del grafo:

$$V^g = \{ {}^0 A \} \cup \{ {}^s p_i^k \in B^g \} \cup \{ Z \} \quad (5)$$

y cada par de eventos consecutivos se considera una arista del grafo:

$$E^g = \left\{ \langle x, y \rangle, x = {}^s p_i^r, y = {}^t p_j^l, t = s + 1, x, y \in B^g \right\} \quad (6)$$

que, de hecho, puede representarse como su matriz de adyacencia ponderada y nos permite representar el invariante grado de entrada.

4.2 LA MATRIZ CARACTERÍSTICA DE UN GRUPO

Este grafo sigue siendo un grafo cíclico (Definición 16) y, como estamos interesados en la representación más esencial del comportamiento de los estudiantes, se eliminarán esos ciclos quitando una cantidad mínima de aristas y conservando invariante el grado de entrada, obteniendo el grafo esencial de la Figura 11b, con su respectiva matriz de distancias (Ecuación 4), que denominamos matriz característica (Definición 10) del grupo g , \mathcal{A}^g . Esta matriz característica es una representación minimal de B^g y a partir de ella se pueden extraer numerosas métricas, siendo la más importante una medida de la complejidad inherente a B^g .

Por lo tanto, $\mathcal{A}^g[x, y] = d > 0$ significaría que x precede a y d veces en B^g . Así pues, cuanto mayor sea d , mayor será la influencia de x sobre y en términos del comportamiento codificado en \mathcal{A}^g . En aras de la simplicidad, además del uso de valores enteros para obtener una fila o columna, también se utilizarán los nombres identificativos de los nodos como, por ejemplo, se muestra en la Ecuación 7.

$$\mathcal{A}[1, 2] = \mathcal{A}[p_2^f, p_3^f] = 4 \quad (7)$$

Podría decirse que \mathcal{A}^g contiene lo que podría haber sido la *experiencia* de resolver todos los problemas, es decir, una especie de huella que codifica la relación entre fracaso y éxito y la fuerza de estas relaciones. Vale la pena decir que en todos los registros (Tabla 9) todas estas matrices son diferentes entre sí. Se derivarán otras dos matrices de \mathcal{A}^g . En primer lugar, obtendremos una matriz de adyacencia característica integrada únicamente por ceros y unos a la que denominaremos \mathcal{A}' . Así pues, $\mathcal{A}'[r, c]$ será 1 únicamente cuando $\mathcal{A}[r, c] > 0$. En segundo lugar, calcularemos la matriz de distancias mínimas que se obtendrá mediante la aplicación del algoritmo de caminos mínimos de Dijkstra, de modo que $\hat{\mathcal{A}}[r, c]$ contiene una especie de longitud de camino mínima entre los diferentes nodos del grafo en función del número de sesiones. Por tanto, estas estructuras son un punto de partida ideal desde el que decodificar la información sobre las experiencias de aprendizaje. En particular, nuestro principal interés será determinar si existe una relación entre esta estructura y el éxito o el fracaso de cada uno de los grupos de alumnos. Es decir, ¿un mal comportamiento de los alumnos produciría un grafo mal estructurado? Y viceversa, ¿las medidas estándar de calidad y entropía defini-

das puramente sobre grafos permiten detectar grupos en riesgo? La respuesta, resulta ser sí (Capítulo 14).

4.3 RESULTADOS OBTENIDOS

En esta sección se muestran los diferentes grafos obtenidos con la herramienta de Minería de Procesos *Graph Miner*. Comparando las Figuras 12 y 9 vemos que los diagramas de la implementación propia y el original obtenido con Disco coinciden, con la salvedad de que en el diagrama de creación propia se muestran el 100 % de las actividades y el 100 % de los caminos.

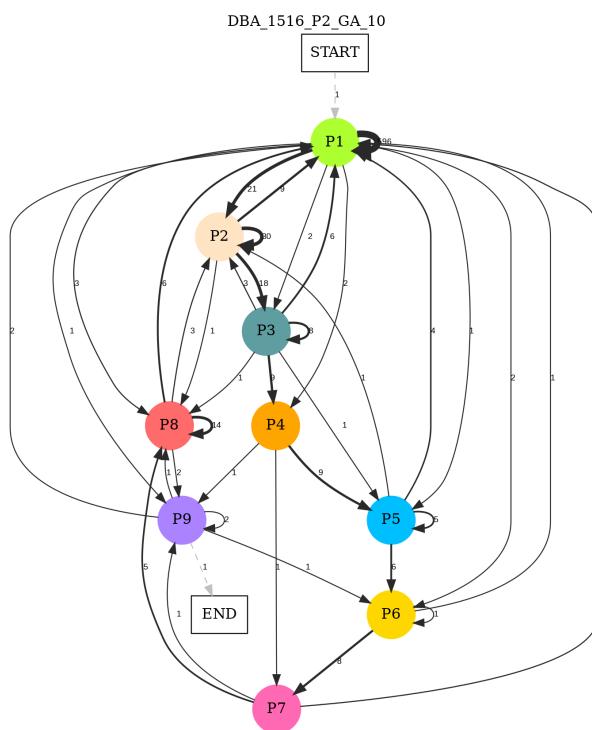


Figura 12: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema y CaseId grupo) obtenido con la implementación propia.

Además, como podemos ver en la Figura 13, hemos obtenido el mismo diagrama que el que se muestra en la Figura 10 con la salvedad de que hemos impedido el retorno a un estado anterior (es decir, se han eliminado ciclos) y como, en el caso anterior, de que en el diagrama de creación propia se muestran el 100 % de las actividades y el 100 % de los caminos.

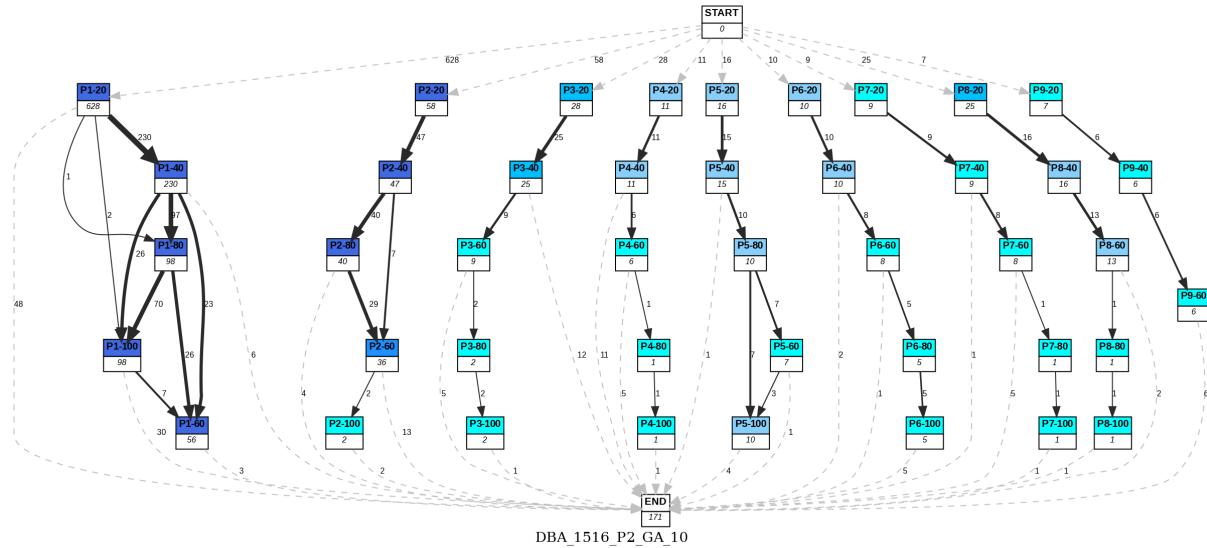


Figura 13: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema-milestone y CaseId sesión) obtenido con la implementación propia.

Además, se ha extendido la implementación de la herramienta de minería de procesos Disco Günther & Rozinat (2012), obteniendo otros diagramas que nos serán de mucha utilidad, como la agrupación en estados exitosos y fallidos (Figura 14) y una serie de grafos parciales considerando sólo la resolución hasta el problema i -ésimo, $i \in \{1, \dots, 9\}$ (Figuras 15, 16, 17 y 18) y considerando todos los registros, incluida toda la actividad realizada por el alumnado después de la resolución del último de los problemas ($i = 10$).

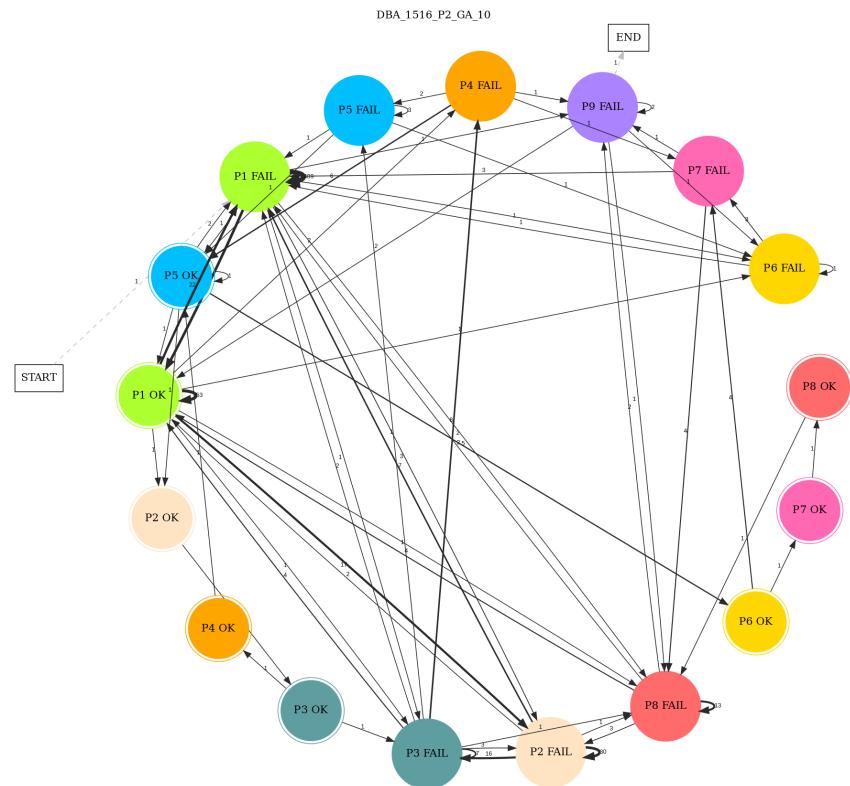


Figura 14: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema-estado y CaseId sesión) obtenido con la implementación propia.

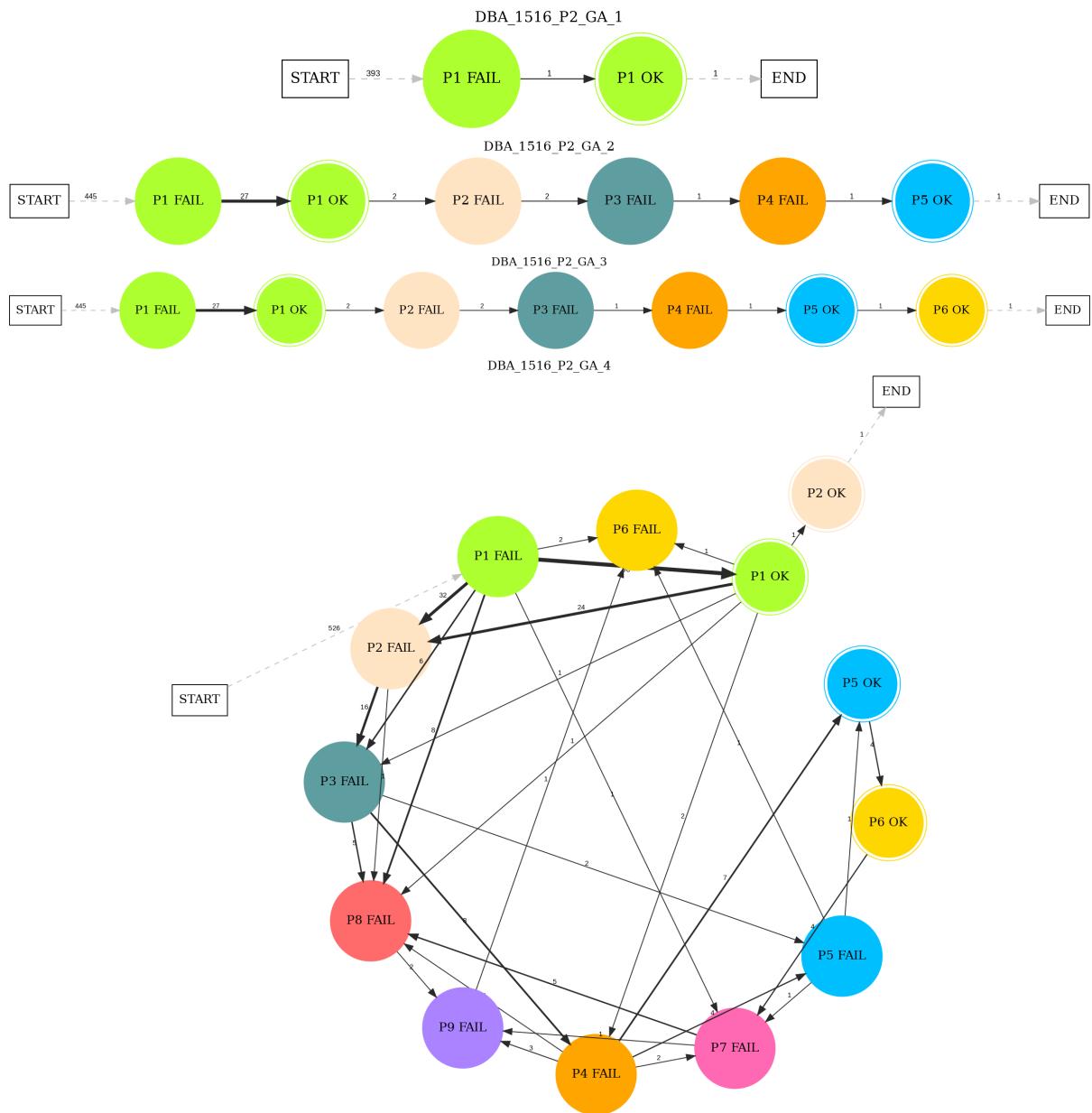


Figura 15: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema-estado y CaseId sesión) obtenido con la implementación propia. En estos grafos los ciclos se han eliminado con el procedimiento explicado en la Sección 4.1.

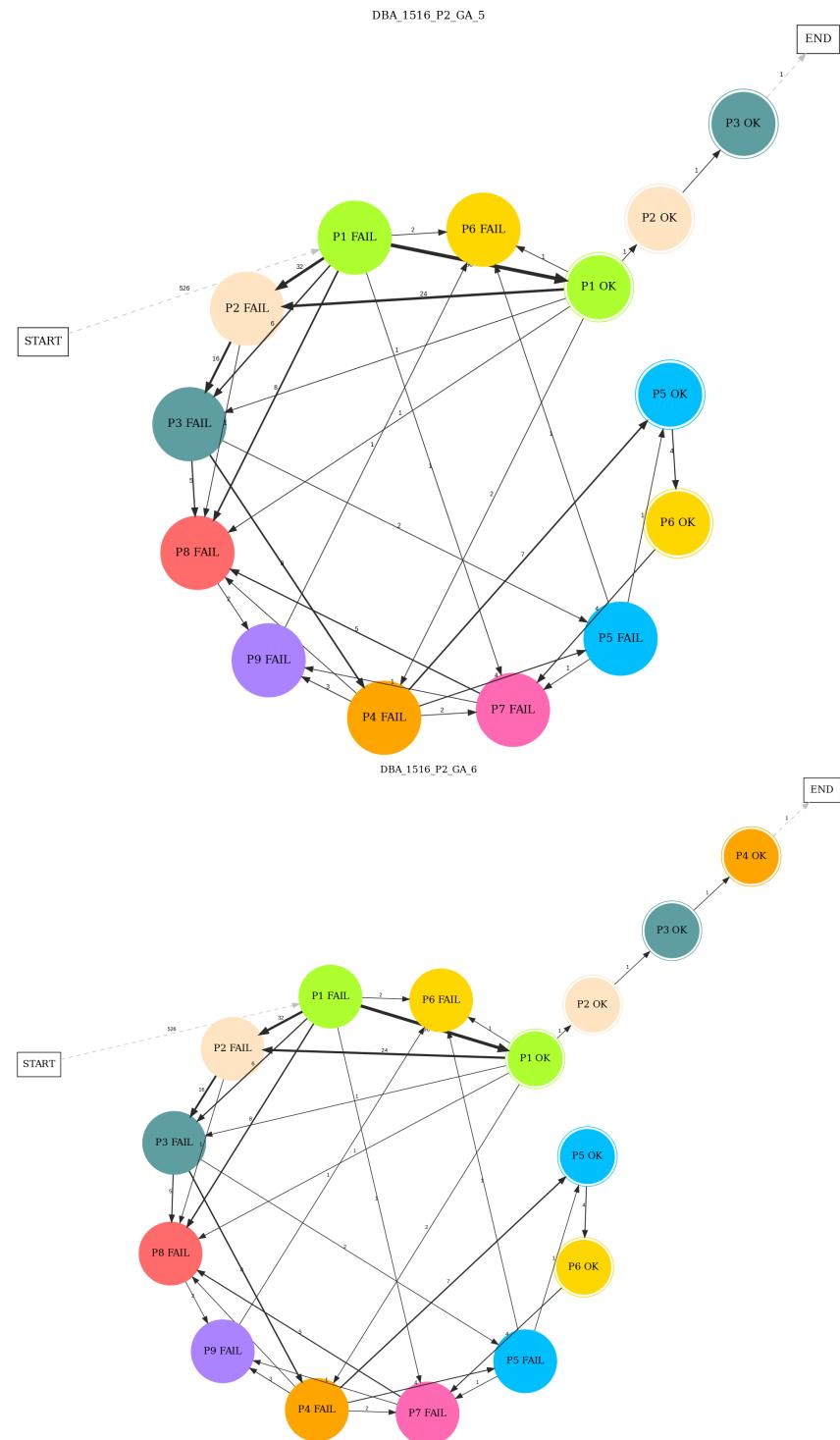


Figura 16: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema-estado y CaseId sesión) obtenido con la implementación propia. En estos grafos los ciclos se han eliminado con el procedimiento explicado en la Sección 4.1.

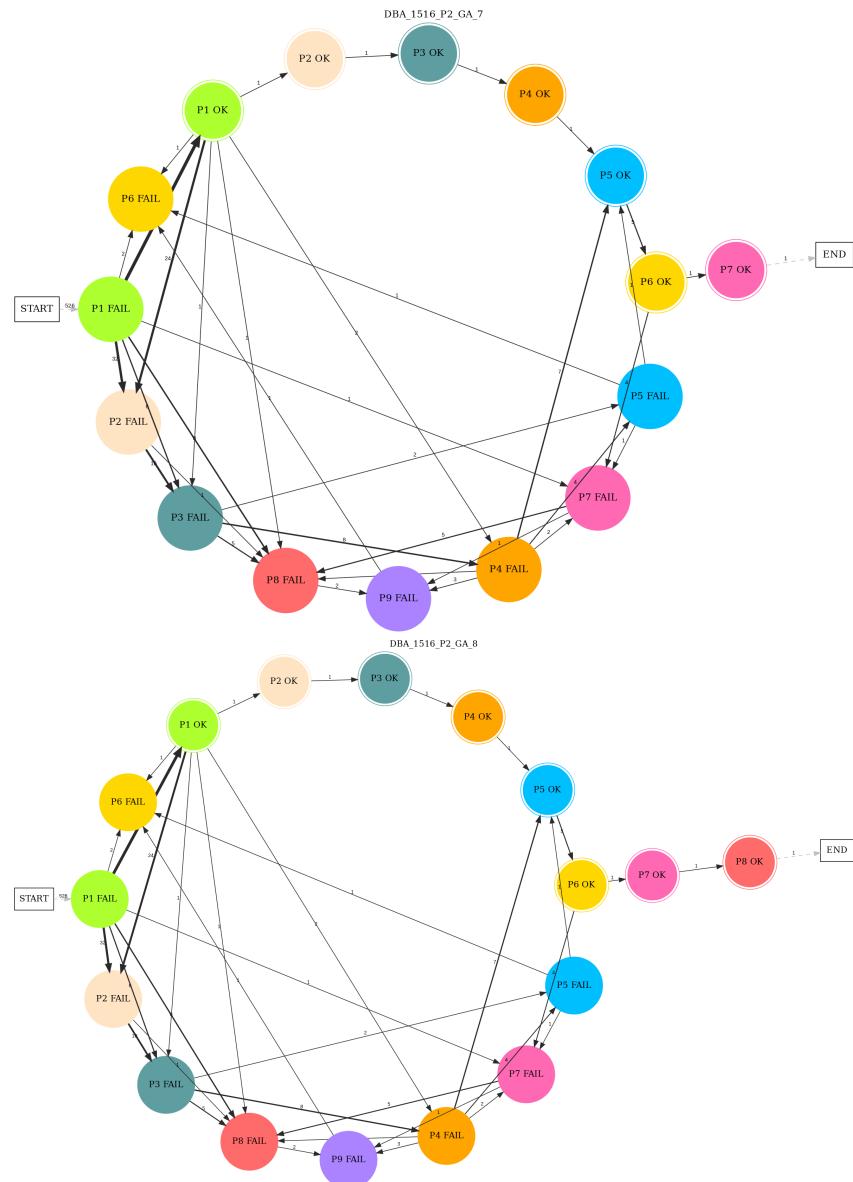


Figura 17: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema-estado y CaseId sesión) obtenido con la implementación propia. En estos grafos los ciclos se han eliminado con el procedimiento explicado en la Sección 4.1.

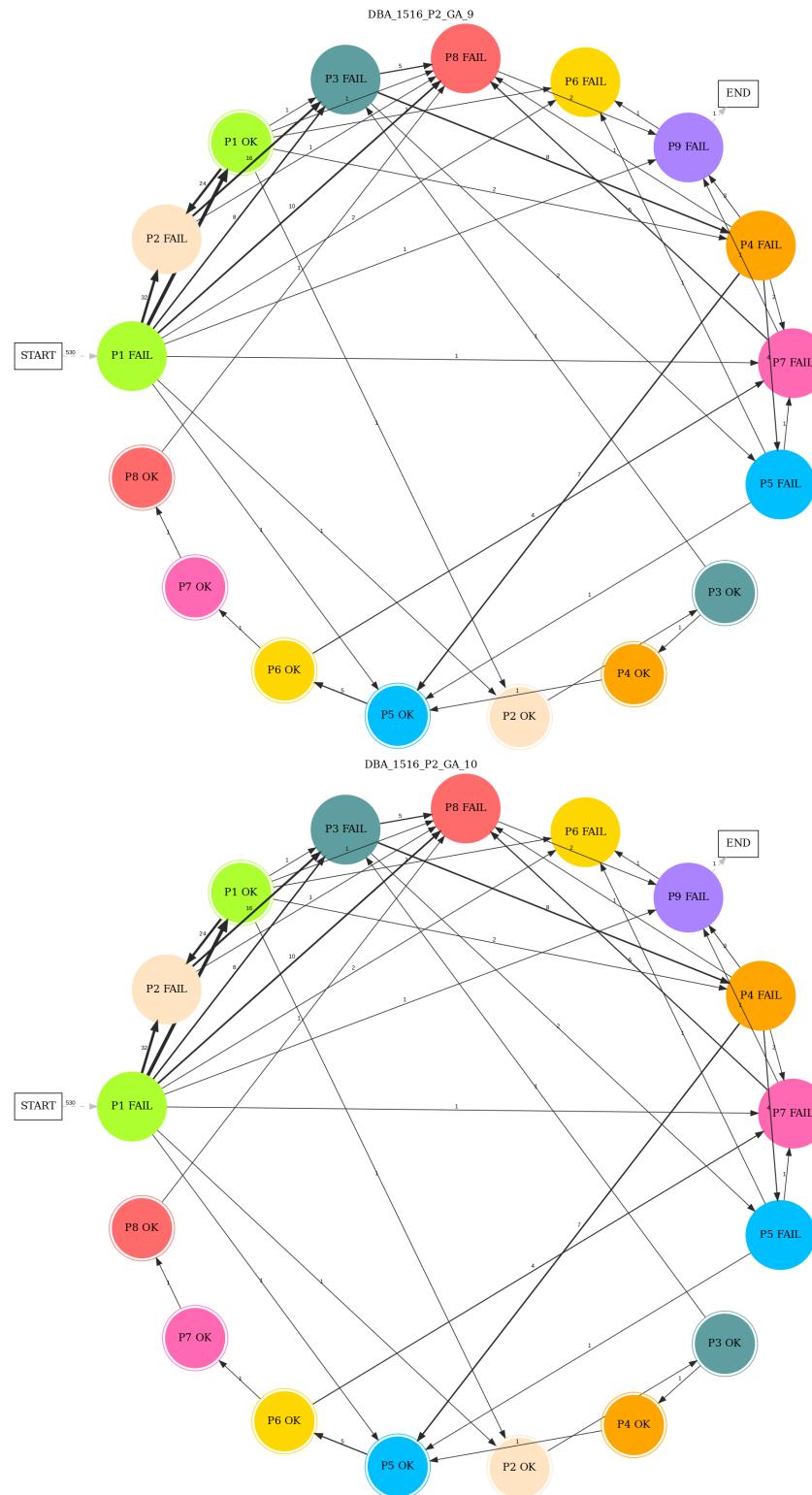


Figura 18: Análisis de procesos del grupo DBA 1516 P2 GA (Activity problema-estado y CaseId sesión) obtenido con la implementación propia. En estos grafos los ciclos se han eliminado con el procedimiento explicado en la Sección 4.1. En este caso, el grupo DBA 1516 P2 GA ha resuelto únicamente 8 de los 9 problemas resueltos, por lo que las gráficas DBA_1516_P2_GA_9 y DBA_1516_P2_GA_10 son idénticas y contendrán toda la actividad registrada en el servidor de dicho grupo.

A partir de ahora, estos diagramas 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.

5

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 [Mondal & De \(2017\)](#).

Como veremos en los Capítulos [12](#) y [14](#), los grafos jugarán un papel fundamental en este trabajo al conseguir representar fielmente el proceso de aprendizaje de los alumnos.

5.1 GRAFOS

En esta sección se introducirán las definiciones 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 $V \neq \emptyset$.

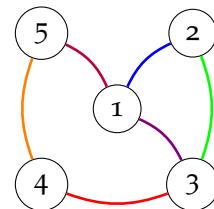


Figura 19: Ejemplo de grafo simple.

Ejemplo 1. El grafo dado en la Figura 19 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 la que se componen los grafos. Los vértices en sí mismos se tratan como objetos indivisibles y sin propiedades. No obstante, pueden tener asociadas una semántica dependiendo del contexto de aplicación del grafo. Por ejemplo, en el grafo [13](#) 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 ponderado* es un grafo cuyas aristas tienen un peso o valor asociado.

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

Ejemplo 2. En la Figura 20 se muestra un ejemplo de grafo ponderado que tiene conjunto de vértices $V = \{1, 2, 3, 4, 5, 6\}$, conjunto de aristas $E = \{(1, 2), (1, 4), (2, 3), (2, 5), (3, 6), (4, 5), (5, 6)\}$ y conjunto de pesos $W = \{2, 1, 1, 4, 3, 4, 2\}$.

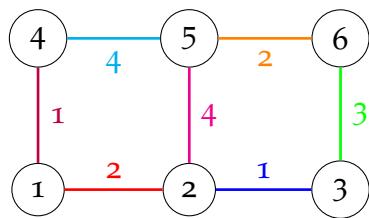


Figura 20: Ejemplo de grafo ponderado.

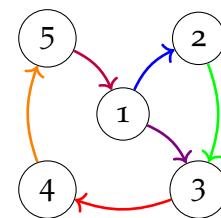


Figura 21: Ejemplo de grafo dirigido.

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 *dagrafo* a aquellos grafos cuyas aristas tengan un sentido definido. En un dagrafo, 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 3. En la Figura 21 se muestra un ejemplo de grafo dirigido mientras que en la Figura 19 tenemos un ejemplo de grafo no dirigido.

Definición 7. Dado un grafo $G = (V, E)$, un *camino* suyo es una secuencia de aristas (e_1, e_2, \dots, e_n) para la cual existe una secuencia de vértices $(v_1, v_2, \dots, v_{n+1})$ tales que $e_i = \{v_i, v_{i+1}\}$ en caso de que el grafo sea *no dirigido*, o bien $e_i = (v_i, v_{i+1})$ en caso de que sea *dirigido*, para todo $i \in \{1, \dots, n\}$. En este caso, la longitud del camino es n . Adicionalmente, un camino se dirá *cerrado* si $v_1 = v_{n+1}$ y *abierto* en caso contrario.

Un *grafo conexo* es un grafo en que todos sus vértices están conectados por un camino.

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

Definición 8. Un *bucle* es una arista que relaciona 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 4. En la Figura 19 se muestra un ejemplo de grafo simple. Por otro lado, en la Figura 22 podemos ver un multigrafo.

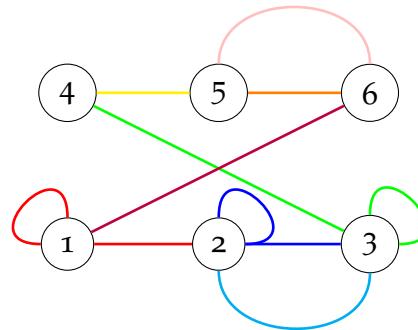


Figura 22: Ejemplo de multigrafo.

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 adyacentes 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 $\mathcal{A} = (a_{ij})_{1 \leq i, j \leq |V|}$.

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 sustituirá el valor de 1 en los casos anteriores por el peso de las aristas correspondientes.

Ejemplo 5. Tenemos que la matriz de adyacencia del grafo de la Figura 19 es:

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix} \quad (8)$$

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

$$N(v) = \{u \in V \mid \{v, u\} \in E\} \quad (9)$$

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 vértice final) y el *grado de salida* $\deg^+(v)$ (número de aristas que tienen a v como vértice inicial).

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 $\frac{n \cdot (n-1)}{2}$ aristas y es un grafo regular de grado $n - 1$.

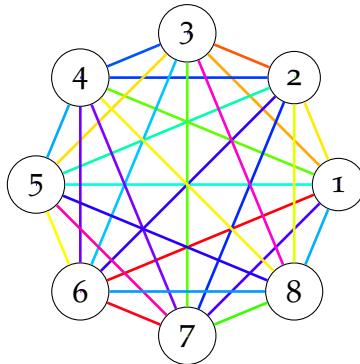


Figura 23: Ejemplo de grafo completo.

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) que se obtiene al añadir un único vértice a un grafo ciclo de $n - 1$ vértices, conectando el nuevo vértice a todos los ya existentes. Es decir, el nuevo vértice será adyacente a todos los vértices del grafo C_{n-1} .

Ejemplo 6. En las Figuras 23, 24 y 25 podemos ver un grafo completo, un grado ciclo y un grafo rueda respectivamente.

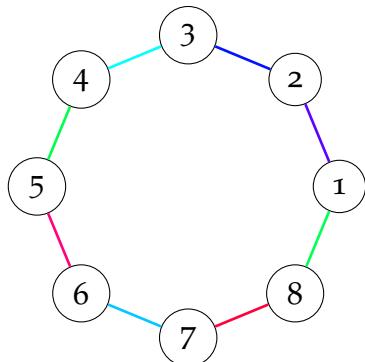


Figura 24: Ejemplo de grafo ciclo.

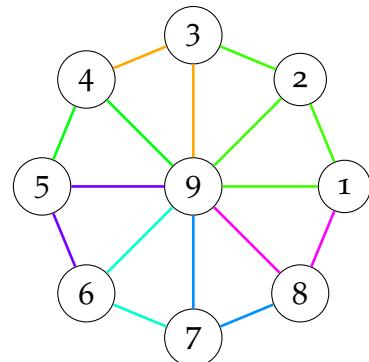


Figura 25: Ejemplo de grafo rueda.

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.

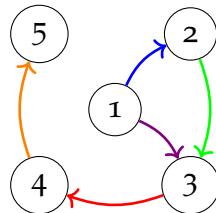


Figura 26: Ejemplo de grafo acíclico dirigido con 5 nodos.

Ejemplo 7. En la Figura 21 tenemos un grafo dirigido con ciclos o cíclico (contiene, por ejemplo, el 1-3-4-5). Sin embargo, eliminando una de las aristas del mismo obtenemos el grafo de la Figura 26, 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 dirigidas, se obtendrá un grafo no dirigido conexo y acíclico.

Definición 18. Un *árbol de expansión* o *árbol generador* de un grafo conexo no dirigido G es un subgrafo suyo que es un 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 Armenta (2015)*, 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 $V = \{v_1, v_2, \dots, v_n\}$, 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 $P(T) = \emptyset$. Supon-

gamos ahora que el código de Prüfer de cualquier árbol con n vértices está definido y sea $T = (V = \{v_1, v_2, \dots, v_n, v_{n+1}\}, E)$ un árbol con $n + 1$ vértices. Sea

$$v = v_{\min\{i \in \{1, 2, \dots, n, n+1\} \mid \deg(v_i) = 1\}} \quad (10)$$

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:

$$P(T) = (u, P(T')) \quad (11)$$

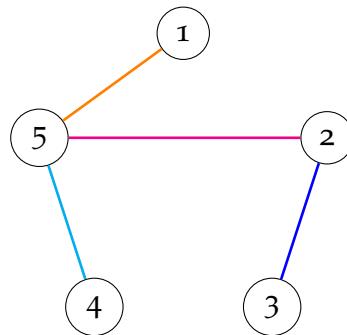


Figura 27: Ejemplo de árbol con 5 nodos.

Ejemplo 8. Consideremos el árbol de la Figura 27. Tenemos que el vértice de grado uno con la numeración más pequeña es el vértice 1. Este vértice únicamente es adyacente 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 adyacente 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 $T - \{1, 3, 2\}$ es un árbol con dos vértices, $P(T - \{1, 3, 2\}) = \emptyset$ y $P(T) = (5, 2, 5)$.

Este ejemplo pone de manifiesto que no es necesario que todos los vértices de un árbol aparezcan en su código de Prüfer asociado y que pudiera ocurrir que un mismo vértice aparezca más de una vez en el mismo. 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 la siguiente Proposición.

Proposición 2. *Sea c_i el número de veces que aparece el número i en el código de Prüfer de un árbol $T = (V = \{1, \dots, n\}, E)$ 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 $T = (V = \{1, \dots, n\}, E)$ con $n \geq 3$ vértices. Sea $T' = (V' = \{1, \dots, n, n+1\}, E')$, $v = \min\{i \in V' \mid \deg(i) = 1\}$ 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, $\deg_{T'}(i) = b_i + 2 = c_i + 1$. \square

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 $\{1, 2, \dots, n\}$.

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 $\{1, \dots, n, n+1\}$ que no aparece en $P(T)$. Por el lema anterior $\deg_T(v) = 1$ y $\deg_{T'}(v) = 1$. 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 $P(T') = (u', P(T - v))$. 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 $T = T'$. \square

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

Proposición 4. *Sea $n \geq 3$. Si $L = (u_1, u_2, \dots, u_{n-2})$ es una lista cuyos elementos pertenecen al conjunto $V = \{1, \dots, n\}$, 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 $L = (u_1, u_2, \dots, u_{n-1})$ 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

$V_{T'} = \{1, \dots, n+1\} - \{v\}$, tal que $P(T') = (u_2, u_3, \dots, u_{n-1})$. 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. \square

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 9. 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 debe ser adyacente al vértice 3 (Figura 28). 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 28 es el árbol asociado a la lista $(3, 5, 3, 1)$.

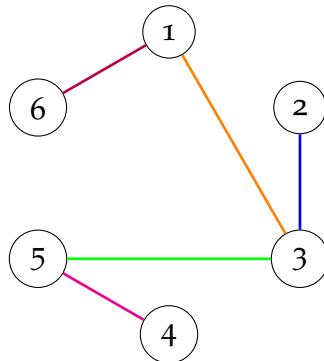


Figura 28: Ejemplo de árbol con 6 nodos.

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 Proposiciones 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 $\{1, 2, \dots, n\}$. 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. \square

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 característicos, valores y vectores propios de las matrices asociadas a los mismos.

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

$$\mathcal{L}_{i,j} = \begin{cases} \deg(v_i) & \text{si } i = j \\ -1 & \text{si } i \neq j \text{ y } v_i \text{ es adyacente a } v_j \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (12)$$

Equivalentemente, se tiene que $\mathcal{L} = \mathcal{D} - \mathcal{A}$ donde \mathcal{D} es la matriz de grados del grafo (matriz diagonal cuyos elementos no nulos son los grados de cada uno de los vértices) y \mathcal{A} es la matriz de adyacencia del grafo (Definición 10).

Ejemplo 10. Se tiene que la matriz laplaciana del grafo simple de la Figura 19 es la siguiente:

$$\begin{pmatrix} 3 & -1 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{pmatrix} \quad (13)$$

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 11. La matriz laplaciana del grafo simple de la Figura 20 será la siguiente:

$$\begin{pmatrix} 3 & -2 & 0 & -1 & 0 & 0 \\ -2 & 7 & -1 & 0 & -4 & 0 \\ 0 & -1 & 4 & 0 & 0 & -3 \\ -1 & 0 & 0 & 5 & -4 & 0 \\ 0 & -4 & 0 & -4 & 10 & -2 \\ 0 & 0 & -3 & 0 & -2 & 5 \end{pmatrix} \quad (14)$$

Ejemplo 12. Las matrices lapacianas del grafo dirigido de la Figura 21 de entrada y de salida se muestran en las Ecuaciones 18 y 19 teniendo en cuenta las matrices de adyacencia (Ecuación 15), del grado de entrada (Ecuación 16) y del grado de salida (Ecuación 17) de dicho grafo.

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (15)$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (16)$$

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (17)$$

$$\begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (18)$$

$$\begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (19)$$

Teorema 1. El Teorema de Kirchhoff [Williamson \(2019\)](#). Sea un grafo conexo G con n vértices numerados y sean $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$ los valores propios no nulos de su matriz laplaciana. Se tiene entonces que el número de árboles de expansión del grafo G es

$$t(G) = \frac{1}{n} \cdot \lambda_1 \cdots \lambda_{n-1} \quad (20)$$

Como podemos ver, se trata de una generalización de la fórmula de Cayley que además demuestra 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.

Demostración. Sea $\mathcal{L}_G[i]$ la matriz laplaciana del grafo G a la que se le ha eliminado su fila y su columna i -ésimas. Se procederá a demostrar que el número de árboles de expansión de un grafo G viene dado por $\det(\mathcal{L}_G[i])$, para cualquier $i \in \{1, \dots, n\}$.

Para ello, usaremos el siguiente hecho:

Proposición 5. Sea $A \in \mathbb{R}^{n \times n}$ y tomemos E_{ii} una matriz con un 1 en su entrada con posición (i, i) (fila i -ésima y columna i -ésima) y ceros en el resto de entradas. Así pues,

$$\det(A + E_{ii}) = \det(A) + \det(A[i]) \quad (21)$$

Este hecho es evidente si consideramos la suma de permutaciones de la definición del determinante de una matriz

$$\det(A = (a_{ij})) = \sum_{\pi} sgn(\pi) \cdot a_{1\pi(1)} \cdot a_{2\pi(2)} \dots a_{n\pi(n)} \quad (22)$$

y tenemos en cuenta que hemos aumentado la entrada (i, i) , a_{ii} , a $a_{ii} + 1$. Así pues, hemos obtenido el determinante original más la suma de todas las permutaciones de $[n] \setminus \{i\}$, evitando la i -ésima fila y columna (es decir, $\det(A[i])$).

La prueba se realizará por inducción en el número de vértices y aristas del grafo G .

Caso base: Si G es un grafo vacío con dos vértices, entonces

$$\mathcal{L}_G = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad (23)$$

Así pues, $\mathcal{L}_G[i] = 0$ y $\det(\mathcal{L}_G[i]) = 0$ tal y como se quería demostrar.

Paso inductivo: Sea $G - e$ el grafo resultante de sustraerle la arista e al grafo inicial y G/e el grafo con la arista e contraída (véase en la Figura 29 una ilustración de la contracción en grafos).

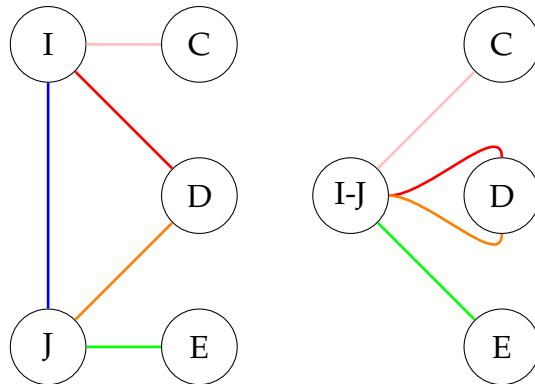


Figura 29: Ejemplo de contracción de los nodos I y J en un único nodo.

Si i es un vértice aislado, entonces G no admite árboles de expansión y la fila y la columna i -ésima de \mathcal{L}_G están conformadas de ceros. Esto es, $\det(\mathcal{L}_G[i]) = \det(\mathcal{L}_{G-i}) = 0$ tal y como queríamos.

Por el contrario, si i no es un vértice aislado podemos suponer que existe una arista $e = (i, j)$ incidente a i . Para cualquier árbol de expansión, T , se tiene que o bien $e \in T$ o bien $e \notin T$. Así pues, tenemos $t(G/e)$ denota el número de árboles de expansión T tal que $e \in T$ mientras que $t(G - e)$ proporcionará el número de árboles de expansión T tal que $e \notin T$. Esto es,

$$t(G) = t(G/e) + t(G - e) \quad (24)$$

Nótese que G/e tiene un vértice menos que G y que $G - e$ tiene una arista menos que G , lo cual servirá de base para realizar la inducción.

En primer lugar, se relacionará \mathcal{L}_G con \mathcal{L}_{G-e} . Observemos que $\mathcal{L}_G[i] = \mathcal{L}_{G-e}[i] + E_{jj}$ (esto es, si eliminamos la arista e la matriz Laplaciana resultante solamente difiere de $\mathcal{L}_G[i]$ en el grado del vértice j). Así pues, empleando la Proposición 5, tenemos que

$$\begin{aligned} \det(\mathcal{L}_G[i]) &= \det(\mathcal{L}_{G-e}[i] + E_{jj}) = \det(\mathcal{L}_{G-e}[i]) + \det(\mathcal{L}_{G-e}[i, j]) = \\ &= \det(\mathcal{L}_{G-e}[i]) + \det(\mathcal{L}_G[i, j]) \end{aligned} \quad (25)$$

donde $\mathcal{L}_G[i, j]$ denota a la matriz \mathcal{L}_G a la que se le han sustraído las filas y columnas i -ésima y j -ésima. La última igualdad se deduce de que, una vez eliminadas las filas y columnas i -ésima y j -ésima, no hay diferencia entre \mathcal{L}_G y \mathcal{L}_{G-e} para $e = (i, j)$.

Ahora relacionaremos \mathcal{L}_G con $\mathcal{L}_{G/e}$. Supongamos que contraemos i en j (entonces $\mathcal{L}_{G/e}$ no tiene ninguna fila o columna correspondiente a i). Entonces, $\mathcal{L}_{G/e}[j] = \mathcal{L}_G[i, j]$.

Así pues, tenemos que

$$\det(\mathcal{L}_G[i]) = \det(\mathcal{L}_{G-e}[i]) + \det(\mathcal{L}_{G/e}[j]) = t(G - e) + t(G/e) = t(G) \quad (26)$$

donde la segunda igual se deduce por hipótesis de inducción. Esto completa la prueba. \square

5.2 MEDIDAS DE COMPLEJIDAD DE PROPÓSITO GENERAL

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, a las que denominaremos características topológicas de los grafos de procesos. Se tratarán de funciones *off-the-shelf* genéricas sobre grafos, lo que permitiría aplicar los resultados aquí obtenidos en otras plataformas educativas.

Esta sección es una recopilación de varias de esas medidas de complejidad Cormen et al. (2022) 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 $\mathcal{A}_{v \times v}$ 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.

5.2.1 Densidad

Es el cociente entre el número de aristas del grafo y el número máximo de aristas permitidas $v \cdot (v - 1)$:

$$De(\mathcal{A}) = \frac{|E^{\mathcal{A}}|}{v^{\mathcal{A}} \cdot (v^{\mathcal{A}} - 1)} \quad (27)$$

5.2.2 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. Así pues, en los grafos dirigidos se tiene que el grado medio es:

$$Dm(\mathcal{A}) = \frac{|E^{\mathcal{A}}|}{v^{\mathcal{A}}} \quad (28)$$

5.2.3 Longitud del camino característico

Es el número medio de aristas, no sus pesos, entre cualquier par de nodos, en promedio. Podríamos aplicar cualquier algoritmo de caminos mínimos como el de Dijkstra [Cormen et al. \(2022\)](#) a la matriz característica \mathcal{A}' para propagar las frecuencias y obtener un cierre de la matriz original $\hat{\mathcal{A}'}$. Esto da una idea de la profundidad media del grafo, no en número de sesiones, sino en número de aristas. Esto es,

$$Le(\mathcal{A}) = \frac{1}{E^{\mathcal{A}}} \cdot \sum_r \sum_c \hat{\mathcal{A}'}[r, c] \quad (29)$$

5.2.4 Diámetro del grafo

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

$$Di(\mathcal{A}) = \max \hat{\mathcal{A}'}[r, c] \quad (30)$$

5.2.5 Conectividad

En este caso, se ha elegido el coeficiente de agrupación Cormen et al. (2022). La conectividad mide la densidad de los subgrafos alrededor de los nodos de una determinada arista. Es decir, es una medida de densidad local. Para el cálculo de la conectividad se usará el concepto de vecindad de un nodo v_i , $N(v_i)$, anteriormente definido en la Definición 11.

Así pues, el coeficiente de agrupación local mide el número de aristas existentes en la vecindad del nodo v_i , denotado por e_i , con respecto al máximo permitido de aristas en esta vecindad $|k_i| \cdot (|k_i| - 1)$:

$$Cl(\mathcal{A}) = \frac{e_i}{|k_i| \cdot (|k_i| - 1)} \quad (31)$$

5.2.6 Betweenness

Es muy conocido en las redes sociales e identifica cuál de los problemas ha recibido más atención e influencia que los demás. Ya está implementada en R y la llamaremos $Be(\mathcal{A})$.

5.2.7 Dag

Es la media de la longitud de todos los caminos diferentes que van del nodo raíz a cada uno de los nodos terminales del grafo. Trata de medir lo “desordenado” que es un grafo.

5.2.8 WDag

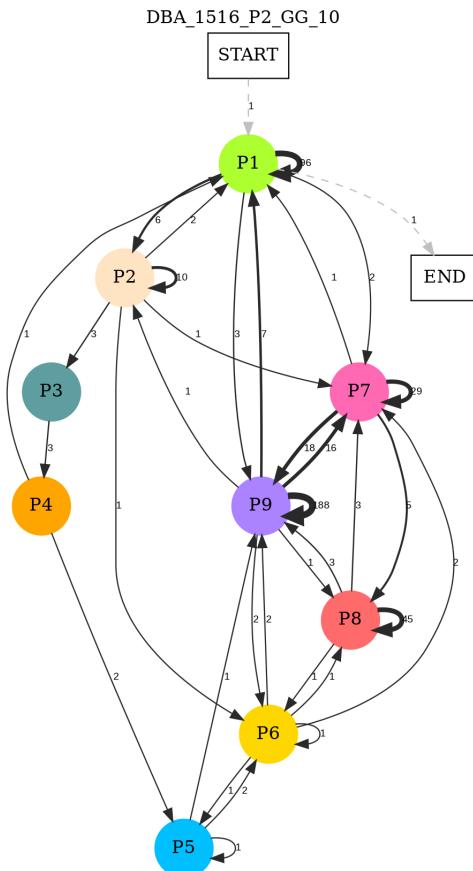
Se basa en Dag pero, en lugar de medir el número de pasos, recoge sus pesos.

5.2.9 St

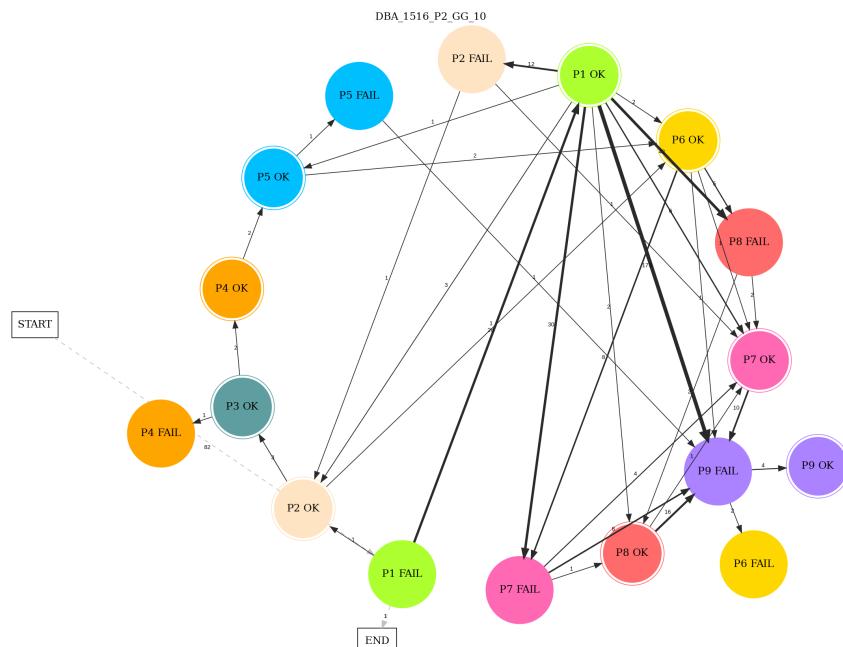
Definida como el factor Laplaciano de la matriz característica, es una medida basada en el análisis espectral de grafos. Este factor no es más que el número de árboles de expansión del grafo (Teorema 1). Se trata también de una medida muy común de la entropía de un grafo.

En primer lugar, se empieza 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 30a podemos ver el recorrido que hizo el grupo DBA 1516 P2 GG. 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 3ob se muestra este nuevo grafo para el grupo de prácticas que estamos considerando. Además, los ciclos han sido eliminados siguiente el procedimiento descrito en la Sección 4.1. Así pues, el coeficiente se calcula sobre la matriz característica de este último grafo.



(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 30: Leaning Path del grupo de prácticas DBA 1516 P2 GG.

5.2.10 Peso del comportamiento

Detecta la cantidad de esfuerzo que 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:

$$We(\mathcal{A}) = \sum_r \sum_c \mathcal{A}[r][c] \quad (32)$$

5.2.11 Eficacia

Detecta cuantos problemas se han resuelto, simplemente contado el número de nodos p_k^s presentes el grafo. Se formula como sigue:

$$Ef(\mathcal{A}) = \left| \left\{ p_i \in P, \sum_r \mathcal{A}[r][p_i^s] > 0 \right\} \right| \quad (33)$$

5.2.12 Balance

Trata de determinar cómo de balanceados están los nodos de un grafo dirigido acíclico. De aquí en adelante se denotará por $Ba(\mathcal{A})$.

El coeficiente Ba se calcula a partir de la matriz de adyacencia de un grafo dirigido acíclico. Éste simplemente consiste en calcular la desviación estándar de las componentes de la matriz de adyacencia no nulas y dividirla por la media de dichas entradas.

Todas las métricas generales aquí presentadas se utilizarán para clasificar los grupos de estudiantes. Se ha de tener en cuenta que todas ellas se han definido sobre grafos dirigidos ponderados y no tienen en cuenta absolutamente nada relacionado con el rendimiento de los alumnos ni con los registros temporales de su progreso excepto las funciones We (número de sesiones) y Ef (número de problemas resueltos) y que se extraen únicamente de la matriz característica \mathcal{A} .

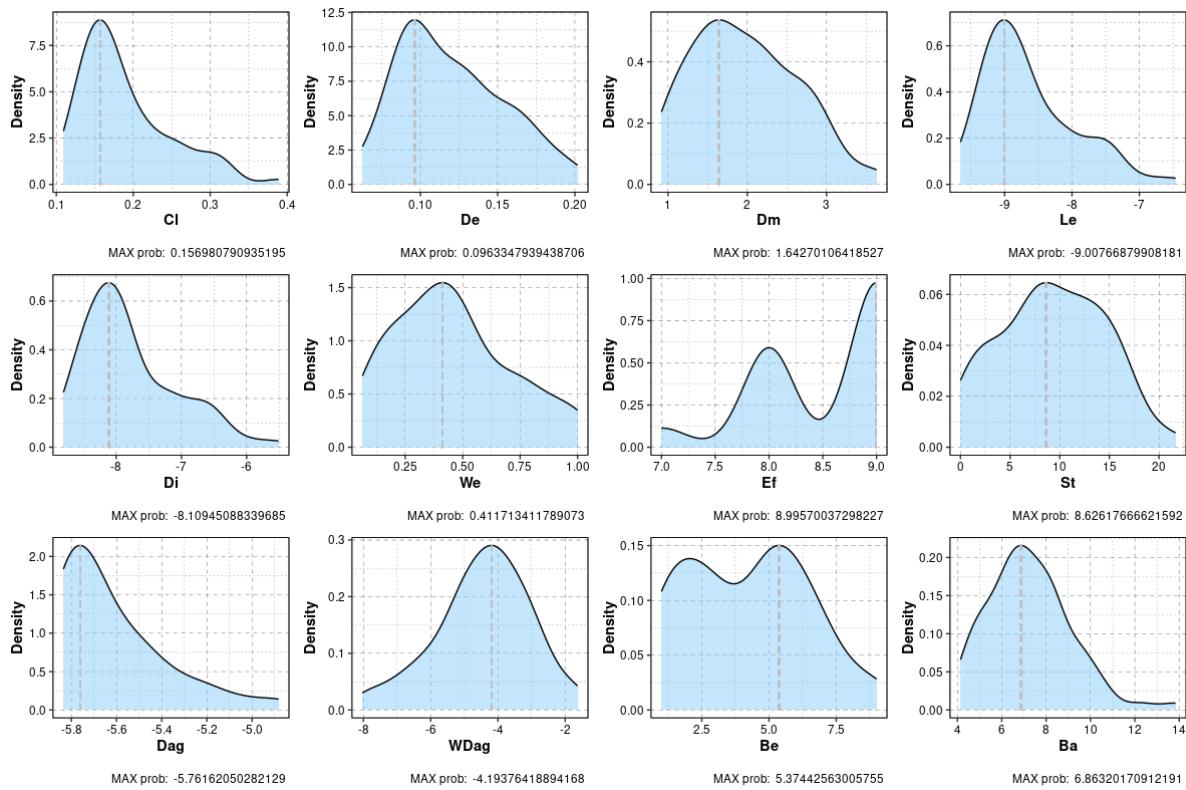


Figura 31: Funciones de densidad de las métricas topológicas utilizadas en el estudio. Las líneas discontinuas marcan los valores con probabilidad máxima.

Al tratar con conjuntos de métricas en el análisis estadístico, siempre es un alivio disponer de funciones con distribuciones normales y podemos aceptar que éste es el caso, tal y como se muestra en la Figura 31. En dicha imagen podemos ver que, excepto *Be*, que tiene rangos discretos (el índice de un problema), todas las métricas restantes podrían considerarse normales, dado el número limitado de grupos que estamos considerando en este estudio. Por otra parte, dado el comportamiento exponencial de algunas de ellas, especialmente las de naturaleza espectral (*Le*, *Di*, *Dag*, *WDag* y *St*), se han reducido con un logaritmo general con la finalidad obtener escalas equivalentes en todas las métricas.

Por otra parte, algunas de ellas se han reducido con un logaritmo general con la finalidad obtener escalas equivalentes en todas las métricas dado su comportamiento exponencial, especialmente las de naturaleza espectral (*Le*, *Di*, *Dag*, *WDag* y *St*).

Parte III

ANÁLISIS DESCRIPTIVO

Análisis estadístico de los datos.

6

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 3.

Cuadro 3: 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

6.1 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 24 y 25 los grupos por curso académico. El número de grupos por año puede consultarse también en la Figura 32.

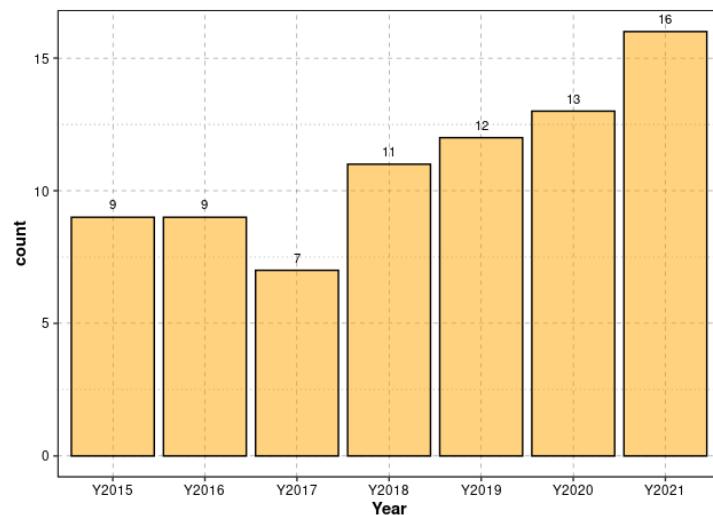


Figura 32: Número de grupos por curso académico estudiado.

6.2 EL PERIODO DE TIEMPO ANALIZADO CADA AÑO

En la Tabla 4 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 4: Número de días que dura la práctica cada año.

Year	Length (days)
Y2015	32
Y2016	23
Y2017	29
Y2018	17
Y2019	27
Y2020	16
Y2021	38

6.3 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.

6.3.1 Dificultad del problema: la tasa de fallo

La apertura de un problema se corresponde con una sesión de trabajo, la cual puede terminar fracasando (fail) si no se consigue resolver el problema, o teniendo 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 y el número total de sesiones de un mismo problema. El boxplot de las tasas de fallo por problema puede verse en la Figura 33.

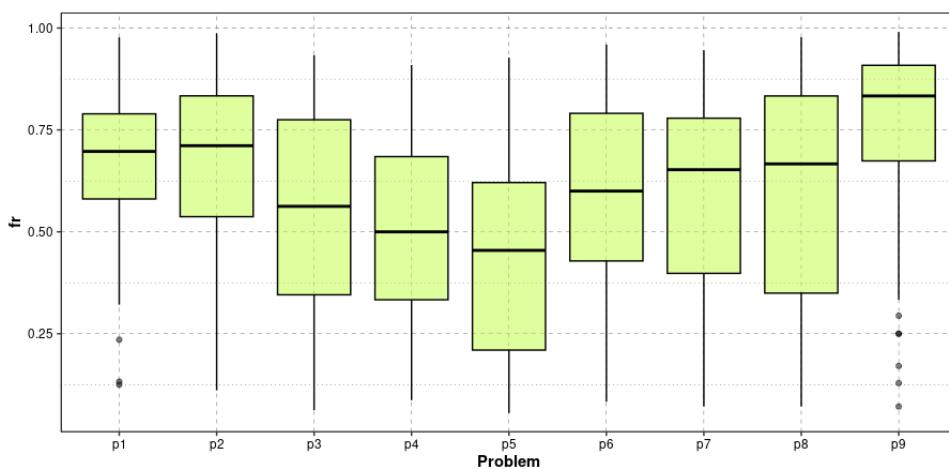


Figura 33: Boxplot de la tasa de fallo (*fail ratio*) por problema.

Tras realizar el test ANOVA de un factor (resultados en la Tabla 5), cuya hipótesis nula establece que la tasa de fallo media de los nueve problemas considerados es la misma, se detecta que las diferencias entre las distribuciones de probabilidad de la tasa de fallo podrían ser estadísticamente significativas en los distintos problemas ($p = 7,72e-12 < 0,05$)¹. Si aplicamos el test de Kruskal-Wallis, obtenemos $p - value = 3,756e-12 < 0,05$, lo que confirma los resultados obtenidos por el test de ANOVA.

Cuadro 5: Resultados del test ANOVA de un solo factor (tasa de fallo).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data[[Variable]]	8	4.00	0.50	9.12	7.72e-12
Residuals	570	31.23	0.05		

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 6). En él se observa que no todos los pares pueden considerarse estadísticamente iguales aunque haya algunos que quizá puedan serlo, como el par $P1 - P2$ ($p \text{ adj} = 1,00$). La Figura 34 muestra los intervalos de confianza de todas las diferencias entre las distintas parejas de años.

Cuadro 6: Test HSD de Tukey (Honestly-significance-difference) de la tasa de fallo por problemas.

	diff	lwr	upr	p adj
p2-p1	-0.00	-0.13	0.12	1.00
p3-p1	-0.11	-0.24	0.01	0.11
p4-p1	-0.17	-0.29	-0.04	0.00
p5-p1	-0.22	-0.35	-0.10	0.00
p6-p1	-0.08	-0.21	0.04	0.47
p7-p1	-0.09	-0.21	0.03	0.36
p8-p1	-0.06	-0.18	0.06	0.81
p9-p1	0.08	-0.05	0.21	0.61
p3-p2	-0.11	-0.24	0.02	0.17
p4-p2	-0.16	-0.29	-0.03	0.00
p5-p2	-0.22	-0.35	-0.09	0.00
p6-p2	-0.08	-0.21	0.05	0.59
p7-p2	-0.08	-0.21	0.04	0.48
p8-p2	-0.06	-0.18	0.07	0.89
p9-p2	0.08	-0.05	0.22	0.57
p4-p3	-0.05	-0.19	0.08	0.95
p5-p3	-0.11	-0.24	0.02	0.20
p6-p3	0.03	-0.10	0.16	1.00
p7-p3	0.03	-0.10	0.15	1.00
p8-p3	0.05	-0.08	0.18	0.94
p9-p3	0.19	0.06	0.33	0.00
p5-p4	-0.06	-0.19	0.08	0.93
p6-p4	0.08	-0.05	0.22	0.57
p7-p4	0.08	-0.05	0.21	0.64
p8-p4	0.10	-0.03	0.24	0.24
p9-p4	0.25	0.11	0.38	0.00
p6-p5	0.14	0.01	0.27	0.03
p7-p5	0.13	0.01	0.26	0.03
p8-p5	0.16	0.03	0.29	0.00
p9-p5	0.30	0.17	0.44	0.00
p7-p6	-0.01	-0.13	0.12	1.00
p8-p6	0.02	-0.11	0.15	1.00
p9-p6	0.16	0.03	0.30	0.01
p8-p7	0.03	-0.10	0.15	1.00
p9-p7	0.17	0.03	0.30	0.00
p9-p8	0.14	0.01	0.27	0.03

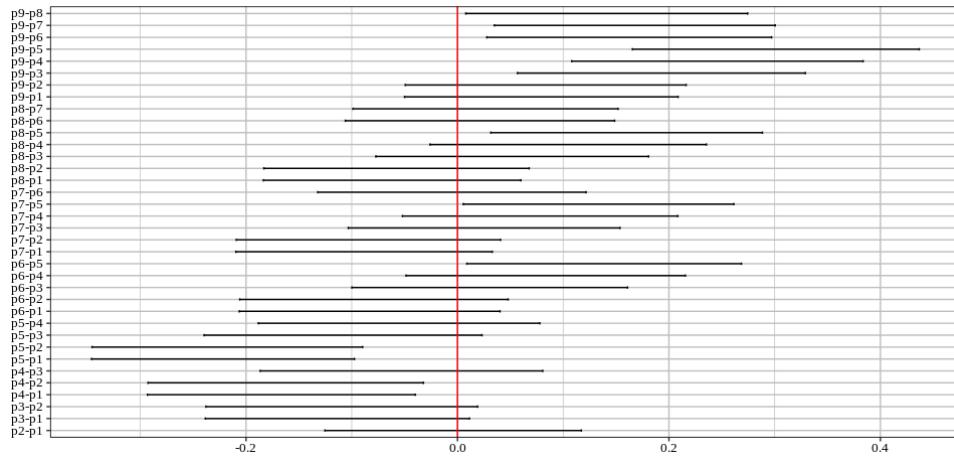


Figura 34: Intervalos de confianza de la tasas de fallo de los problemas.

Estos datos hay que interpretarlos de manera incremental, pues para resolver un problema P_i se requieren las habilidades de los problemas P_j , $0 \leq j < i$ más habilidades nuevas propias del problema P_i . Así pues, periódicamente se incrementa notablemente el nivel de dificultad.

6.3.2 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. El boxplot de los tiempos de resolución por problema puede verse en la Figura 35.

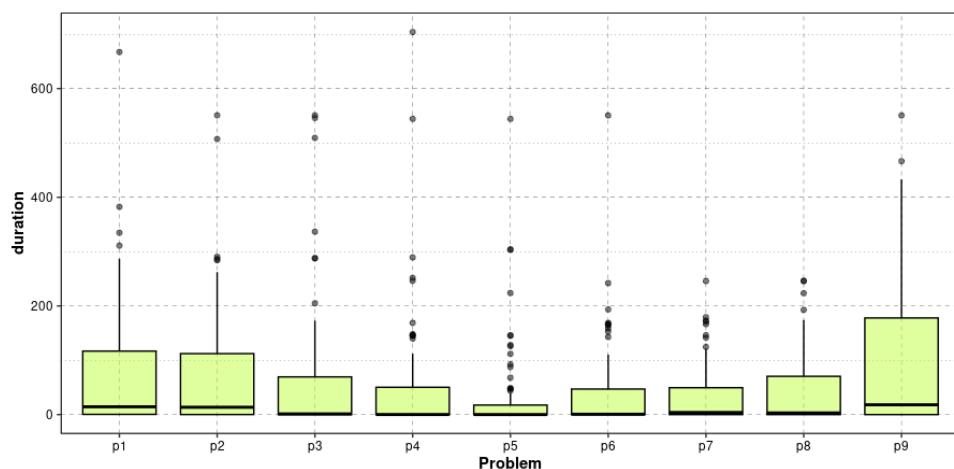


Figura 35: Boxplots del tiempo necesario para resolver cada uno de los problemas.

Podemos ver un resumen del test ANOVA de un factor en la Tabla 7. De nuevo, los tests detectan comportamientos diferentes: obtenemos $p = 0,0074$ con el test ANOVA de un solo factor y $p = 0,0002763$ con la prueba de Kruskal-Wallis).

Cuadro 7: Resultados del test ANOVA de un solo factor (tiempo en resolver los problemas por primera vez).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data[[Variable]]	8	222548.19	27818.52	2.64	0.0074
Residuals	642	6757495.31	10525.69		

Adicionalmente, se aportan los resultados del test de Tukey por pares de problemas (Tabla 8). En ellos se detecta, por ejemplo, que los pares de problemas $P9 - P7$ y $P9 - P5$ son estadísticamente diferentes ($p_{adj} = 0,02$ en ambos casos). En la Figura 36 se muestran los intervalos de confianza de todas las diferencias entre las distintas parejas de años.

Cuadro 8: Test HSD de Tukey (Honestly-significance-difference) del tiempo de resolución por problemas.

	diff	lwr	upr	p adj
p2-p1	-8.38	-59.84	43.08	1.00
p3-p1	-14.68	-66.66	37.31	0.99
p4-p1	-22.83	-74.99	29.33	0.91
p5-p1	-41.17	-92.97	10.63	0.25
p6-p1	-34.32	-86.85	18.22	0.52
p7-p1	-41.10	-93.26	11.07	0.26
p8-p1	-27.84	-80.38	24.70	0.78
p9-p1	19.54	-35.45	74.52	0.97
p3-p2	-6.30	-58.28	45.68	1.00
p4-p2	-14.45	-66.61	37.71	0.99
p5-p2	-32.79	-84.60	19.01	0.56
p6-p2	-25.94	-78.48	26.60	0.84
p7-p2	-32.72	-84.88	19.45	0.58
p8-p2	-19.46	-72.00	33.07	0.97
p9-p2	27.91	-27.07	82.90	0.82
p4-p3	-8.15	-60.83	44.52	1.00
p5-p3	-26.49	-78.81	25.83	0.82
p6-p3	-19.64	-72.69	33.41	0.97
p7-p3	-26.42	-79.10	26.25	0.83
p8-p3	-13.17	-66.21	39.88	1.00
p9-p3	34.21	-21.26	89.68	0.60
p5-p4	-18.34	-70.84	34.16	0.98
p6-p4	-11.49	-64.71	41.74	1.00
p7-p4	-18.27	-71.12	34.59	0.98
p8-p4	-5.01	-58.24	48.21	1.00
p9-p4	42.36	-13.28	98.01	0.30
p6-p5	6.85	-46.02	59.73	1.00
p7-p5	0.07	-52.43	52.57	1.00
p8-p5	13.33	-39.55	66.20	1.00
p9-p5	60.70	5.40	116.01	0.02
p7-p6	-6.78	-60.00	46.45	1.00
p8-p6	6.47	-47.12	60.07	1.00
p9-p6	53.85	-2.14	109.85	0.07
p8-p7	13.25	-39.97	66.48	1.00
p9-p7	60.63	4.99	116.27	0.02
p9-p8	47.38	-8.62	103.37	0.17

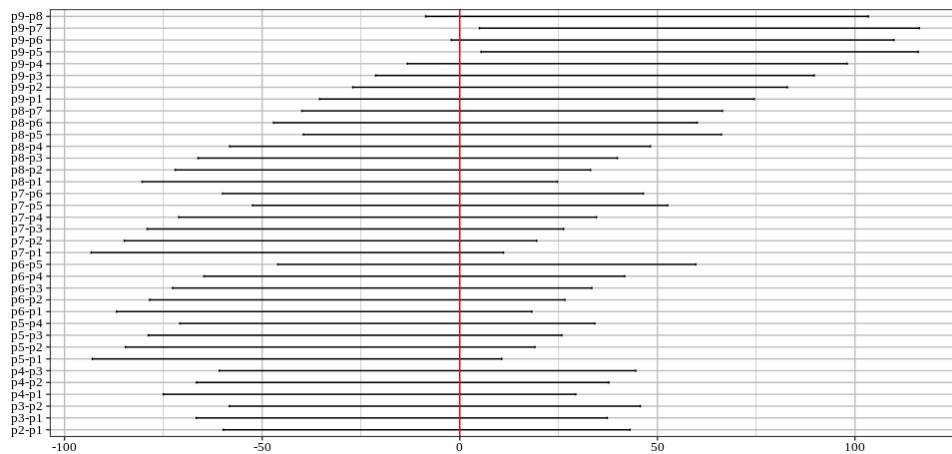


Figura 36: Intervalos de confianza del tiempo empleado en la resolución de los problemas propuestos por primera vez.

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

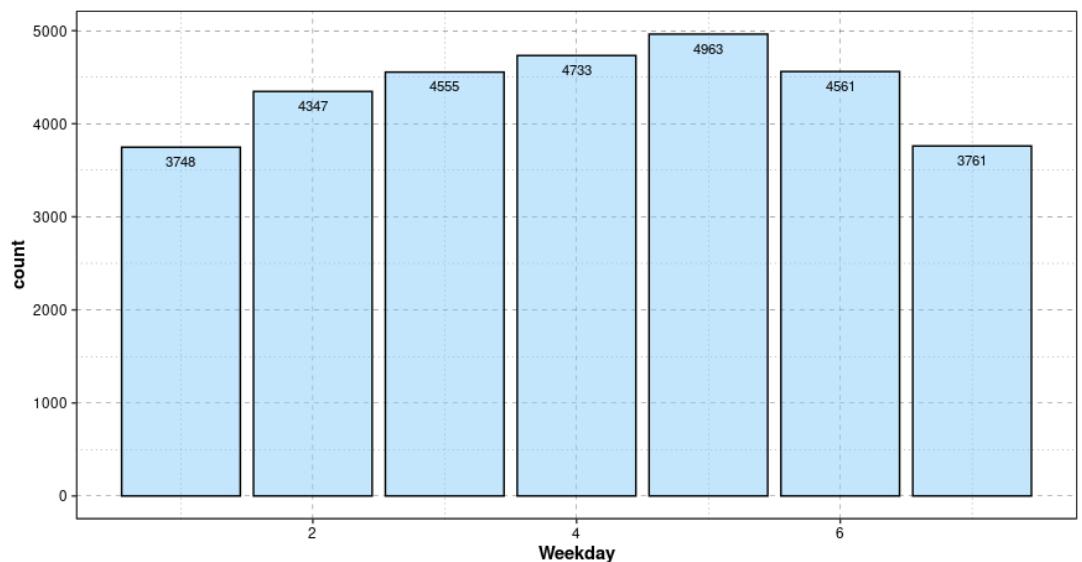
6.4 ACTIVIDAD REGISTRADA

El número de registros y de sesiones de trabajo de cada uno de los años analizados se muestran en la Tabla 9. Como podemos ver, aunque el curso académico 2020-2021 (Y2020) registra más actividad que los demás, no es el que presenta el mayor número de sesiones.

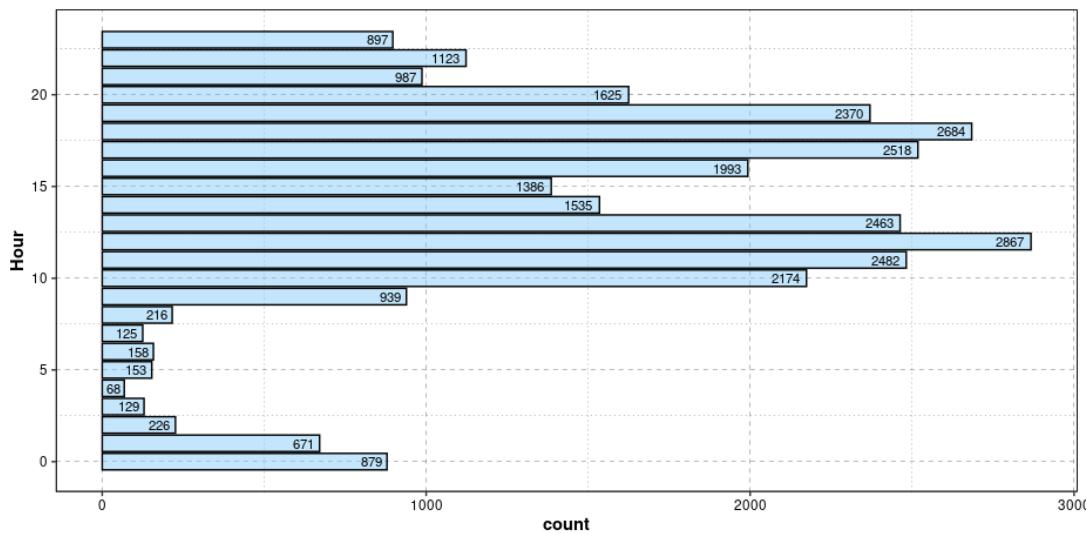
Cuadro 9: Número de registros y sesiones almacenados en el servidor por años.

Year	Activity Records	Sessions
Y2015	12088	4489
Y2016	12525	4538
Y2017	9088	3661
Y2018	5705	2811
Y2019	14475	5156
Y2020	21188	3904
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 37a y a cualquier hora del día (Figura 37b).



(a) Histograma de los días de la semana.



(b) Histograma de las horas del día.

Figura 37: 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 26.

6.4.1 Análisis de la normalidad de la distribución del número de sesiones

En las Figuras 38a y 38b podemos ver el boxplot de los residuos de la variable número de sesiones (s) junto con el histograma de los mismos.

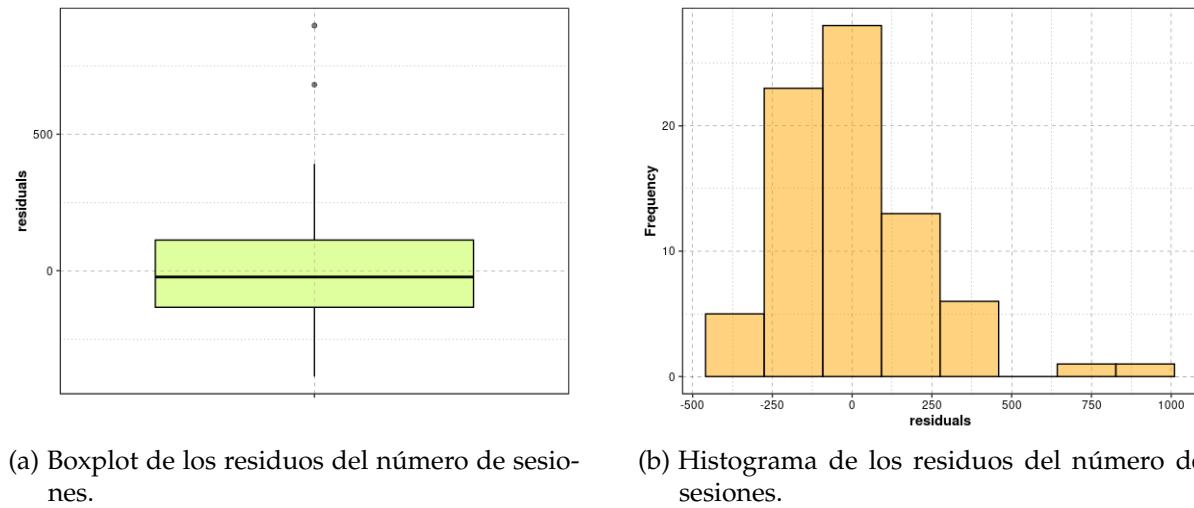


Figura 38: Distribución de los residuos del número de sesiones.

A continuación, en las Figuras 39 y 40, podemos observar que la distribución del número de sesiones no es perfectamente normal pero es casi-normal si eliminamos algunos outsiders. La línea discontinua vertical marca el valor más probable (336 sesiones), lo que muestra un gran esfuerzo por parte del alumnado teniendo en cuenta la duración de la práctica (Tabla 4).

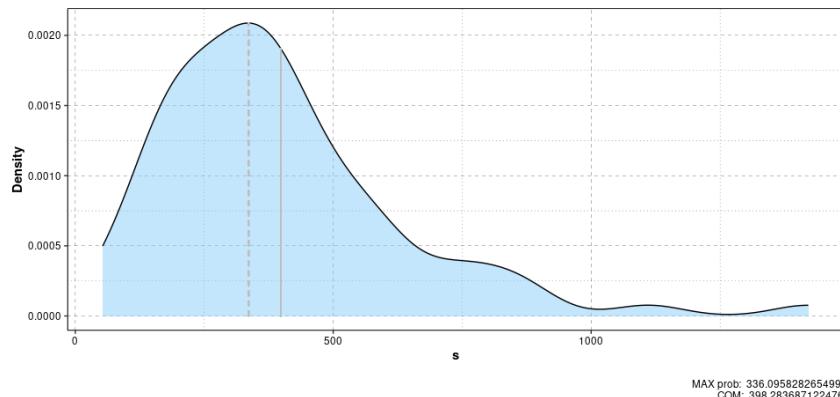


Figura 39: Función de densidad de probabilidad del número de sesiones.

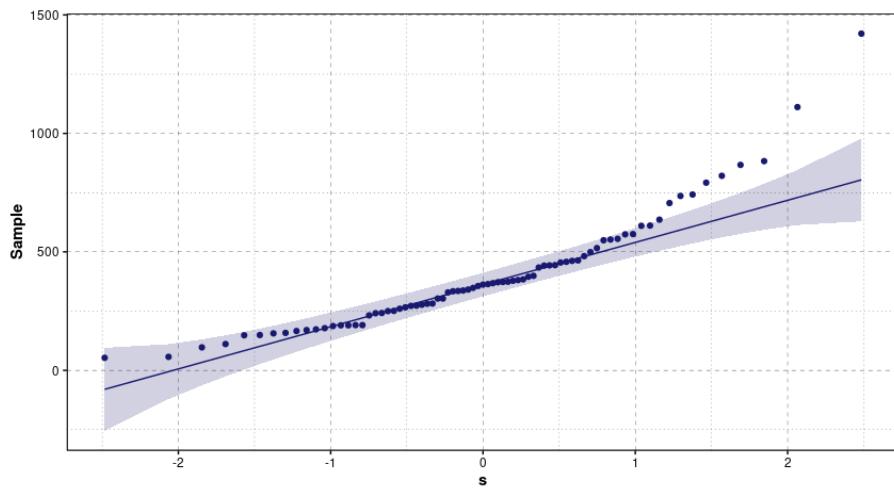


Figura 40: Gráfico Q-Q del número de sesiones.

Además, podemos ver en la Figura 42 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 41. 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 43.

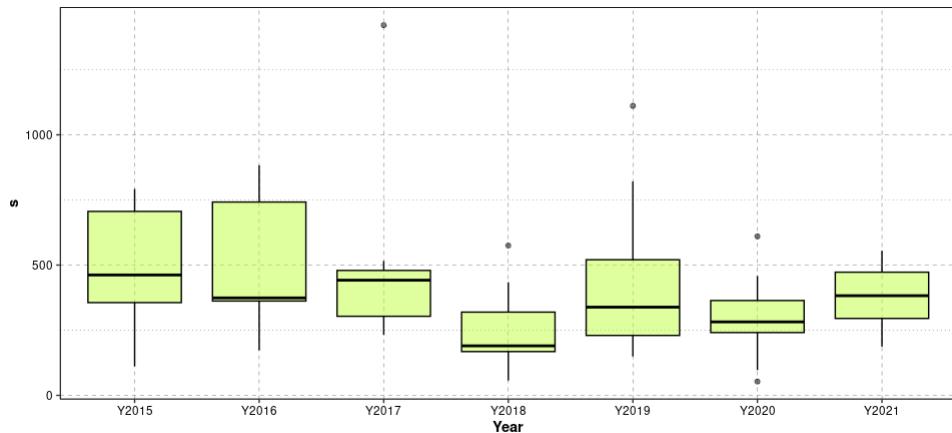


Figura 41: Boxplot del número de sesiones por año inicialmente.

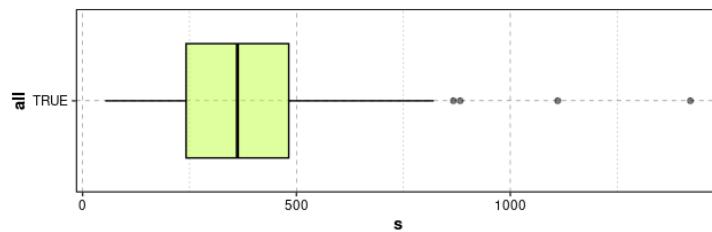


Figura 42: Distribución del número de sesiones inicial.

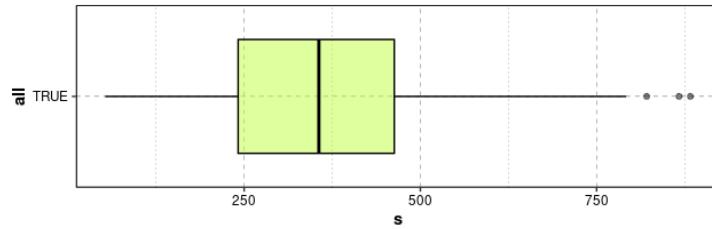


Figura 43: Distribución del número de sesiones tras la eliminación de algunos outliers.

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 44. Los resultados obtenidos pueden verse en las Figuras 45a y 45b.

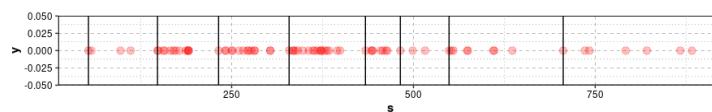
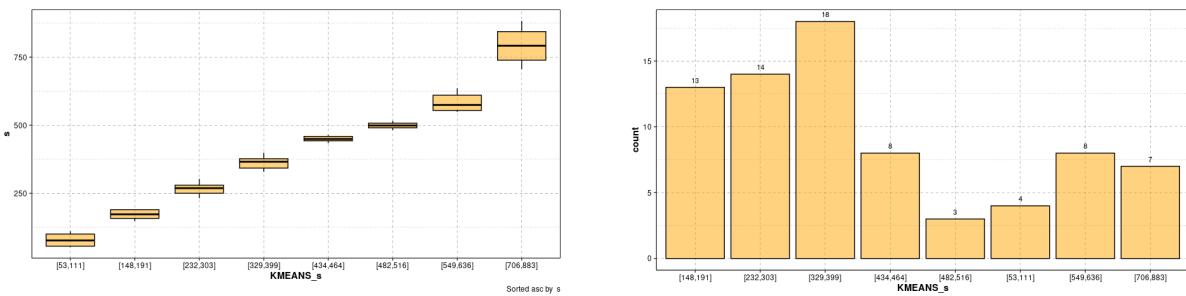


Figura 44: Particiones obtenidas con $K = 8$.



(a) Boxplot de cada una de las particiones.

(b) Número de grupos por partición.

Figura 45: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 8$.

Nótese que como hemos obtenido una precisión del $97,95579\% > 95\%$, 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 6.4.4 (podemos ver la nueva función de densidad en la Figura 46) 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.

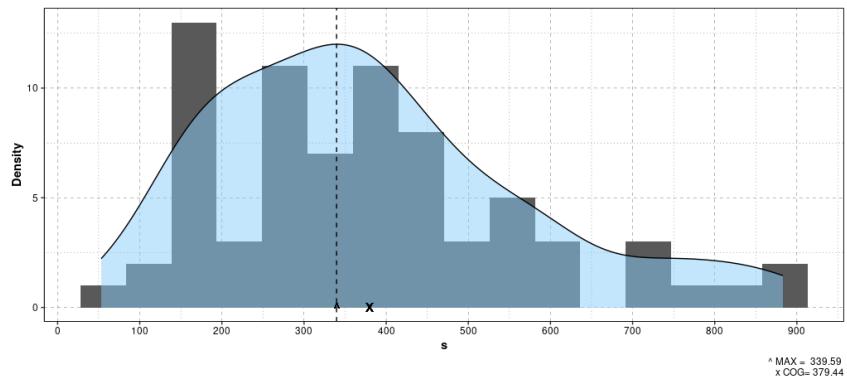


Figura 46: Función de densidad de probabilidad del número de sesiones tras eliminar algunos outliers.

6.4.2 Sesiones por cada problema

En la Figura 47 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, la diferencia entre el número de sesiones abiertas del problema P1 y el número de sesiones abiertas de los problemas 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 de la práctica, sino durante todo el desarrollo de la misma.

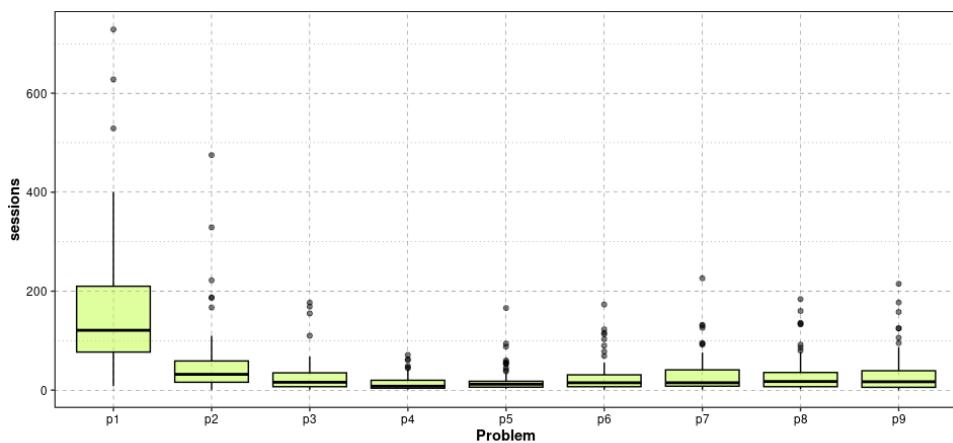


Figura 47: Boxplot del número de sesiones por problema.

6.4.3 Sesiones cada año

Como podemos ver en la Figura 48, las sesiones de trabajo abiertas en el servidor año tras año, parecen seguir la misma distribución de probabilidad.

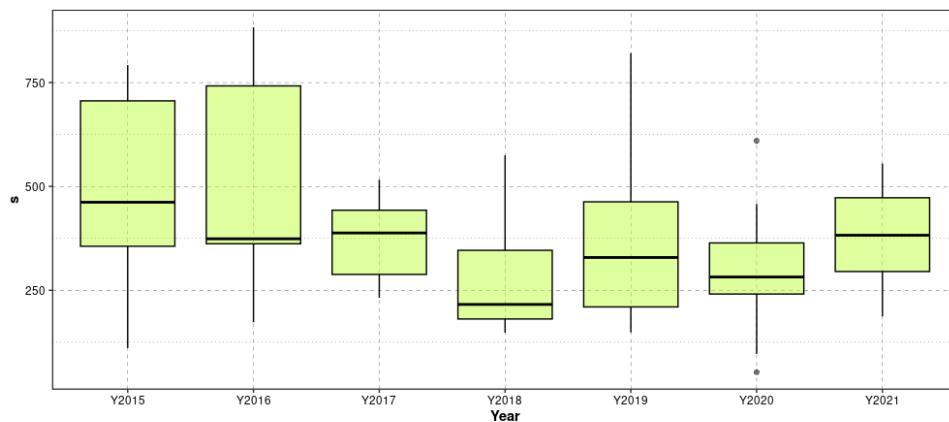


Figura 48: Boxplot del número de sesiones por año tras la eliminación de algunos outliers.

Un resumen de los resultados obtenidos al realizar el test ANOVA se muestra en la Tabla 10. 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 $p = 0,0412 < 0,05$, por lo que 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 10: Resultados del test ANOVA de un solo factor (número de sesiones).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data[[Variable]]	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 11). 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 49 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 básico de los alumnos, que viene dado por cuántas veces se conectan al servidor, es el mismo en todos los cursos académicos considerados.

Cuadro 11: 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

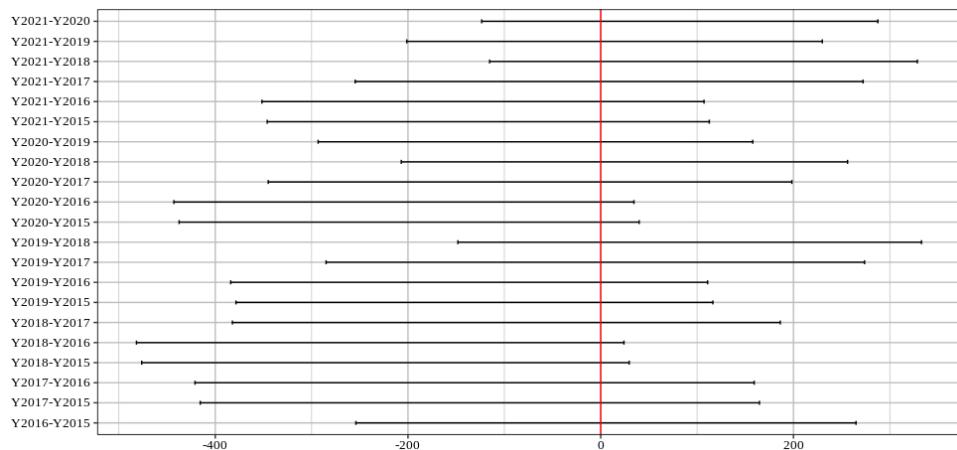


Figura 49: Intervalos de confianza del número de sesiones por años.

6.4.4 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 50 podemos ver que el número de problemas resueltos oscila entre 6 y 9.

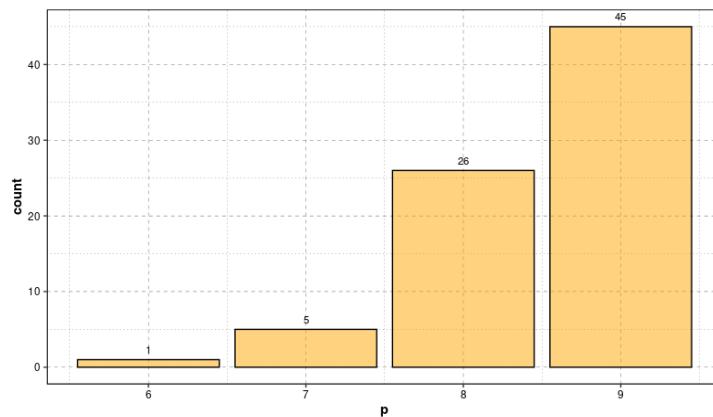


Figura 50: Distribución del número de problemas resueltos.

Adicionalmente, podemos observar en la Figura 51 la existencia de un elemento extremo (6) en la distribución del número de problemas resueltos por grupo de alumnos. Segmentando por años, obtenemos los boxplots que se muestran en la Figura 52. Así pues, eliminaremos el outlier encontrado puesto que se trata de un valor extremo en todos los años incluidos en este estudio.

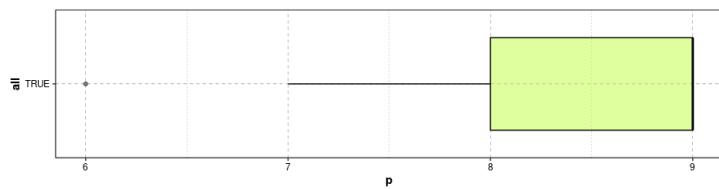


Figura 51: Distribución del número de problemas resueltos inicial.

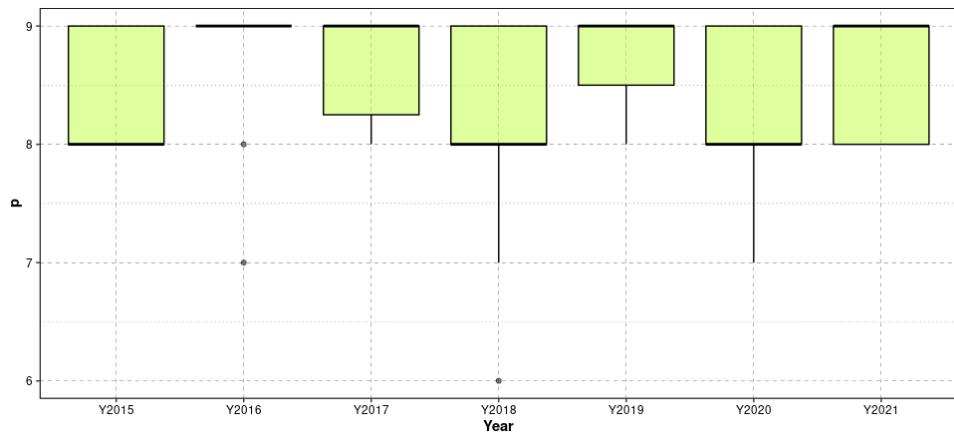


Figura 52: Boxplot del número de problemas resueltos por año.

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 profesorado 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.

7.1 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 clásicas del rendimiento del estudiantado se han representado en la Figura 53.

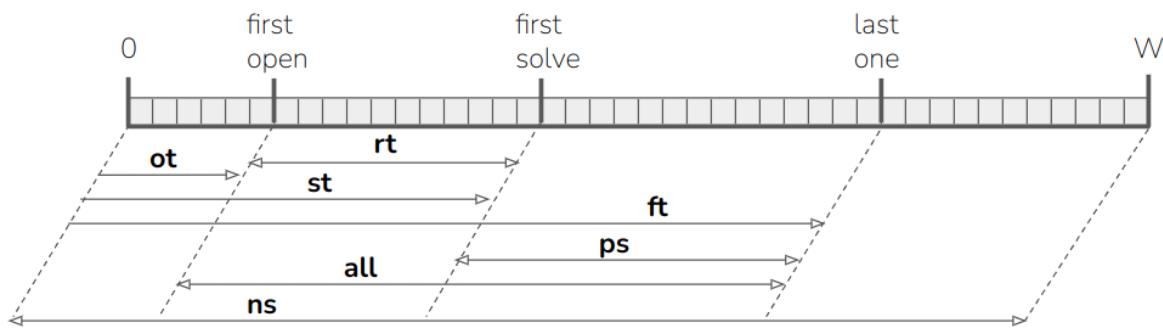


Figura 53: Representación gráfica de las medidas de rendimiento empleadas que se extraen directamente de los registros del servidor (no se incluyen las medidas derivadas del análisis espectral de grafos).

7.1.1 *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 53 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. Éste 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 53, 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 6.4. En la Figura 53, se representa el número de sesiones normalizado (*ns*).

7.1.2 *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 53.
- 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.

- 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 53 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 promedio de resolución de los problemas por primera vez. Se representará por st en la Figura 53.
- 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.
- Todas las medidas de complejidad definidas en la Sección 5.2.

8

RENDIMIENTO OBSERVADO DE LOS ALUMNOS

8.1 CALIFICACIONES OBTENIDAS (GRADE)

La primera medida de rendimiento que tendremos en cuenta son las calificaciones de los grupos de prácticas. Nótese 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 54.

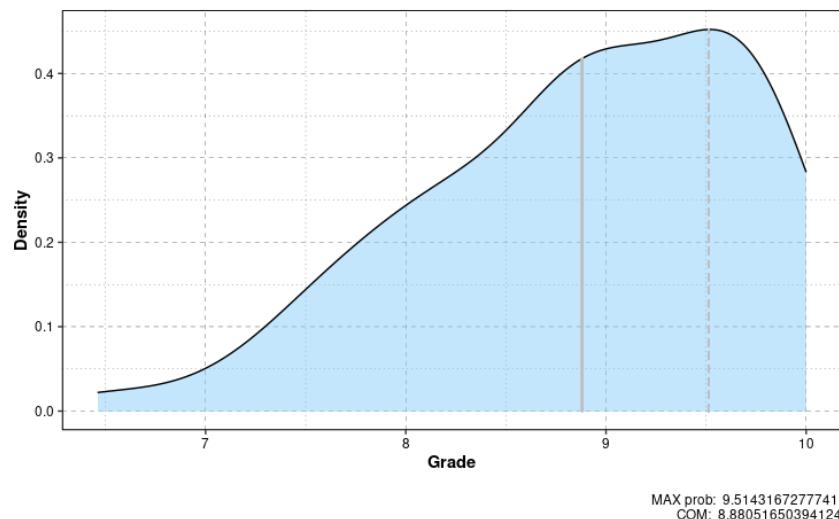


Figura 54: Función de densidad de probabilidad de las calificaciones obtenidas por los distintos grupos de prácticas.

Además, la media está bastante balanceada (Figuras 55a y 55b) y no se detecta la presencia de outliers (Figura 56).

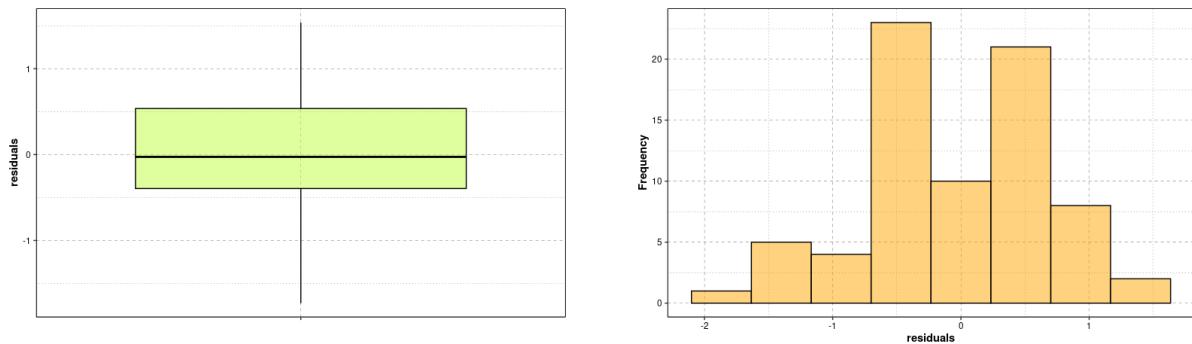


Figura 55: Distribución de los residuos de las calificaciones.

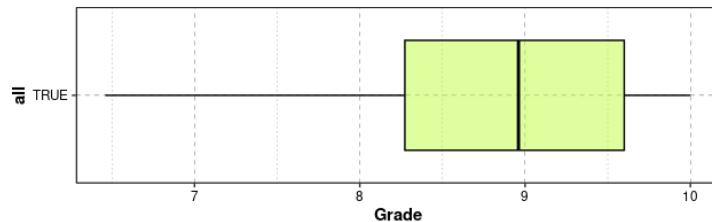


Figura 56: Distribución de las calificaciones obtenidas por los distintos grupos de alumnos inicial.

Más aún, la regresión es muy aceptable (Figura 57).

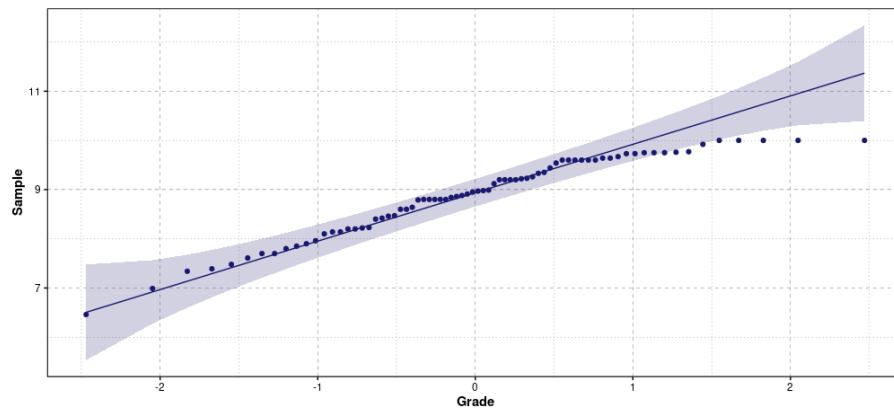


Figura 57: Gráfico Q-Q de las calificaciones.

A continuación se realizará un estudio por años de las calificaciones obtenidas. El boxplot de las calificaciones por años puede verse en la Figura 58. Como se puede observar, 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 diferencias significativas entre ellas. Tras realizar el test ANOVA de un factor (resultados en la Tabla 12) 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 13). Podemos ver que el año 2017 es el principal elemento de disruptión, pero por poco margen. Lo mismo indica el análisis de los intervalos de confianza de la Figura 59.

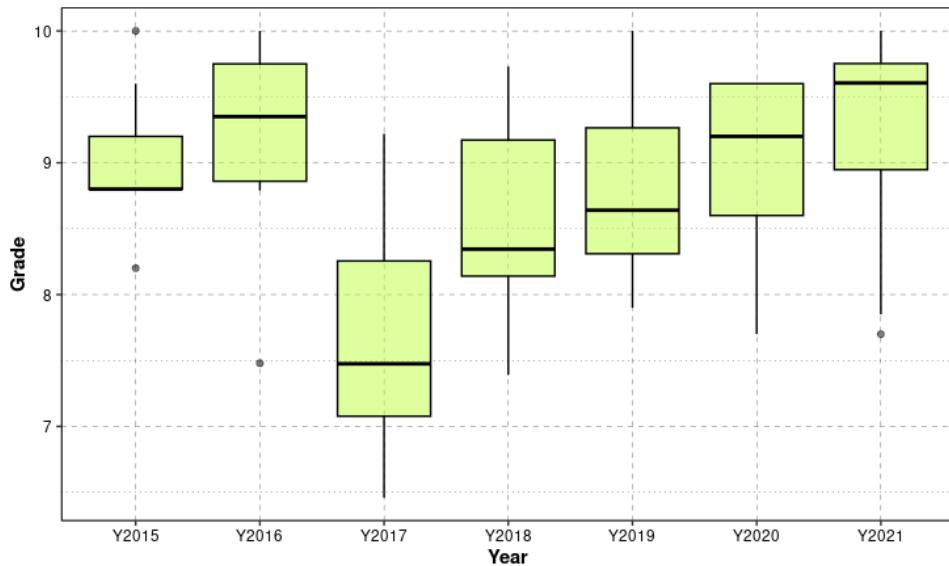


Figura 58: Boxplot de las calificaciones por año.

Cuadro 12: Resultados del test ANOVA de un solo factor (calificaciones).

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

Cuadro 13: Test HSD de Tukey (Honestly-significance-difference) de las calificaciones por años.

	diff	lwr	upr	p adj
Y ₂₀₁₆ -Y ₂₀₁₅	0.18	-0.87	1.22	1.00
Y ₂₀₁₇ -Y ₂₀₁₅	-1.35	-2.52	-0.19	0.01
Y ₂₀₁₈ -Y ₂₀₁₅	-0.45	-1.47	0.57	0.83
Y ₂₀₁₉ -Y ₂₀₁₅	-0.24	-1.23	0.76	0.99
Y ₂₀₂₀ -Y ₂₀₁₅	-0.10	-1.06	0.85	1.00
Y ₂₀₂₁ -Y ₂₀₁₅	0.22	-0.70	1.14	0.99
Y ₂₀₁₇ -Y ₂₀₁₆	-1.53	-2.70	-0.36	0.00
Y ₂₀₁₈ -Y ₂₀₁₆	-0.63	-1.64	0.39	0.51
Y ₂₀₁₉ -Y ₂₀₁₆	-0.41	-1.41	0.58	0.87
Y ₂₀₂₀ -Y ₂₀₁₆	-0.28	-1.24	0.68	0.97
Y ₂₀₂₁ -Y ₂₀₁₆	0.05	-0.88	0.97	1.00
Y ₂₀₁₈ -Y ₂₀₁₇	0.90	-0.24	2.05	0.21
Y ₂₀₁₉ -Y ₂₀₁₇	1.12	-0.00	2.24	0.05
Y ₂₀₂₀ -Y ₂₀₁₇	1.25	0.16	2.34	0.01
Y ₂₀₂₁ -Y ₂₀₁₇	1.57	0.52	2.63	0.00
Y ₂₀₁₉ -Y ₂₀₁₈	0.21	-0.75	1.18	0.99
Y ₂₀₂₀ -Y ₂₀₁₈	0.34	-0.59	1.28	0.92
Y ₂₀₂₁ -Y ₂₀₁₈	0.67	-0.22	1.56	0.27
Y ₂₀₂₀ -Y ₂₀₁₉	0.13	-0.78	1.04	1.00
Y ₂₀₂₁ -Y ₂₀₁₉	0.46	-0.41	1.32	0.68
Y ₂₀₂₁ -Y ₂₀₂₀	0.33	-0.50	1.15	0.89

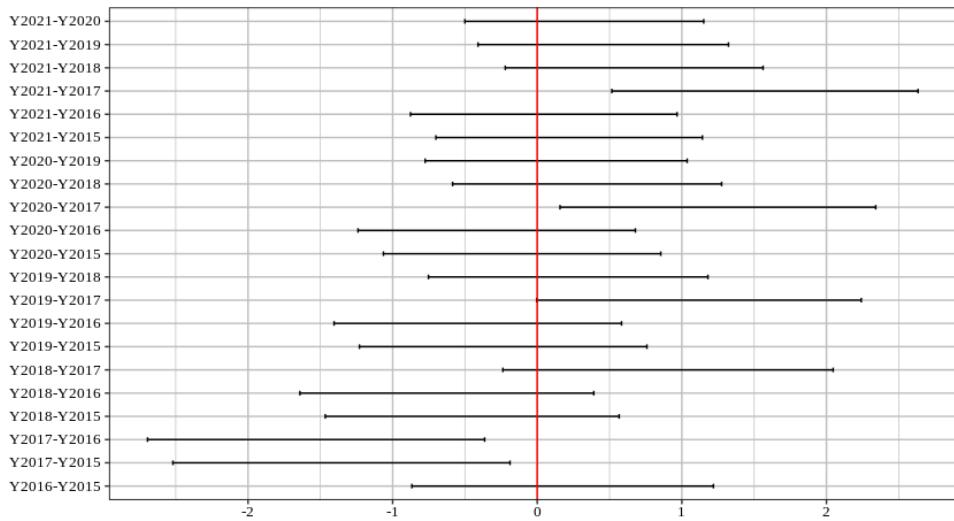


Figura 59: Intervalos de confianza de las calificaciones por años.

8.2 NÚMERO TOTAL DE PROBLEMAS RESUELTOS (P)

Podemos observar que ha habido variaciones perceptibles durante los distintos cursos académicos estudiados, siendo un caso especial el año 2018 en el que hubo muchos grupos que no resolvieron todos los problemas (Figura 60).

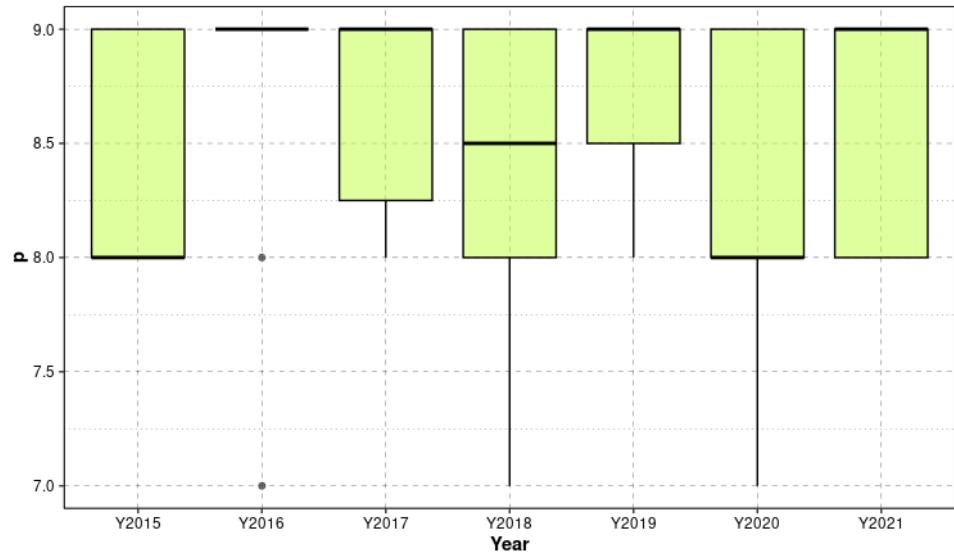


Figura 60: Boxplot del número de problemas resueltos por año.

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 14 ($p = 0,2835 >$

0,05). 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 es uniforme (cualquier variación es debida al azar).

Además, realizando un test de Tukey por pares de años (Tabla 15) se observa que todos los pares pueden considerarse estadísticamente iguales. La Figura 61 muestra los intervalos de confianza de todas las diferencias entre las distintas parejas de años.

Cuadro 14: Resultados del test ANOVA de un solo factor (número de problemas resueltos).

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

Cuadro 15: 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

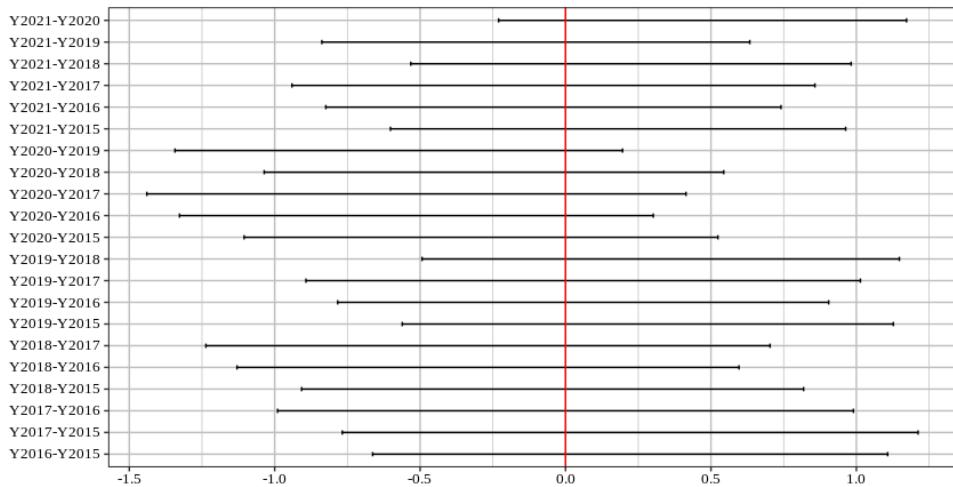


Figura 61: Intervalos de confianza del número de problemas resueltos por años.

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 62a y 62b).

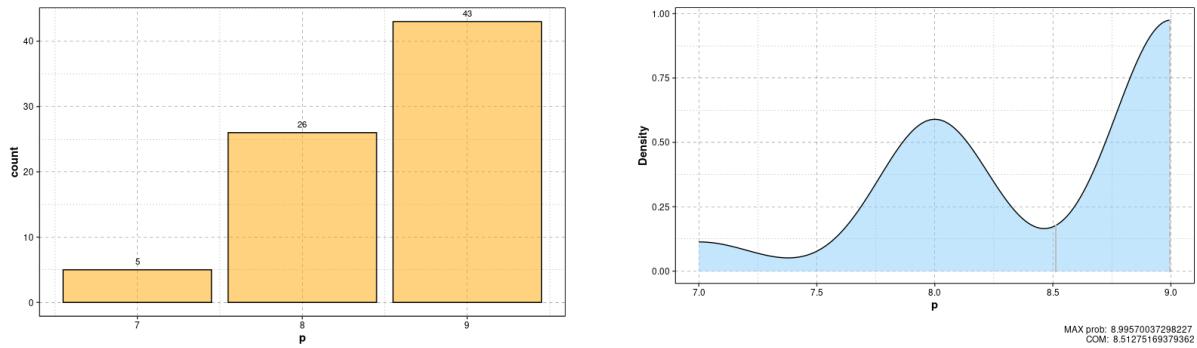


Figura 62: Distribución del número de problemas resueltos por cada grupo de alumnos.

8.3 FINALIZAR LA PRÁCTICA (FT)

Como vemos en la Figura 63, el punto de finalización de la práctica normalizado tiende a estar cerca del final de la misma.

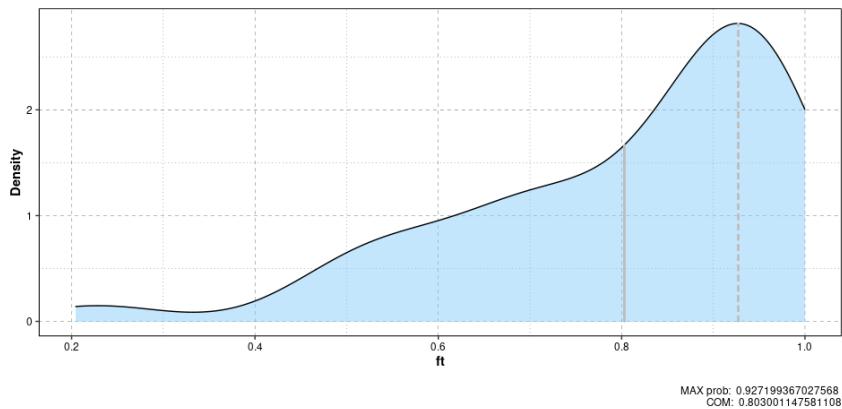


Figura 63: Función de densidad de probabilidad del momento en el que los distintos grupos de prácticas finalizan la práctica.

8.4 NÚMERO DE SESIONES REALIZADAS (s)

El número de sesiones realizadas por los diferentes grupos de prácticas se ha analizado con anterioridad. En la Sección 6.4.1 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 (Sección 6.4.3).

8.5 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 (se normaliza porque la duración de la práctica no es la misma todos los años). Como podemos ver en su función de densidad (Figura 64) y en su gráfico Q-Q (Figura 65), se aproxima a una distribución normal, un poco ladeada hacia la derecha. Esto puede deberse a que los distintos equipos tienden a abrir los problemas cuando se va aproximando la fecha de entrega de la práctica.

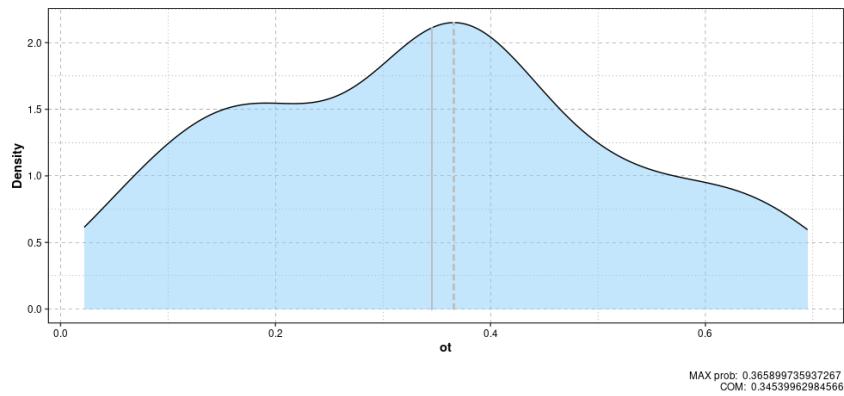


Figura 64: Función de densidad de probabilidad del momento en el que los distintos grupos de prácticas abren por primera vez un problema.

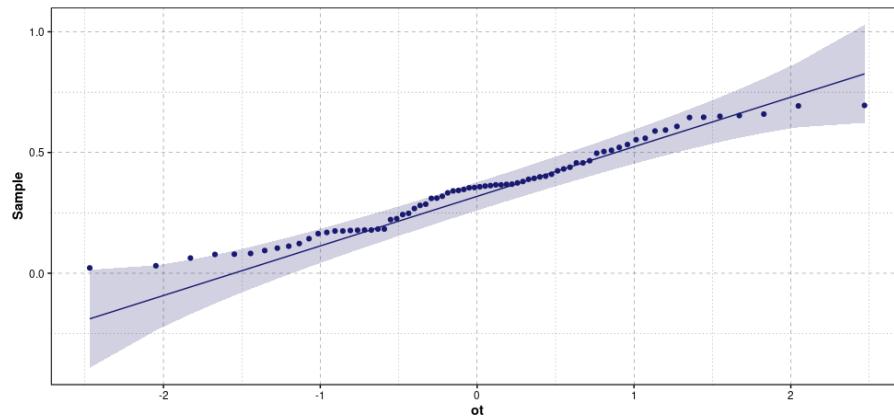
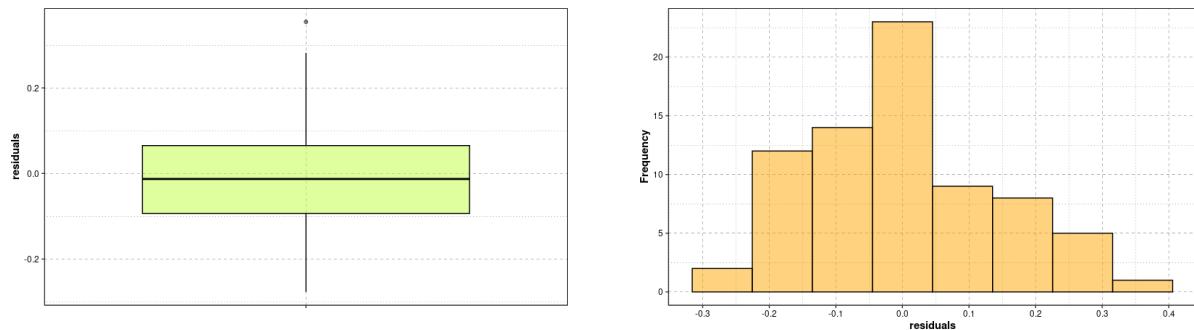


Figura 65: Gráfico Q-Q del instante temporal en el que se abren los problemas por primera vez normalizado.

En las Figuras 66a y 66b pueden verse el boxplot de los residuos y el histograma de los mismos respectivamente.



(a) Boxplot de los residuos del instante temporal en el que se abren los problemas por primera vez.

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

Figura 66: 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 67). El test ANOVA de un sólo factor ha confirmado lo observado (se ha obtenido $p = 4,05e - 06 < 0,05$ como puede verse en la Tabla 16). Igualmente, el test de Kruskal-Wallis coincide con lo anteriormente visto ($p - value = 4,633e - 05 < 0,05$).

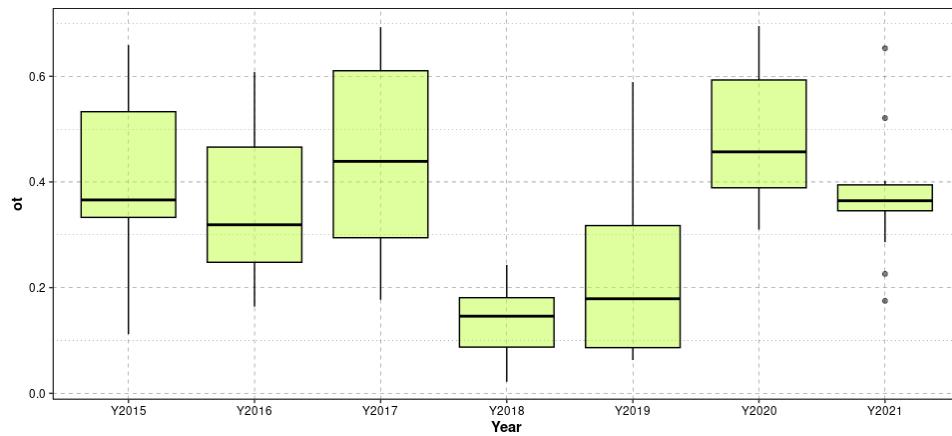


Figura 67: Boxplot del momento en el que se abre un problema por primera vez por año.

Cuadro 16: 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)
data[[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 17 y Figura 68).

Cuadro 17: Test HSD de Tukey (Honestly-significance-difference) del momento exacto en el que se abre un problema por primera por año.

	diff	lwr	upr	p adj
Y ₂₀₁₆ -Y ₂₀₁₅	-0.02	-0.23	0.18	1.00
Y ₂₀₁₇ -Y ₂₀₁₅	0.05	-0.18	0.29	0.99
Y ₂₀₁₈ -Y ₂₀₁₅	-0.26	-0.46	-0.06	0.00
Y ₂₀₁₉ -Y ₂₀₁₅	-0.15	-0.35	0.04	0.22
Y ₂₀₂₀ -Y ₂₀₁₅	0.09	-0.10	0.28	0.74
Y ₂₀₂₁ -Y ₂₀₁₅	-0.02	-0.20	0.17	1.00
Y ₂₀₁₇ -Y ₂₀₁₆	0.08	-0.15	0.31	0.94
Y ₂₀₁₈ -Y ₂₀₁₆	-0.23	-0.44	-0.03	0.01
Y ₂₀₁₉ -Y ₂₀₁₆	-0.13	-0.33	0.07	0.42
Y ₂₀₂₀ -Y ₂₀₁₆	0.12	-0.07	0.31	0.50
Y ₂₀₂₁ -Y ₂₀₁₆	0.01	-0.18	0.19	1.00
Y ₂₀₁₈ -Y ₂₀₁₇	-0.31	-0.54	-0.09	0.00
Y ₂₀₁₉ -Y ₂₀₁₇	-0.21	-0.43	0.01	0.08
Y ₂₀₂₀ -Y ₂₀₁₇	0.04	-0.18	0.26	1.00
Y ₂₀₂₁ -Y ₂₀₁₇	-0.07	-0.28	0.14	0.94
Y ₂₀₁₉ -Y ₂₀₁₈	0.10	-0.09	0.29	0.66
Y ₂₀₂₀ -Y ₂₀₁₈	0.35	0.17	0.54	0.00
Y ₂₀₂₁ -Y ₂₀₁₈	0.24	0.06	0.42	0.00
Y ₂₀₂₀ -Y ₂₀₁₉	0.25	0.07	0.43	0.00
Y ₂₀₂₁ -Y ₂₀₁₉	0.14	-0.04	0.31	0.21
Y ₂₀₂₁ -Y ₂₀₂₀	-0.11	-0.28	0.05	0.38

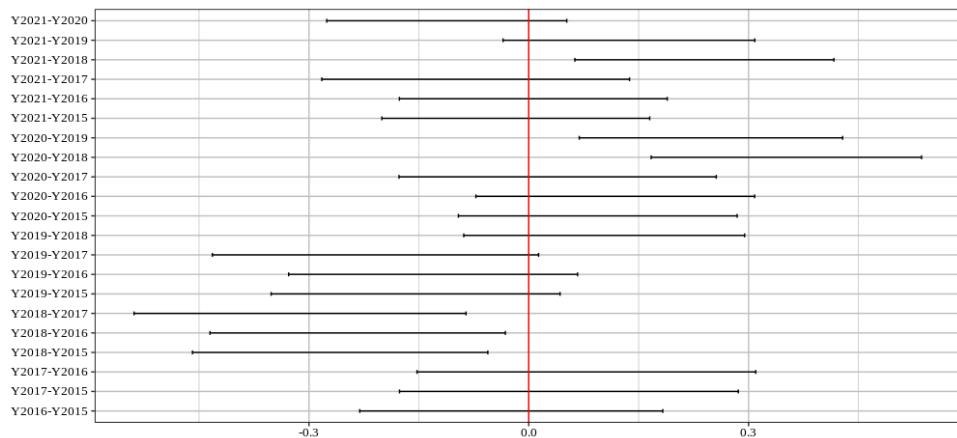


Figura 68: Intervalos de confianza del momento en el que se abre un problema por años.

8.6 TASA DE FALLO (FR)

Podemos ver que la distribución de la tasa de fallo o el número de intentos fallidos para resolver un problema dividido el número de intentos totales de resolución de ese mismo problema se aproxima a una distribución normal, un poco ladeada hacia la derecha a través de su función de densidad (Figura 69) y su gráfico Q-Q (Figura 70).

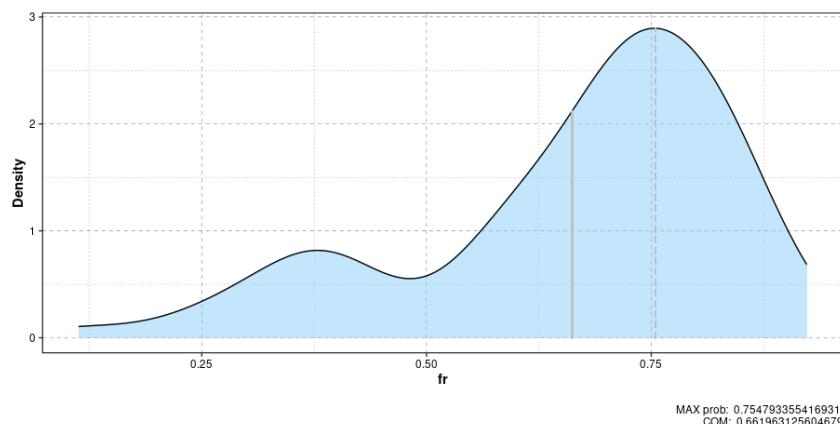


Figura 69: Función de densidad de probabilidad de la tasa de fallo en la resolución de problemas.

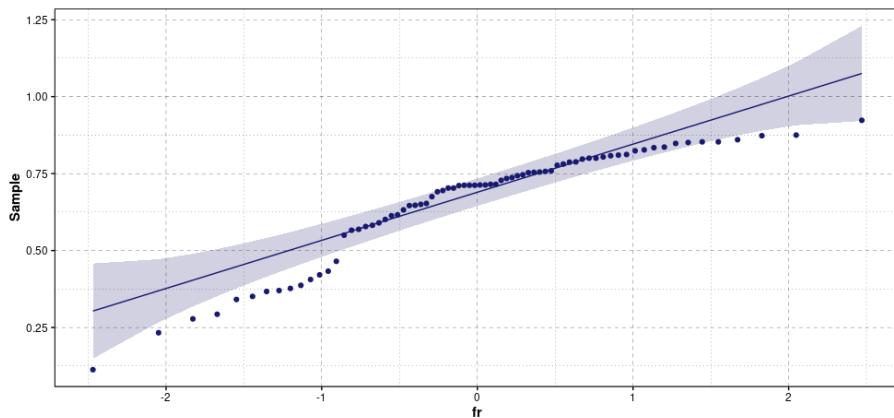
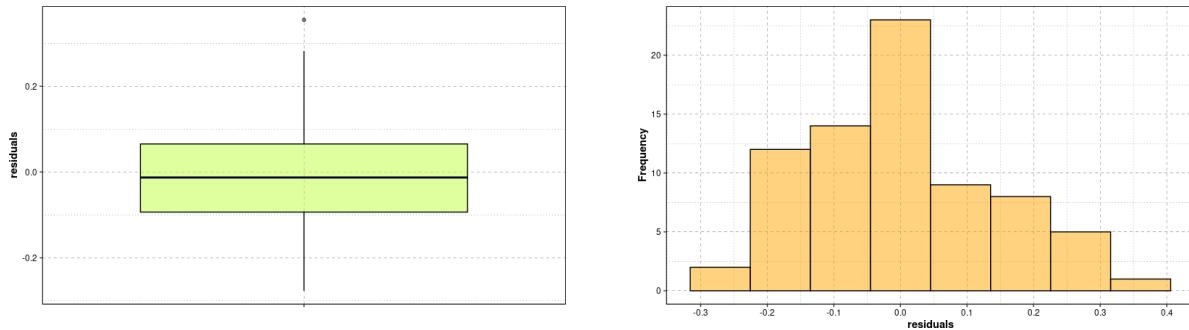


Figura 70: Gráfico Q-Q de la tasa de fallo en la resolución de problemas.

En las Figuras 71a y 71b podemos ver el boxplot de los residuos de la variable tasa de fallo (fr) junto con el histograma de los mismos.



(a) Boxplot de los residuos de la tasa de fallo en la resolución de problemas.

(b) Histograma de los residuos de la tasa de fallo en la resolución de problemas.

Figura 71: Distribución de la tasa de fallo en la resolución de problemas por parte de las diferentes agrupaciones de alumnos.

A continuación, se realizará un estudio por años de la tasa de fallo en la resolución de problemas por parte del alumnado. Como podemos observar en la Figura 72, hay una variación muy perceptible en la tasa de fallo a lo largo de los años. La realización del test ANOVA de un factor, cuyos resultados se muestran en la Tabla 18, ratifica que hay diferencias significativas ($p = 6,55e - 09 < 0,05$). Igualmente, los resultados del test de Kruskal-Wallis confirman lo anterior ($p - value = 0,0001689 < 0,05$). Posteriormente, se ha realizado el test de Tukey por pares y se ha visto que hay diferencias entre algunos de los cursos académicos considerados (Tabla 19 y Figura 73).

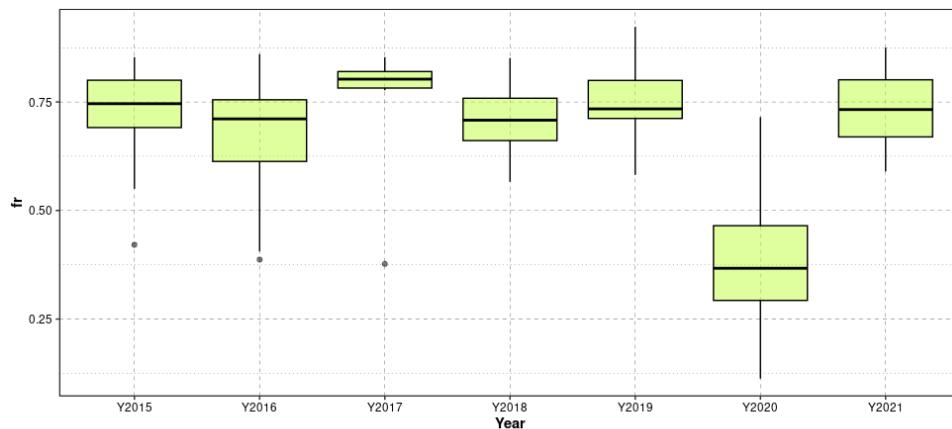


Figura 72: Boxplot de la tasa de fallo en la resolución de problemas por año.

Cuadro 18: Resultados del test ANOVA de un solo factor (tasa de fallo en la resolución de los problemas propuestos).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data[[Variable]]	6	1.19	0.20	11.68	6.55e-09
Residuals	67	1.14	0.02		

Cuadro 19: Test HSD de Tukey (Honestly-significance-difference) de la tasa de fallo en la resolución de los problemas propuestos por año.

	diff	lwr	upr	p adj
Y ₂₀₁₆ -Y ₂₀₁₅	-0.05	-0.24	0.14	0.98
Y ₂₀₁₇ -Y ₂₀₁₅	0.03	-0.18	0.24	1.00
Y ₂₀₁₈ -Y ₂₀₁₅	0.00	-0.18	0.19	1.00
Y ₂₀₁₉ -Y ₂₀₁₅	0.03	-0.14	0.21	1.00
Y ₂₀₂₀ -Y ₂₀₁₅	-0.32	-0.49	-0.15	0.00
Y ₂₀₂₁ -Y ₂₀₁₅	0.03	-0.14	0.19	1.00
Y ₂₀₁₇ -Y ₂₀₁₆	0.08	-0.13	0.29	0.91
Y ₂₀₁₈ -Y ₂₀₁₆	0.05	-0.13	0.24	0.97
Y ₂₀₁₉ -Y ₂₀₁₆	0.08	-0.09	0.26	0.78
Y ₂₀₂₀ -Y ₂₀₁₆	-0.27	-0.44	-0.10	0.00
Y ₂₀₂₁ -Y ₂₀₁₆	0.07	-0.09	0.24	0.81
Y ₂₀₁₈ -Y ₂₀₁₇	-0.03	-0.23	0.18	1.00
Y ₂₀₁₉ -Y ₂₀₁₇	0.01	-0.20	0.21	1.00
Y ₂₀₂₀ -Y ₂₀₁₇	-0.35	-0.54	-0.15	0.00
Y ₂₀₂₁ -Y ₂₀₁₇	-0.00	-0.19	0.19	1.00
Y ₂₀₁₉ -Y ₂₀₁₈	0.03	-0.14	0.20	1.00
Y ₂₀₂₀ -Y ₂₀₁₈	-0.32	-0.49	-0.15	0.00
Y ₂₀₂₁ -Y ₂₀₁₈	0.02	-0.14	0.18	1.00
Y ₂₀₂₀ -Y ₂₀₁₉	-0.35	-0.51	-0.19	0.00
Y ₂₀₂₁ -Y ₂₀₁₉	-0.01	-0.16	0.15	1.00
Y ₂₀₂₁ -Y ₂₀₂₀	0.34	0.19	0.49	0.00

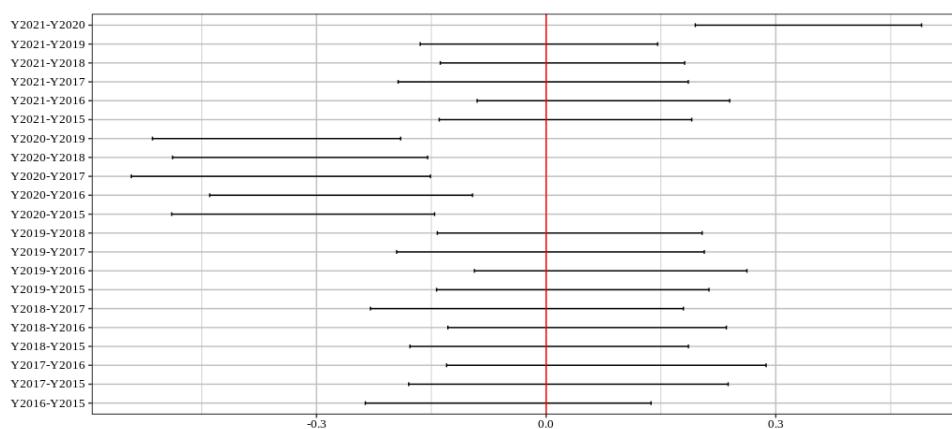


Figura 73: Intervalos de confianza de la tasa de fallo en la resolución de los problemas propuestos por años.

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

Número de veces que se abre un mismo problema sin resolver hasta que se resuelve por primera vez normalizado. 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 74, lo que indica que los grupos suelen emplear poco tiempo en resolver un problema por primera vez.

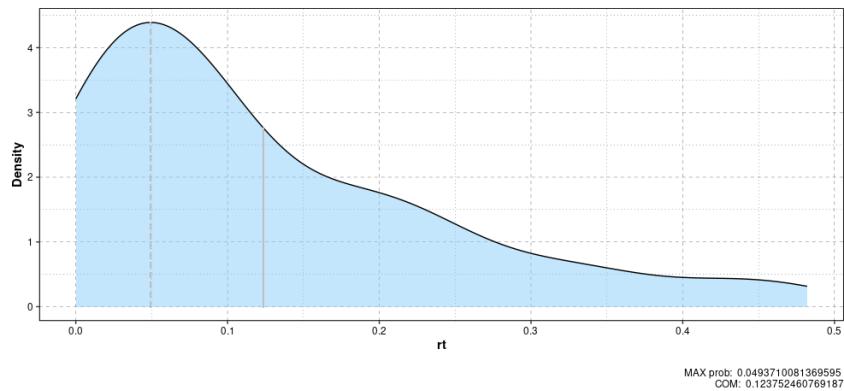
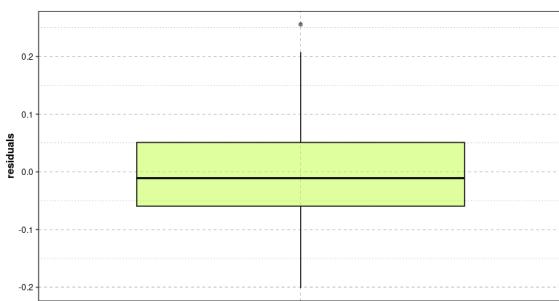
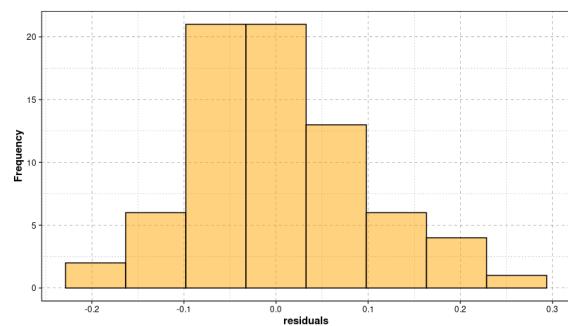


Figura 74: Función de densidad de probabilidad del tiempo que los distintos grupos de prácticas dedican a la resolución de problemas por primera vez.

Sin embargo, podemos ver que hay algunos grupos que emplean muy pocos intentos para resolver un problema o que, por el contrario, emplean más intentos que la mayoría (Figuras 75a y 75b).



(a) Boxplot de los residuos del número de intentos necesarios para resolver los problemas por primera vez normalizado.



(b) Histograma de los residuos del número de intentos necesarios para resolver los problemas por primera vez normalizado.

Figura 75: Distribución de los residuos del número de intentos necesarios para resolver los problemas por primera vez normalizado.

No obstante, podemos considerar que la regresión es aceptable (Figura 76).

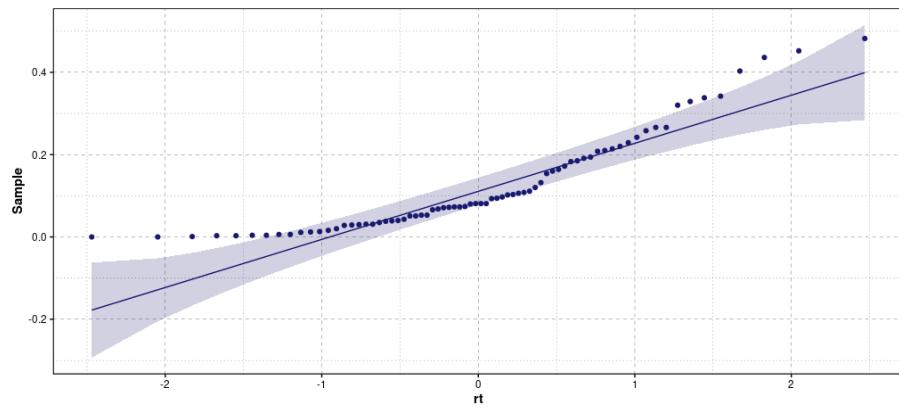


Figura 76: Gráfico Q-Q del número de intentos necesarios para resolver los problemas por primera vez normalizado.

A continuación, se realizará un estudio por años del número de intentos antes de resolver un problema la primera vez realizados. Como podemos observar en la Figura 77, 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 20) ratifica que hay diferencias significativas ($p = 8,85e - 07 < 0,05$). Igualmente, los resultados del test de Kruskal-Wallis confirman lo anterior ($p - value = 9,767e - 07 < 0,05$). Posteriormente, se ha realizado el test de Tukey por pares y se ha visto que hay diferencias entre algunos de los cursos académicos considerados (Tabla 21). Esto mismo puede observarse en la Figura 78.

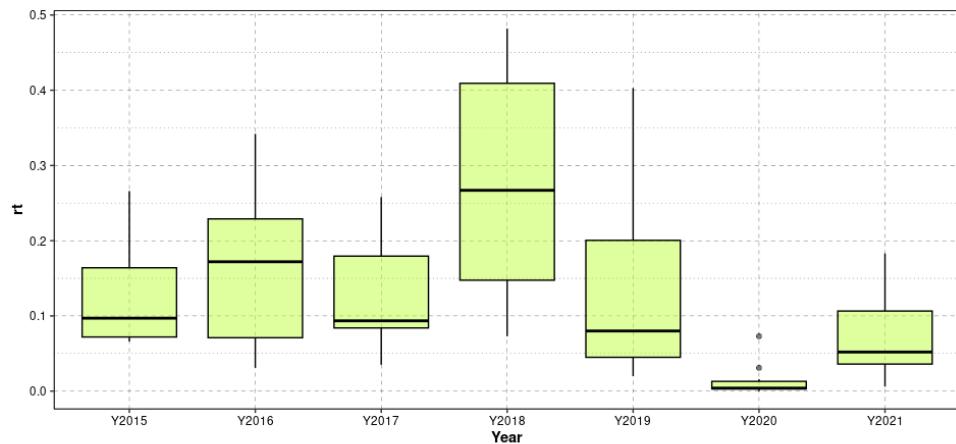


Figura 77: Boxplot del número de intentos para resolver un problema por año.

Cuadro 20: Resultados del test ANOVA de un solo factor (número de intentos necesarios para resolver un problema por primera vez normalizado).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data[[Variable]]	6	0.45	0.07	8.37	8.85e-07
Residuals	67	0.60	0.01		

Cuadro 21: Test HSD de Tukey (Honestly-significance-difference) del número de intentos realizados para resolver un problema normalizado por año.

	diff	lwr	upr	p adj
Y2016-Y2015	0.04	-0.10	0.17	0.98
Y2017-Y2015	-0.00	-0.15	0.15	1.00
Y2018-Y2015	0.15	0.01	0.28	0.02
Y2019-Y2015	0.02	-0.11	0.15	1.00
Y2020-Y2015	-0.12	-0.24	0.01	0.08
Y2021-Y2015	-0.05	-0.17	0.07	0.81
Y2017-Y2016	-0.04	-0.19	0.11	0.99
Y2018-Y2016	0.11	-0.02	0.24	0.19
Y2019-Y2016	-0.02	-0.15	0.11	1.00
Y2020-Y2016	-0.15	-0.28	-0.03	0.01
Y2021-Y2016	-0.09	-0.21	0.03	0.24
Y2018-Y2017	0.15	-0.00	0.29	0.06
Y2019-Y2017	0.02	-0.13	0.16	1.00
Y2020-Y2017	-0.12	-0.26	0.03	0.19
Y2021-Y2017	-0.05	-0.19	0.08	0.90
Y2019-Y2018	-0.13	-0.25	-0.00	0.05
Y2020-Y2018	-0.26	-0.38	-0.14	0.00
Y2021-Y2018	-0.20	-0.32	-0.08	0.00
Y2020-Y2019	-0.13	-0.25	-0.02	0.02
Y2021-Y2019	-0.07	-0.19	0.04	0.45
Y2021-Y2020	0.06	-0.05	0.17	0.58

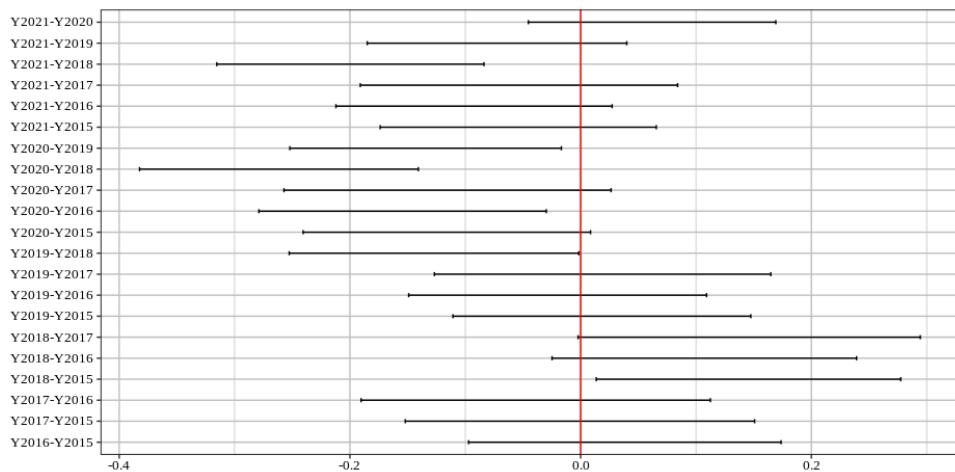


Figura 78: Intervalos de confianza del número de intentos antes de resolver un problema normalizado por años.

8.8 EXPLORACIÓN DE NUEVAS VÍAS Y MEJORAS (PS)

Curiosamente, algunos grupos siguen trabajando en un problema que ya han resuelto (Figura 79).

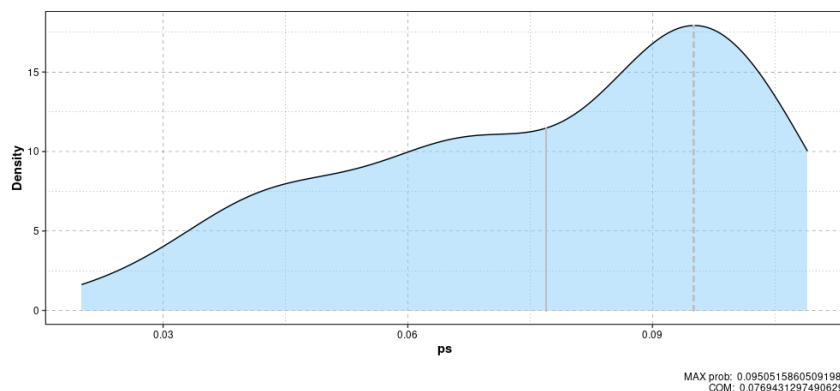


Figura 79: Función de densidad de probabilidad del tiempo que los distintos grupos de prácticas emplean en un problema tras su resolución.

8.9 RESOLVER UN PROBLEMA POR PRIMERA VEZ (ST)

Momento exacto en el que se consigue resolver cada problema por primera vez, normalizado para poder compararlo (ya que la duración de la práctica no es la misma todos los años). Como podemos ver en la Figura 80, 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.

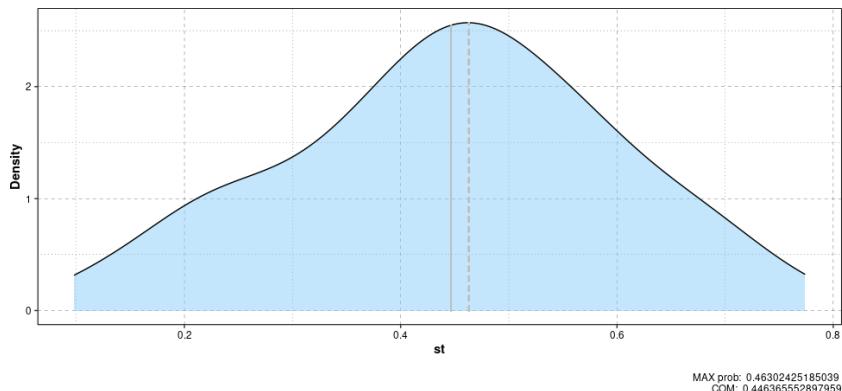
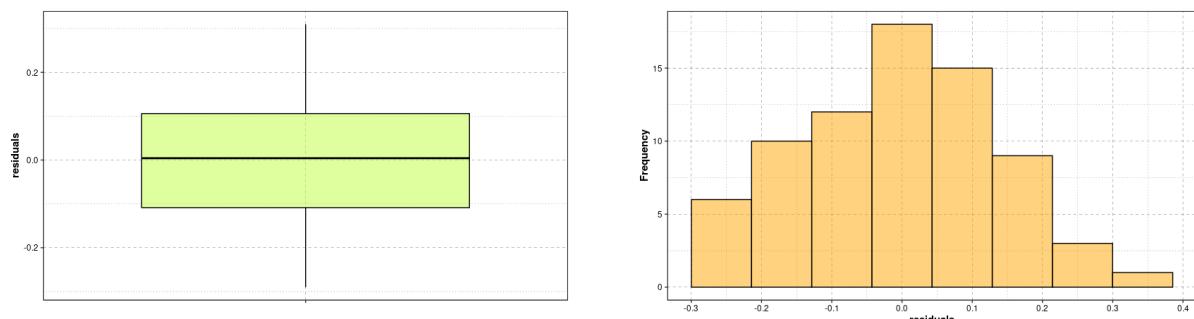


Figura 80: Función de densidad de probabilidad del momento en el que los distintos grupos de prácticas resuelven por primera vez un problema.

Podemos observar que no hay presencia de outliers (Figuras 81a y 81b).



(a) Boxplot de los residuos del momento en el que se resuelven los problemas por primera vez normalizado.

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

Figura 81: Distribución de los residuos del número de intentos desde el comienzo de la práctica hasta la resolución de los problemas por primera vez normalizado.

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 82). Esto se confirma tras la realización del test ANOVA de un solo factor (Tabla 22) 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 23).

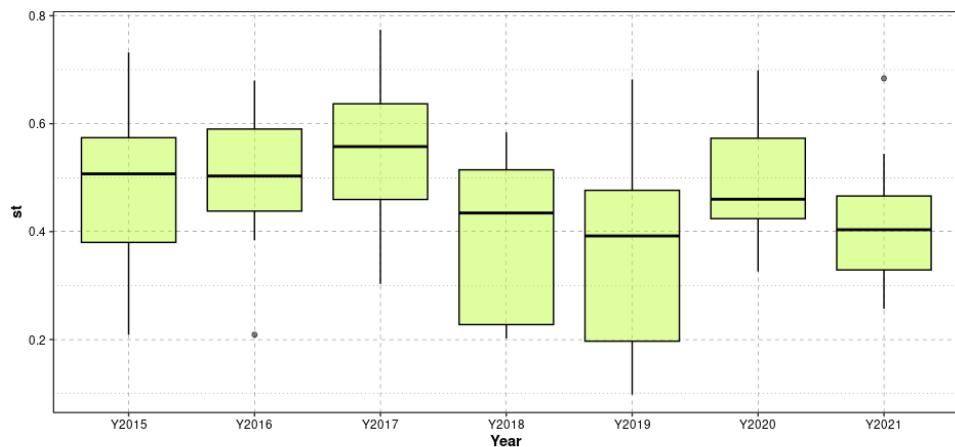


Figura 82: Boxplot del momento en el que se resuelve un problema por primera vez por año.

Cuadro 22: Resultados del test ANOVA de un solo factor (momento en el que se resuelve un problema por primera vez normalizado).

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

Cuadro 23: Test HSD de Tukey (Honestly-significance-difference) del momento en el que se resuelve un problema por primera vez normalizado, por año.

	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 83.

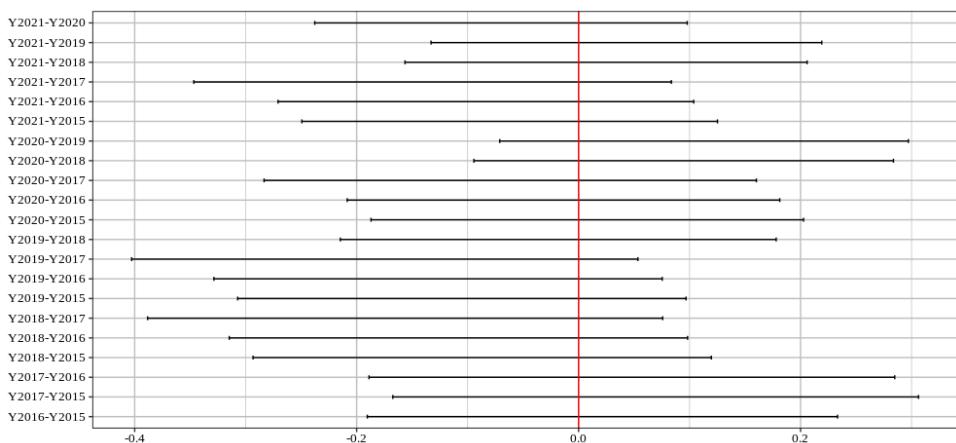


Figura 83: Intervalos de confianza del momento en el que se resuelve un problema por primera vez normalizado, por años.

8.10 SIGUIENDO EL PLAN DEL PROFESOR (SQ)

Se incorpora una medida de similaridad, en este caso la medida de similaridad del coseno (a la que denotaremos por 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 ($P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9$). 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 $P_1 P_3 P_4 P_5 P_7 P_2 P_6 P_8 P_9$).

La medida de similaridad del coseno se define sobre dos vectores definidos sobre un espacio prehilbertiano¹. Así pues, la medida de similaridad del coseno es el coseno del ángulo entre los vectores. Es decir, es el producto escalar de los vectores dividido por el producto de sus longitudes (Ecuación 34, donde A y B son dos vectores n dimensionales representando la sucesión de resolución de n problemas de dos grupos de prácticas). De ello se deduce que la semejanza coseno no depende de las magnitudes de los vectores, sino sólo de su ángulo. En la Figura 84 se representa gráficamente la situación descrita.

¹ Espacio vectorial provisto de un producto escalar.

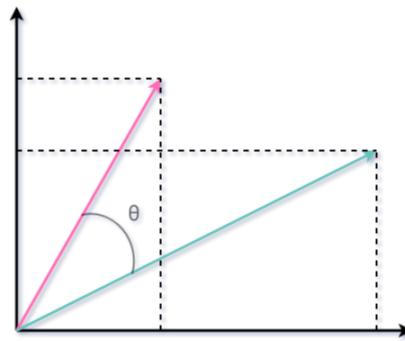


Figura 84: Representación gráfica del ángulo entre dos vectores con magnitudes positivas en un espacio prehilbertiano de dimensión dos.

$$\text{cosine similarity } (A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^n A_i \cdot B_i}{\sqrt{\sum_{i=1}^n A_i^2 \cdot \sum_{i=1}^n B_i^2}} \quad (34)$$

Trivialmente, la similitud coseno pertenece siempre al intervalo $[-1, 1]$. Por ejemplo, dos vectores proporcionales tienen una similitud coseno de 1, dos vectores ortogonales tienen una similitud de 0 y dos vectores opuestos tienen una similitud de -1 . Sin embargo, en nuestro contexto, como los valores componentes de los vectores no pueden ser negativos (serán números naturales entre el uno y el nueve representando a cada uno de los problemas), la semejanza coseno está acotada en el intervalo $[0, 1]$.

Nótese que se podría haber definido una medida de similaridad basada en la distancia euclídea. Sin embargo, ésta sólo es de utilidad en espacios de dimensiones pequeñas.

En la Figura 85 podemos ver la función de densidad de probabilidad de la medida de rendimiento que estamos considerando.

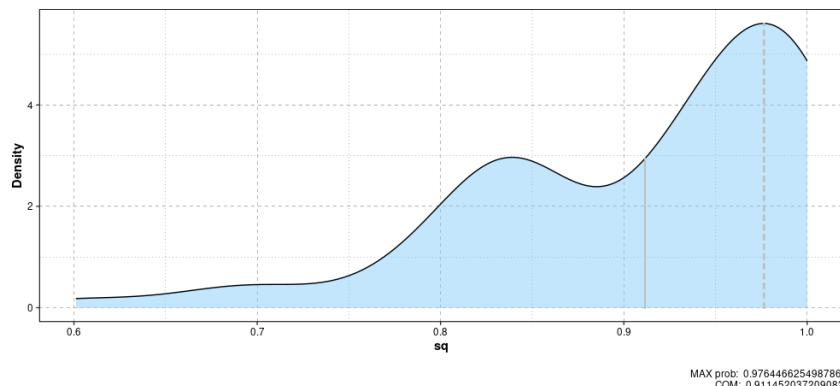


Figura 85: Función de densidad de probabilidad del seguimiento del orden de resolución previsto por el profesor que realizan los diferentes grupos de prácticas.

Parte IV

PLANIFICACIÓN DEL PROYECTO

Estructuración en objetivos, división en sprints y seguimiento.

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 las siguientes iteraciones:

- Realización de un estudio multianual y segmentado por calificaciones y primeros resultados de homogeneidad de las muestras transversal por años mediante la realización 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.

10

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

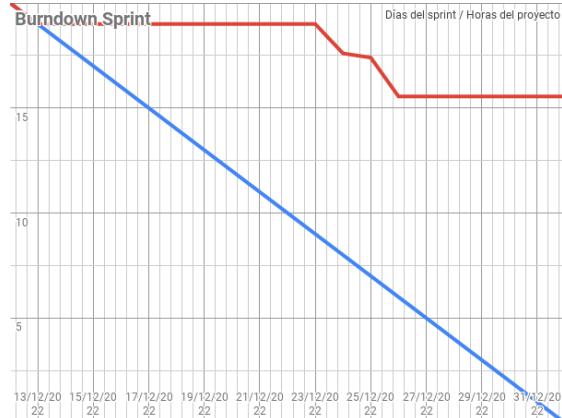
El proyecto se ha dividido en diez sprints de dos semanas cada uno. Además, durante el transcurso de cada sprint se ha realizado un seguimiento del trabajo realizado en el mismo, aportando *burndown charts* de cada uno de ellos.

Un *burndown chart* es una gráfica en la que se muestra el progreso de un proyecto durante cierto periodo de tiempo preestablecido (un sprint, una release o el proyecto completo, por ejemplo). Para la construcción de los mismos se requieren los siguientes elementos:

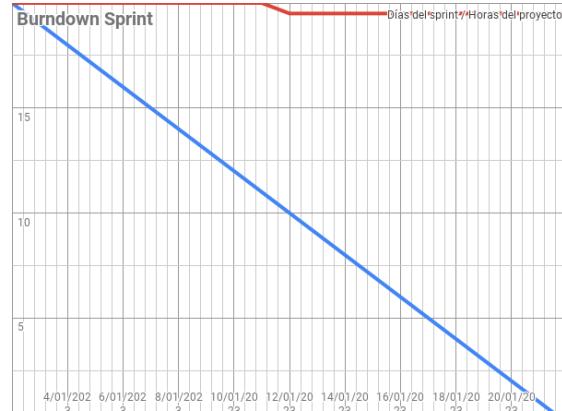
- El periodo de *tiempo a analizar*, que se corresponderá con el eje X de la gráfica. El inicio del periodo temporal vendrá representado por $x = 0$ mientras que el final del periodo temporal será representado por $x = t$ donde t es la duración del periodo. Si observamos el burndown chart global del proyecto (Figura global), estos dos valores se corresponderán, respectivamente, con las fechas de inicio y finalización del mismo.
- La cantidad de trabajo a realizar, que se corresponderá con el eje Y de la gráfica y representa el trabajo planificado que se deberá realizar en el periodo de tiempo a analizar. Cuando se realiza la estimación de las tareas, independientemente de la unidad de estimación empleada, se obtiene una cantidad de trabajo a realizar (en horas, por ejemplo). Así pues, la cantidad de trabajo restante irá decreciendo conforme el tiempo vaya avanzando.
- Una referencia ideal, que será la línea diagonal trazada desde la esquina superior izquierda hasta la esquina inferior derecha de la gráfica. Se trata de una representación de la relación ideal entre la disminución de la cantidad de trabajo y el tiempo dedicado durante el transcurso de la fase correspondiente. Esto es, cuanto más se aproxime la línea real a la ideal se estará trabajando de mejor manera en relación a la consecución de los objetivos marcados.

Adicionalmente, si se poseen conocimientos y experiencia trabajando con herramientas ágiles, se puede obtener información útil sobre el seguimiento del proyecto con el fin de impulsar buenas prácticas e impulsar el desarrollo del mismo.

10.1 ANÁLISIS DE CADA SPRINT



(a) Burndown chart del sprint 1 (periodo del 12/12/2022 al 01/01/2023).



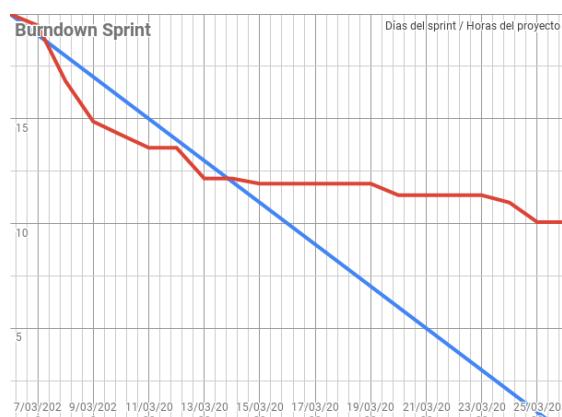
(b) Burndown chart del sprint 2 (periodo del 02/01/2023 al 22/01/2023).



(c) Burndown chart del sprint 3 (periodo del 23/01/2023 al 12/02/2023).



(d) Burndown chart del sprint 4 (periodo del 13/02/2023 al 05/03/2023).



(e) Burndown chart del sprint 5 (periodo del 06/03/2023 al 26/03/2023).



(f) Burndown chart del sprint 6 (periodo del 27/03/2023 al 16/04/2023).

Figura 86: Burndown charts de los seis primeros sprints del proyecto.

10.1.1 Sprint 1 (Figura 86a)

Este primer sprint representa una situación anómala que no se debería dar. La explicación es sencilla: las líneas rojas horizontales representan periodos de tiempo en los cuales no se ha estado trabajando en el proyecto. No obstante, esto es correcto puesto que aunque se definieron unos objetivos al principio, todavía no tenía claro el procedimiento a seguir para la consecución de los mismos. Aunque podría haberse omitido este periodo de seguimiento, se ha decidio incluirlo como muestra de una tendencia no deseable en relación al desarrollo de un proyecto.

10.1.2 Sprint 2 (Figura 86b)

En este segundo sprint tampoco se han alcanzado los objetivos de trabajo preestablecidos. No obstante, es correcto porque este periodo coincide con la realización de exámenes y se planificó previamente para el estudio de los mismos y no para continuar con la realización de este trabajo fin de grado. Como se ha comentado anteriormente, durante el desarrollo de un proyecto, esta situación no debería ocurrir bajo ninguna circunstancia.

10.1.3 Sprint 3 (Figura 86c)

En este caso, la Figura 86c muestra que se realizó una gran cantidad de trabajo focalizada en días concretos distribuidos a lo largo de todo el sprint con algunos parones entre medias. Sin embargo, a pesar de todo, no se consiguió llegar a la cantidad estipulada de trabajo al final del sprint.

10.1.4 Sprint 4 (Figura 86d)

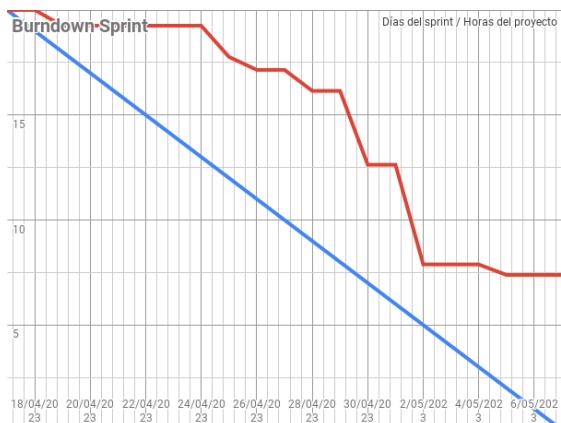
Este cuarto sprint consiste en un periodo inicial inactivo seguido de varios picos de productividad en días concretos del sprint. Finalmente, podemos observar que se realizó una gran cantidad de trabajo al final del sprint, quedándose muy cerca la cantidad de trabajo realizada de la referencia ideal.

10.1.5 Sprint 5 (*Figura 86e*)

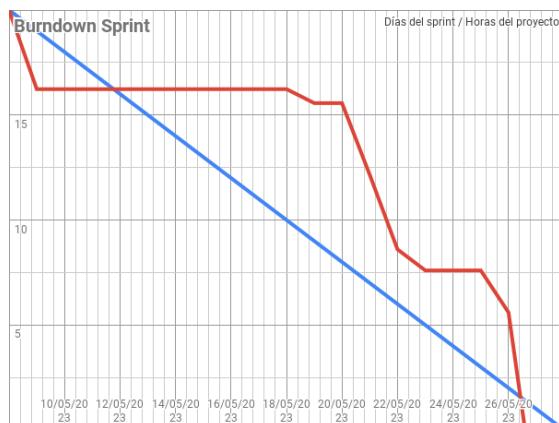
En este sprint podemos ver que se realizó más trabajo del requerido durante el primer tercio del mismo. Sin embargo, a partir de entonces, a pesar de que se avanza en días concretos, el ritmo de trabajo es mucho más lento y se pasa a no realizar la cantidad de trabajo planificado durante dicho periodo temporal.

10.1.6 Sprint 6 (*Figura 86f*)

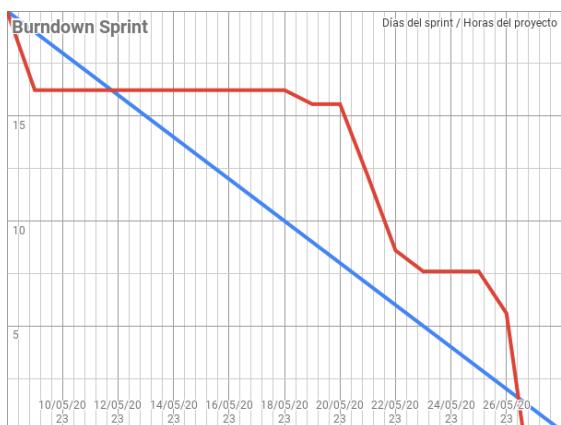
La gráfica de este sprint (Figura 86f) muestra que se realizó una gran cantidad de trabajo al principio del sprint llegando la curva de trabajo real a cortar a la curva ideal un poco antes de la mitad de la misma. No obstante, al final del sprint hubo algún día de parón y el ritmo de trabajo se redujo. En consecuencia, no se alcanzaron los objetivos preestablecidos.



(a) Burndown chart del sprint 7 (periodo del 17/04/2023 al 07/05/2023).



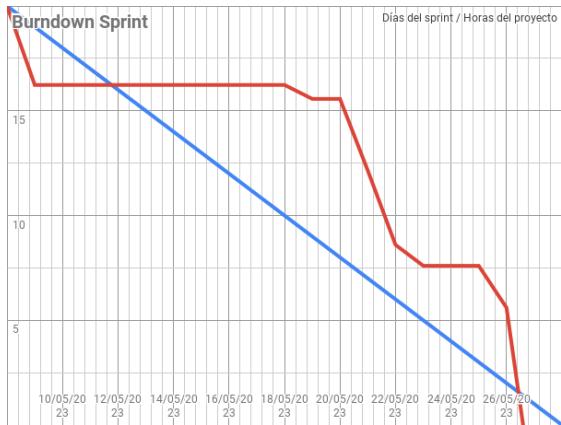
(b) Burndown chart del sprint 8 (periodo del 08/05/2023 al 28/05/2023).



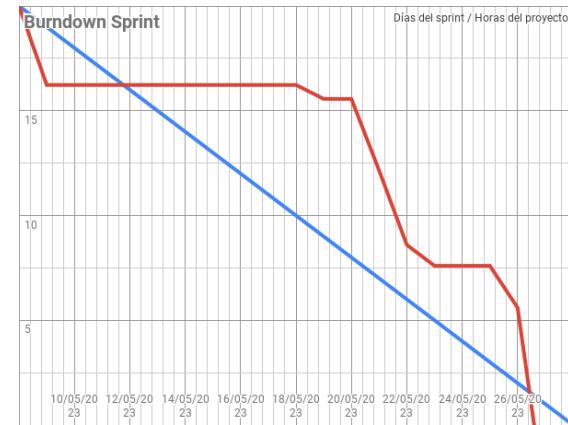
(c) Burndown chart del sprint 9 (periodo del 29/05/2023 al 18/06/2023).



(d) Burndown chart del sprint 10 (periodo del 19/06/2023 al 09/07/2023).



(e) Burndown chart del sprint 11 (periodo del 10/07/2023 al 30/07/2023).



(f) Burndown chart del sprint 12 (periodo del 27/03/2023 al 16/04/2023).

Figura 87: Burndown charts de los seis últimos sprints del proyecto.

10.1.7 Sprint 7 (Figura 87a)

En este caso, la Figura 87a revela que hubo un ritmo de trabajo insuficiente tanto al principio como al final del sprint. Sin embargo, durante todo el tramo central del mismo se estuvo trabajando de manera intensa aunque insuficiente para abarcar toda la cantidad de trabajo asignada a este sprint.

10.1.8 Sprint 8 (Figura 87b)

En este octavo sprint podemos apreciar que se realiza más trabajo del planificado en los extremos del periodo considerado gracias a la realización de grandes esfuerzos en días concretos. Esto, en cierto modo, ha servido para compensar parte del trabajo no realizado durante la parte central del sprint.

10.1.9 Sprint 9 (Figura 87c)

10.1.10 Sprint 10 (Figura 87d)

10.1.11 Sprint 11 (Figura 87e)

10.1.12 Sprint 12 (Figura 87f)

10.2 ANÁLISIS GLOBAL DEL PROYECTO

Parte V

RESULTADOS OBTENIDOS

Análisis de los resultados obtenidos.

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 88 podemos ver que no todas las métricas clásicas correlan con la misma.

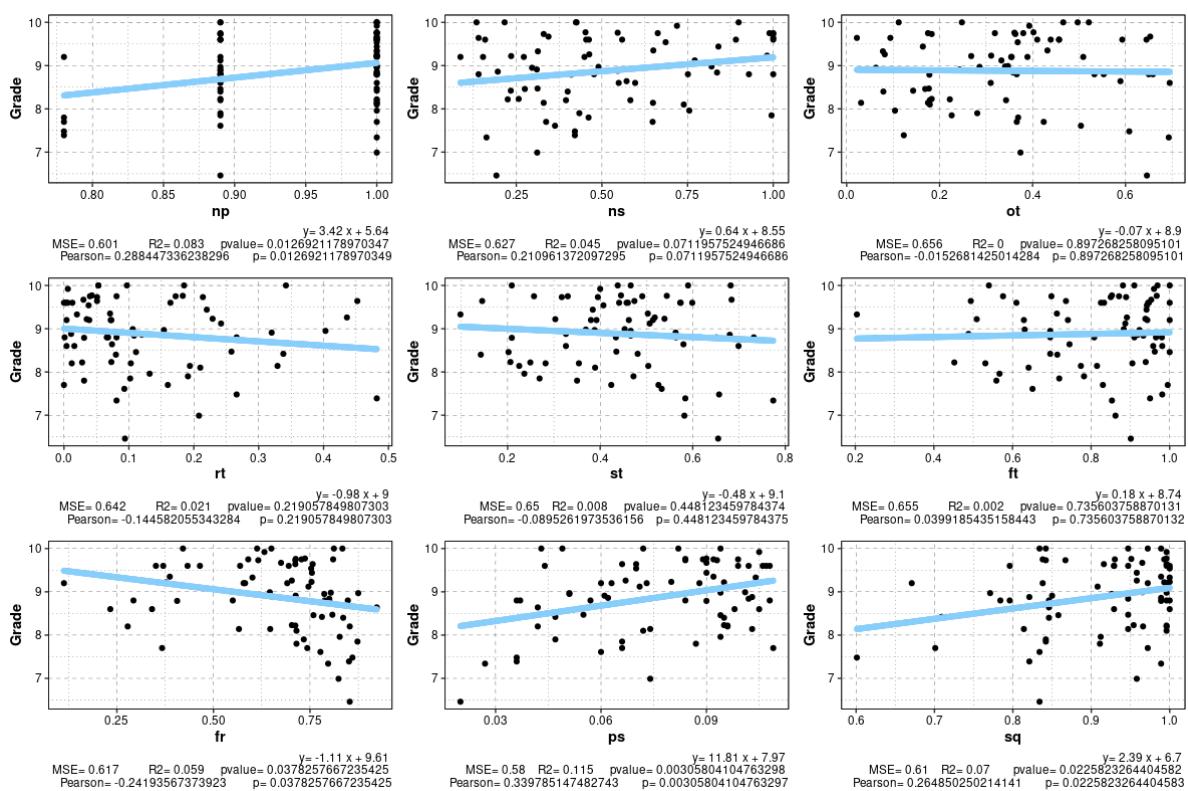


Figura 88: Correlaciones existentes entre las distintas métricas clásicas y la variable *Grade*.

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 ($p = 0,0712 < 0,1$).

Además, realizando una regresión polinomial entre dichas medidas y la variable *Grade* obtenemos que existe una relación no lineal entre las variables *np* ($p = 0,0137 < 0,05$), *fr* ($p = 0,0016 < 0,05$) y *ps* ($p = 0,0006 < 0,05$) y las calificaciones obtenidas tal y como se muestra en la Figura 89. Adicionalmente, podría existir una relación no lineal entre la variable *sq* y la variable *Grade* aunque con un grado de certeza menor ($p = 0,0695 < 0,1$).

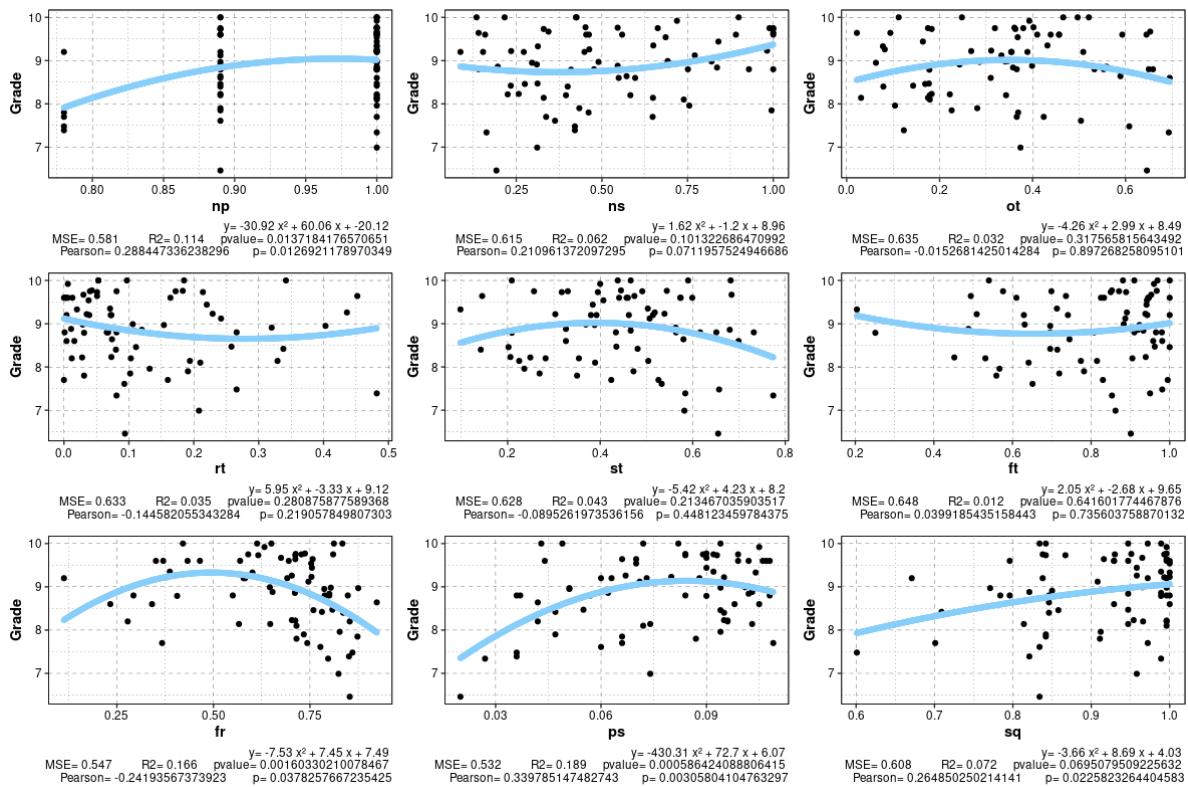


Figura 89: Regresión polinomial entre las distintas métricas clásicas y la variable *Grade*.

También estudiaremos las correlaciones existentes entre las medidas basadas en el análisis espectral de grafos y la variable *Grade*. Empezaremos estudiando la correlación lineal entre las medidas anteriormente presentadas en la Sección 5.2 y la calificación obtenida por los distintos grupos de prácticas. Como vemos en la Figura 90, únicamente existe una relación lineal positiva entre las variables *We* y *Grade* ($p = 0,0374 < 0,05$).

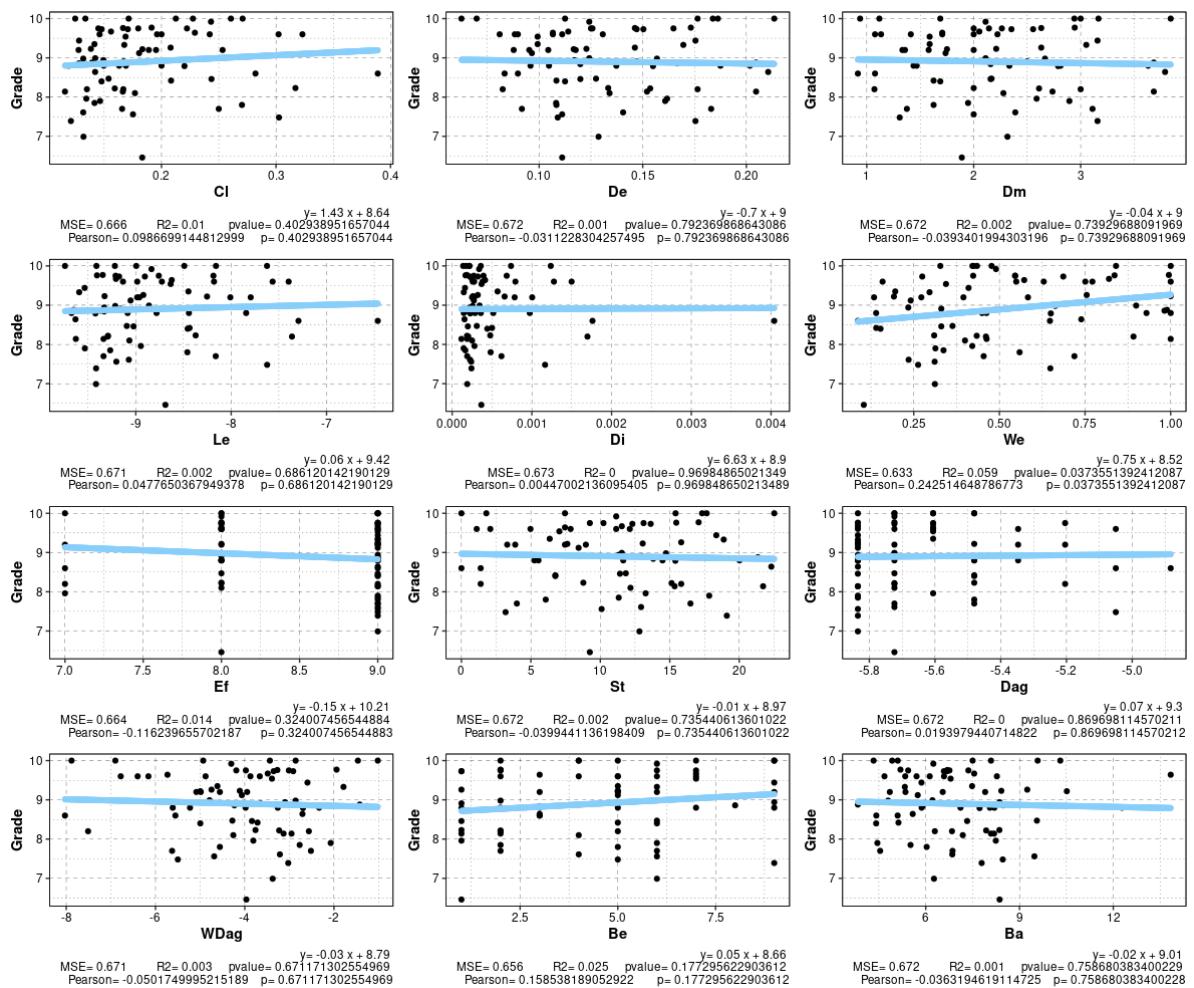


Figura 90: Correlaciones existentes entre las medidas de complejidad de propósito general y la variable *Grade*.

Por último, estudiaremos si existe alguna relación no lineal entre alguna de las medidas topológicas estudiadas y las calificaciones obtenidas por el alumnado. Así pues, en la Figura 91 podemos concluir que podría existir una relación no lineal entre las variables *We* y *Grade* ($p = 0,0774 < 0,1$).

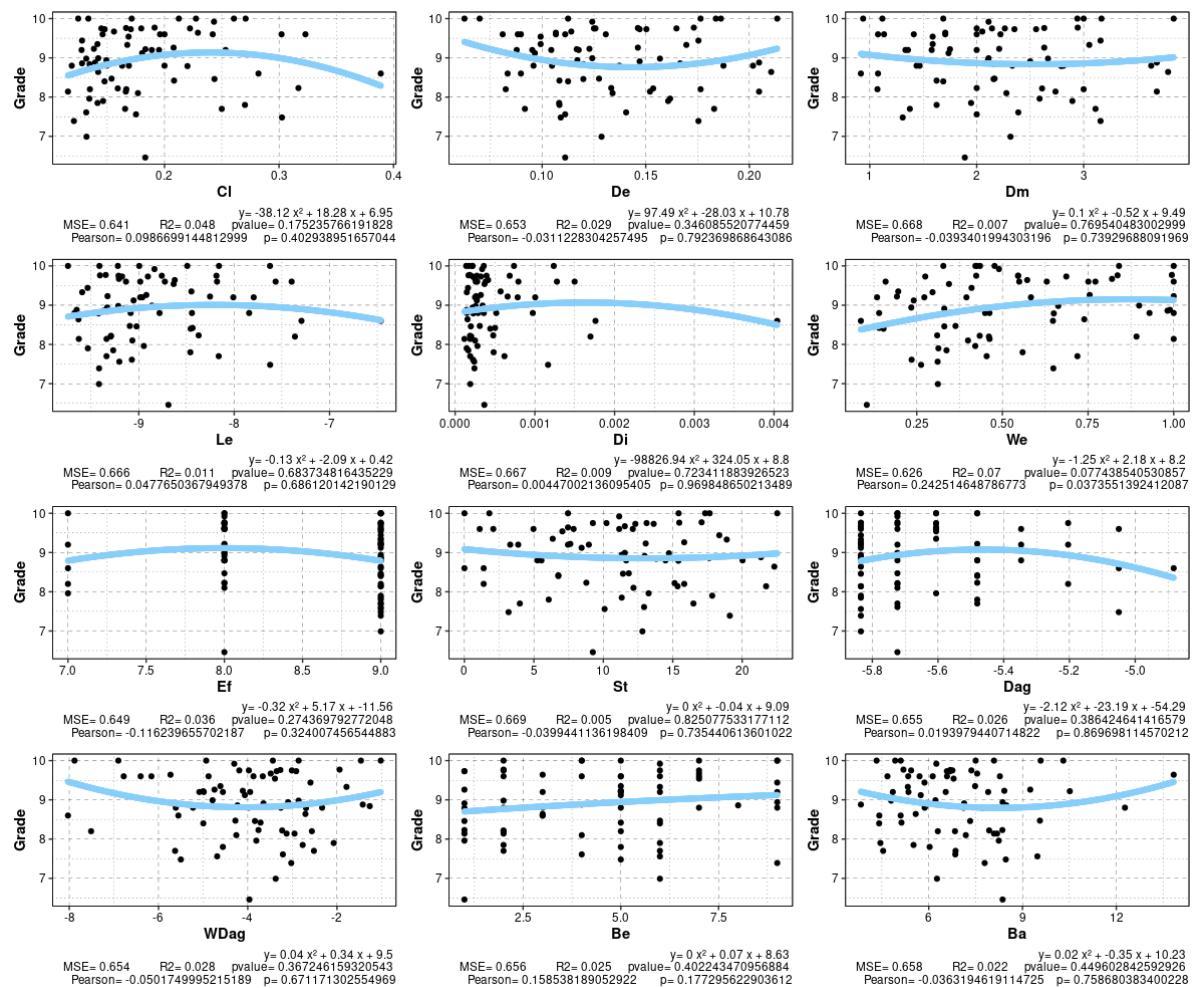


Figura 91: Regresión polinomial entre las medidas de complejidad de propósito general y la variable Grade.

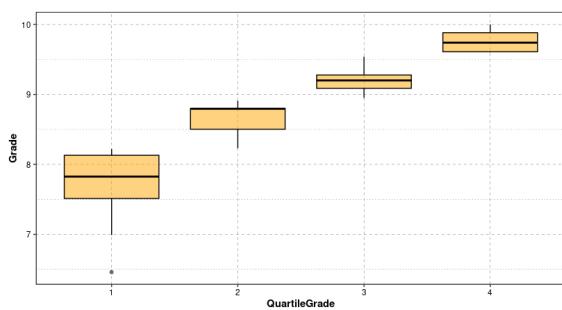
12

PERFILES DE ESTUDIANTES SEGÚN SU RENDIMIENTO

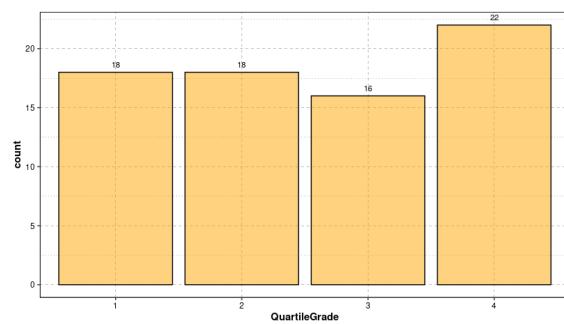
Debido a la dificultad de predecir de manera exacta la calificación obtenida por los alumnos a partir de las medidas de rendimiento descritas anteriormente, se agruparán las notas en clusters significativos y trataremos de predecir en qué cluster se encuentra la nota de un determinado grupo de alumnos.

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 92a, 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 92b el número de grupos que hay en cada cuartil.



(a) Boxplot de las calificaciones por cuartil.



(b) Número de grupos por cuartil.

Figura 92: Resultados obtenidos tras agrupar las calificaciones por cuartiles.

Sin embargo, en la Figura 92a y en la Figura 93, donde se representan cómo de frecuentes son cada una de las calificaciones obtenidas, notamos la presencia de un outlier.

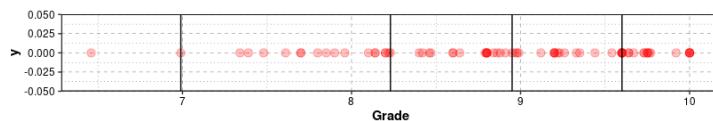


Figura 93: Calificaciones obtenidas por los distintos grupos. El límite de los cuartiles se ha indicado con líneas verticales negras.

En la Figura 94 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.

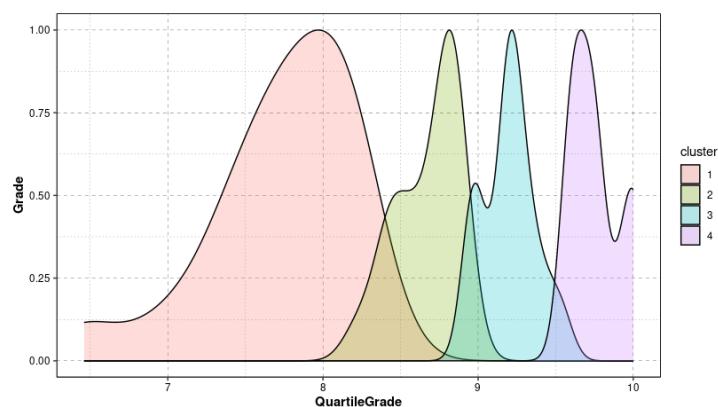


Figura 94: Funciones de densidad de las calificaciones obtenidas por cluster.

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 95 y 96, 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.

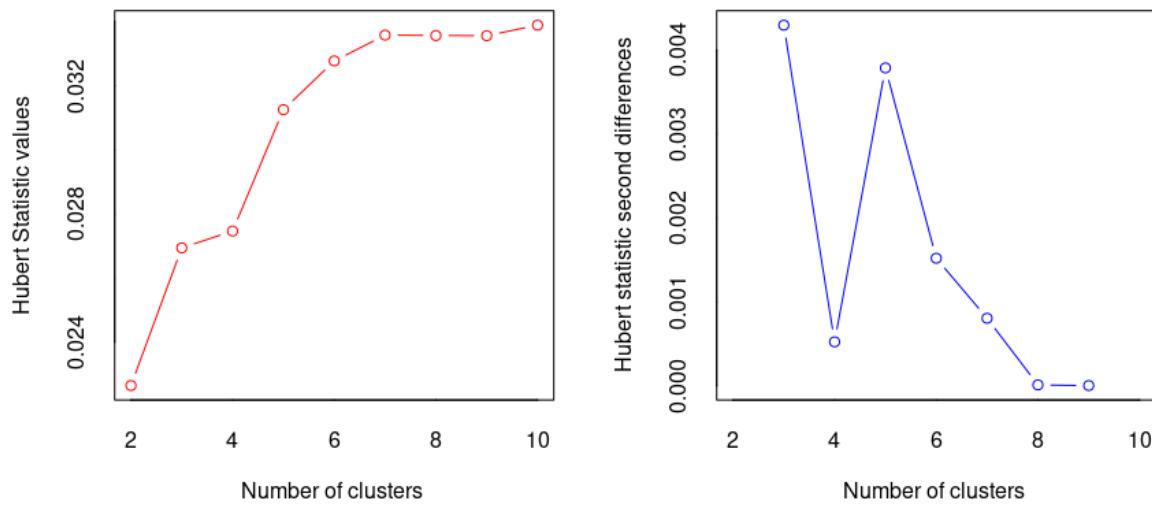


Figura 95: Valores estadísticos de Hubert.

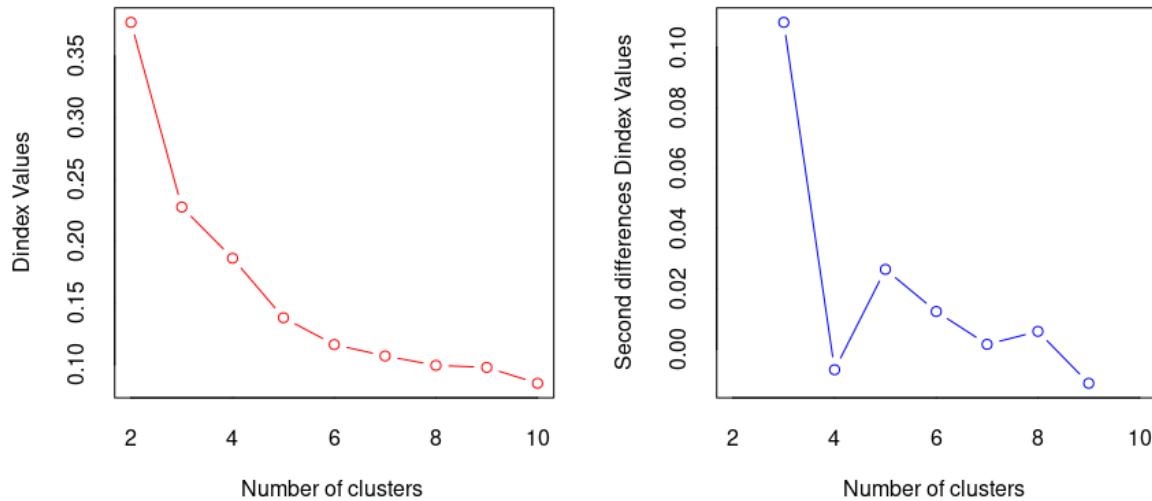
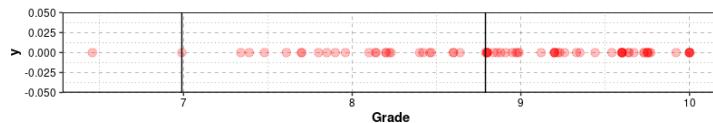
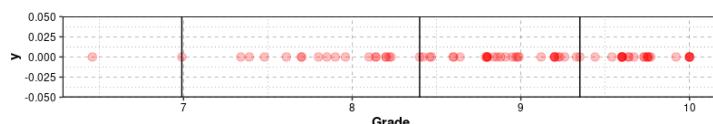
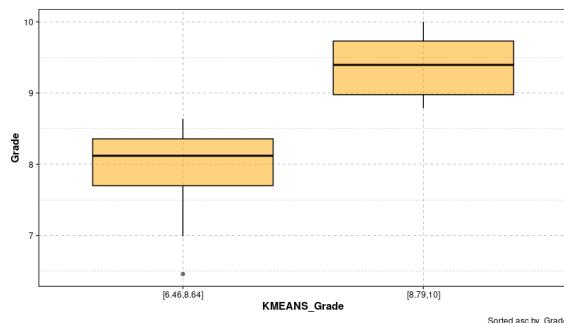


Figura 96: Valores de Dindex.

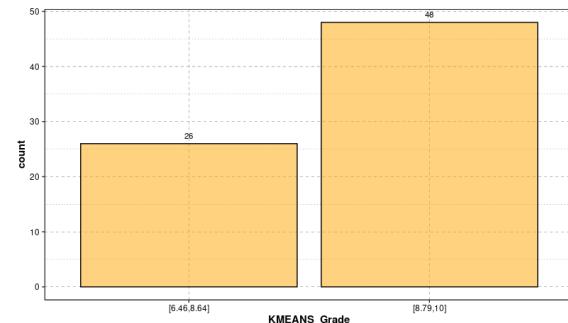
Aplicando el algoritmo de las K-medias para $K = 2$ (Figura 97) y para $K = 3$ (Figura 98), vemos que hay una gran diferencia entre la precisión de uno y otro (para $K = 2$ se tiene accuracy = 0,6978412 mientras que para $K = 3$ se tendrá accuracy = 0,8585982). Además, como puede apreciarse en las Figuras 97 y 98, seguimos teniendo un outlier en ambos casos.

Figura 97: Particiones obtenidas con $K = 2$.Figura 98: Particiones obtenidas con $K = 3$.

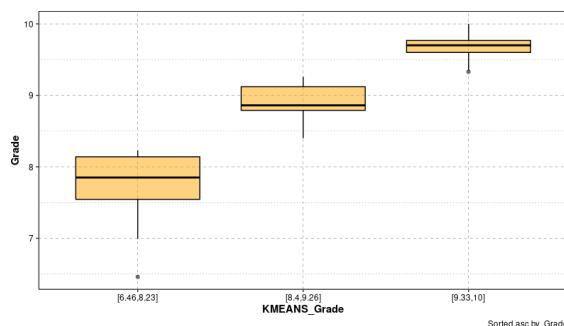
La distribución de la variable *Grade* dentro de cada partición puede verse en las Figuras 99a y 100a mientras que el número de grupos que hay en las particiones puede verse en las Figuras 99b y 100b.



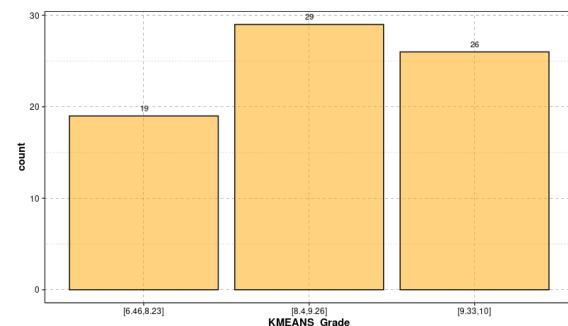
(a) Boxplot de cada una de las particiones.



(b) Número de grupos por partición.

Figura 99: 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 100: Resultados obtenidos tras aplicar el algoritmo de las *K*-Medias con $K = 3$.

Por último, para cinco particiones (Figura 101) se tendrá $\text{accuracy} = 0,9474439$. 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).

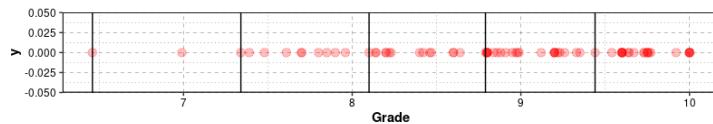
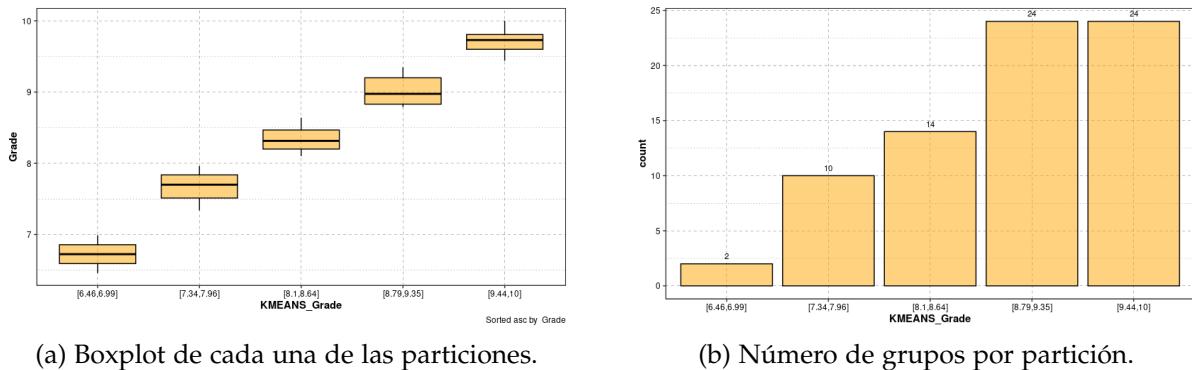


Figura 101: Particiones obtenidas con $K = 5$.



(a) Boxplot de cada una de las particiones.

(b) Número de grupos por partición.

Figura 102: 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 8, nos quedaremos con aquellas que correlan con la variable *Grade* (*np*, *fr*, *ps*, *sq* y *ns*). Así pues, se definirá una nueva métrica, a la que denotaremos por *fm*, como la suma de las medidas de rendimiento *p*, *fr*, *ps*, *sq* y *s*. En la Figura 103 vemos que *fm* no correla con la variable *Grade*.

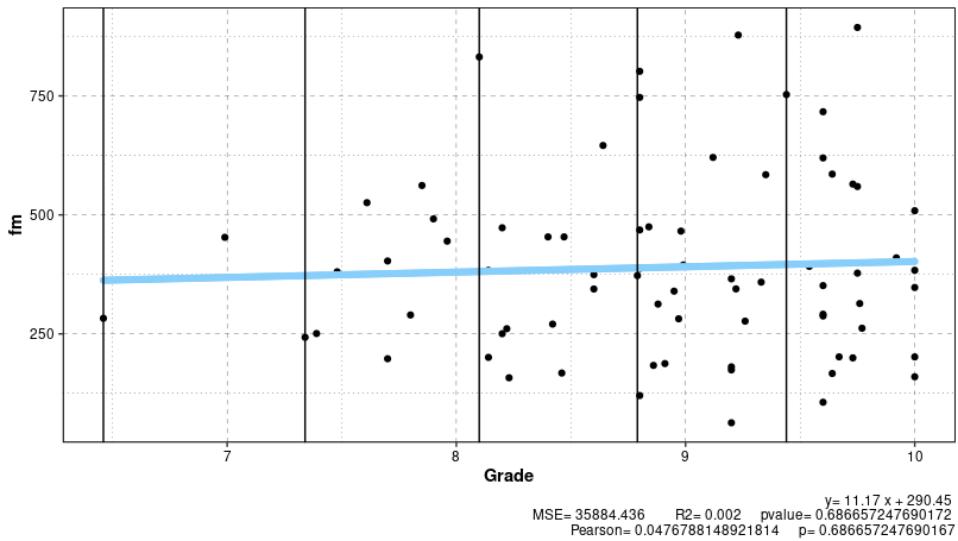


Figura 103: Regresión lineal para aproximar la relación de dependencia entre la variable *fm* y la variable *Grade*.

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 104 y 105, 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.

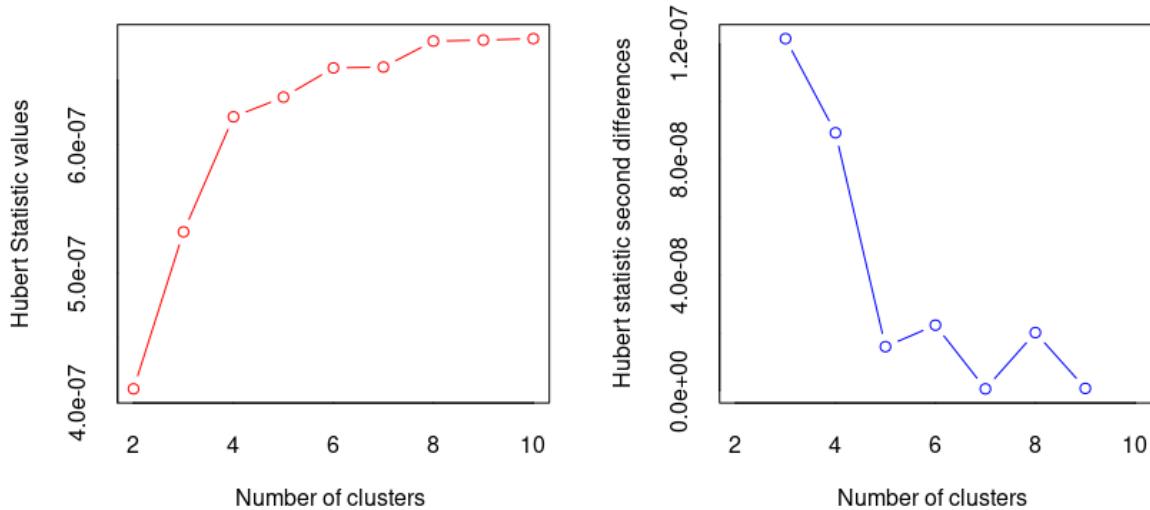


Figura 104: Valores estadísticos de Hubert.

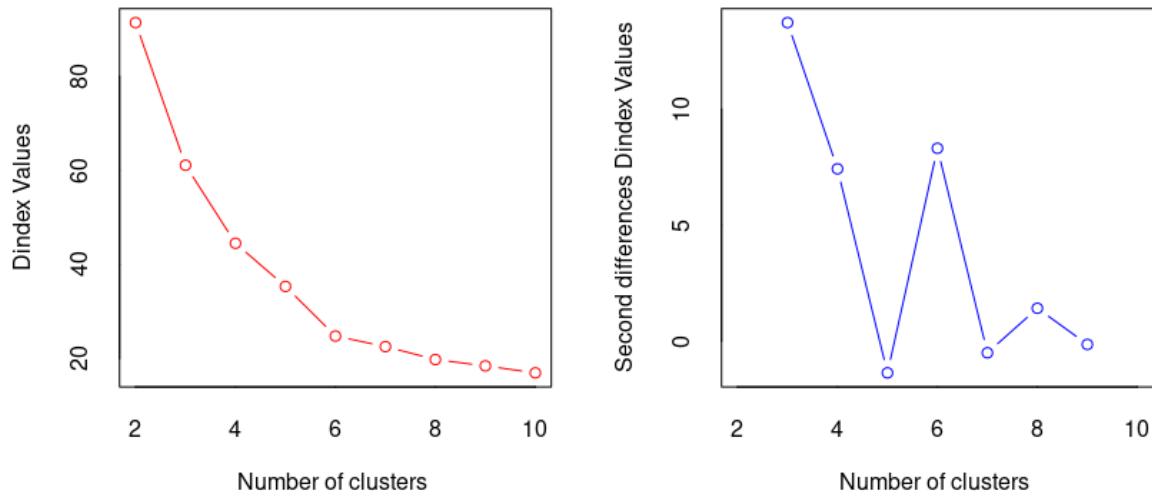


Figura 105: Valores de Dindex.

Aplicando el algoritmo de las K-medias para $K = 4$ (Figura 106) y para $K = 7$ (Figura 107), 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 106 podemos notar la presencia de outliers mientras que en la Figura 107 no.

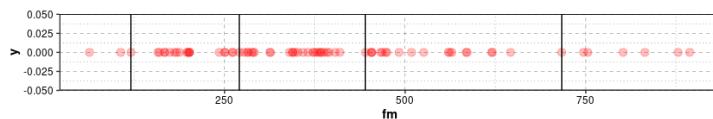


Figura 106: Particiones obtenidas con $K = 4$. Como podemos ver, se observa la presencia de outliers (260,804, 63,173, 106,379 y 261,823).

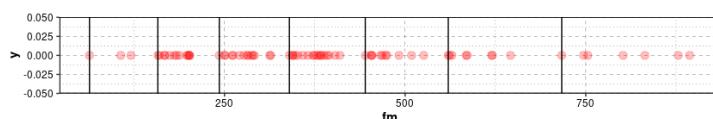
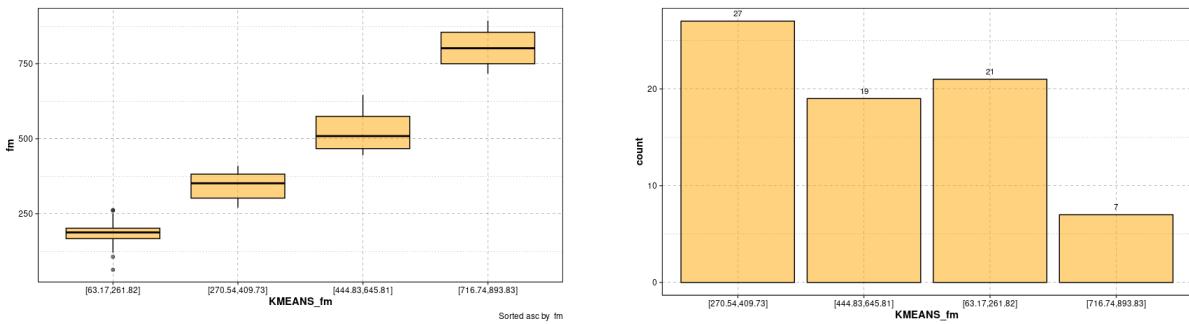


Figura 107: Particiones obtenidas con $K = 7$. No hay ningún outlier.

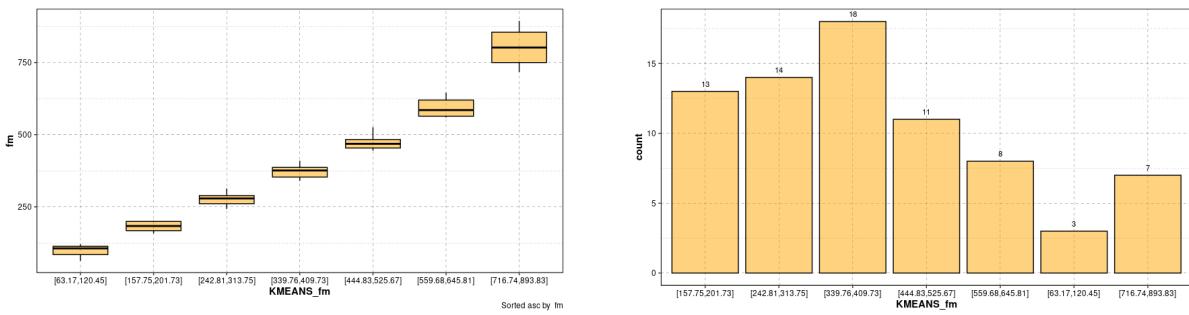
La distribución de la variable *Grade* dentro de cada partición puede verse en las Figuras 108a y 109a mientras que el número de grupos que hay en las particiones puede verse en las Figuras 108b y 109b.



(a) Boxplot de cada una de las particiones.

(b) Número de grupos por partición.

Figura 108: 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 109: Resultados obtenidos tras aplicar el algoritmo de las K -Medias con $K = 7$.

12.4 CLUSTERING MEDIANTE LAS PROPIEDADES ESPECTRALES DE LOS GRAFOS

12.4.1 Clustering mediante el coeficiente LOGLAPo9

Ahora, se ha decidido se asociar los datos usando el algoritmo de las K -medias sobre la variable $LOGLAPo9$. Para decidir el número de clusters en el que agruparemos los datos, se usarán métodos gráficos. Como podemos ver en las Figuras 110 y 111, se ha decidido agrupar los datos en 5 clusters (Figura 112, accuracy = 0,9673813).

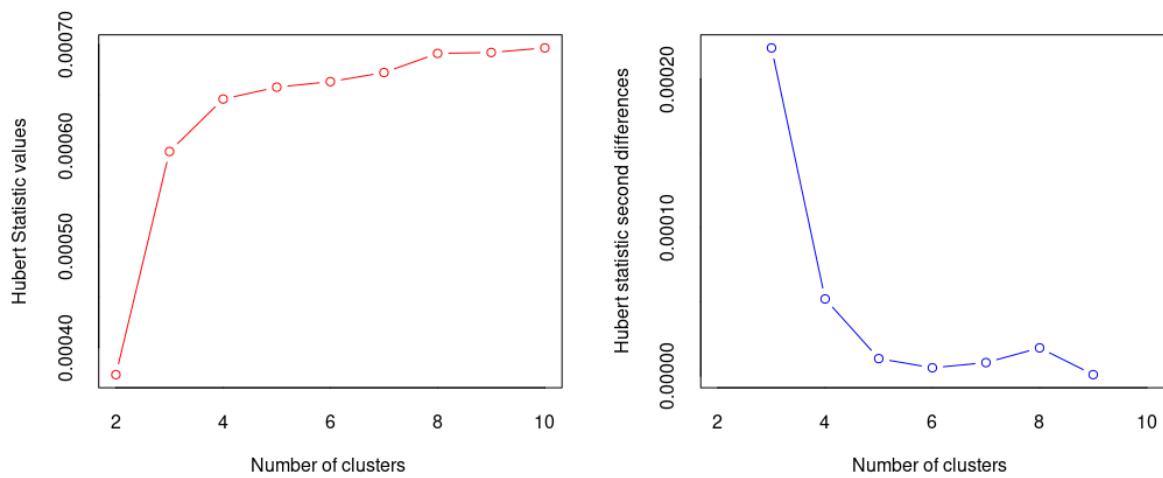


Figura 110: Valores estadísticos de Hubert.

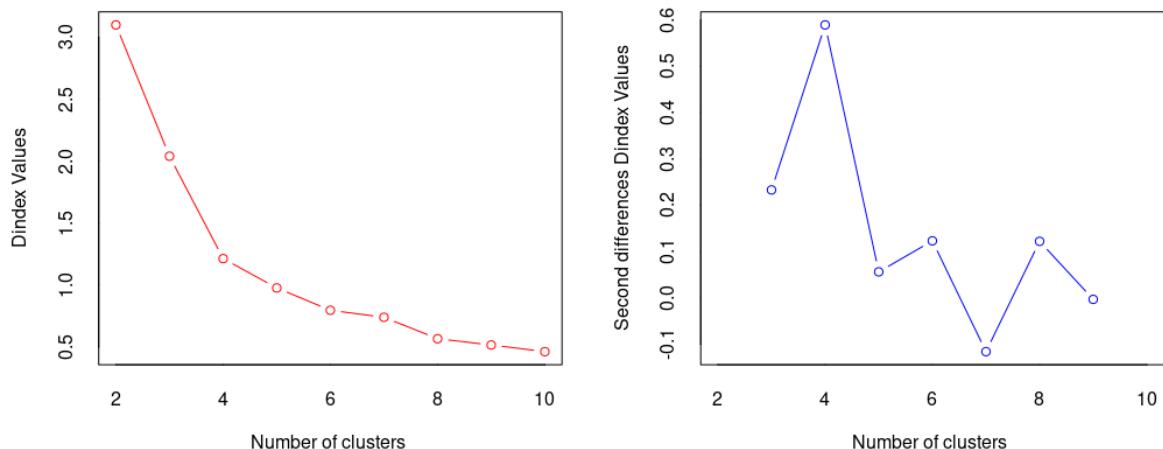
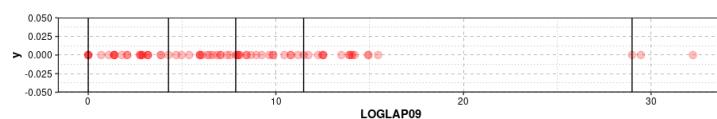
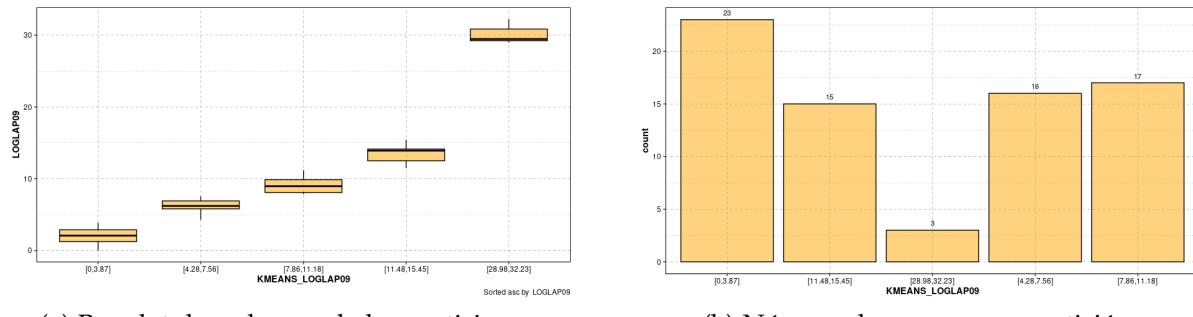


Figura 111: Valores de Dindex.

Figura 112: Particiones obtenidas con $K = 5$. No hay ningún outlier.

La distribución de la variable *Grade* dentro de cada partición puede verse en la Figura 113a mientras que el número de grupos que hay en las particiones puede verse en la Figura 113b.



(a) Boxplot de cada una de las particiones.

(b) Número de grupos por partición.

Figura 113: Resultados obtenidos tras aplicar el algoritmo de las K-Medias con $K = 5$.

12.4.2 Clustering mediante el coeficiente DAG

12.4.3 ¿Quiénes son los grupos en riesgo?

El objetivo principal de este enfoque es tratar de identificar a los estudiantes que tienen dificultades en resolver las tareas propuestas en el laboratorio virtual porque ellos, y no los alumnos ordinarios, requieren la mayor parte de la atención de del profesor para poder alcanzar un 100 % de éxito. Pero, ¿quiénes son?

En principio nos vamos a centrar en los alumnos con el cuartil más bajo de la nota, es decir, aquellos cuya nota está por debajo, como mínimo, del 75 % de las demás calificaciones (Figura 93). Sin embargo, dada la dinámica de las calificaciones cada año (Figura 58), podemos ver que este primer cuartil podría no ser muy preciso. En su lugar, se ha usado la función KMeans en R para dividir el rango de calificaciones en 5 intervalos con una precisión del 95 % en las interdependencias e intradependencias de los 5 clusters. En la partición obtenida mediante KMeans, hay dos grupos cuyas calificaciones son estrictamente inferiores a 8 (Figura 101). Vamos a centrarnos en estos a grupos, que se denominarán “*LOW*”. Por el contrario, el resto de grupos se llamarán “*GOOD*”.

Este conjunto de datos podría procesarse como un único cluster de datos y sólo tendríamos una visión del final del periodo de las prácticas, que podría ser informativo, pero no útil para una intervención temprana. Con el fin de preparar el conjunto de datos para que sea útil cuanto antes y no sólo al final de la práctica, se va a estratificar por niveles donde cada nivel $i \in [3, 9]$ está marcado por la primera vez que el problema (o asignación) i se ha resuelto (Figura 114). Por lo tanto, alcanzar nivel 9 significa que se han completado todas las tareas.

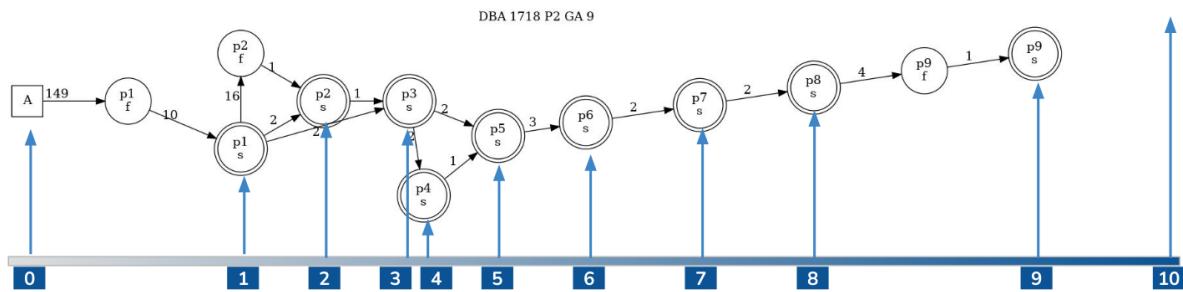


Figura 114: Estratificación del conjunto de datos en 10 episodios consecutivos. Uno para la consecución de cada uno de los 9 problemas, más un último episodio hasta el final de la práctica.

Así pues, tenemos un conjunto de datos muy fiable, reforzado con evidencias de que no hay sesgos a lo largo de los años. También tenemos una partición preliminar sólida y significativa de las calificaciones y un nuevo conjunto de funciones en forma de campana de uso general para identificar estas particiones (Sección 5.2). ¿Qué podría salir mal?

13

CAPACIDAD DE LAS MÉTRICAS

En este capítulo se muestran una serie de estudios estadísticas que muestran la capacidad de todas y cada una de las métricas para diferenciar entre los grupos de prácticas en riesgo y los que no tienen problemas para resolver los problemas propuestos por el profesor.

13.0.1 *Estudio de la capacidad de separación de las medidas clásicas del rendimiento*

13.0.2 *Estudio de la capacidad de separación de las medidas de complejidad de propósito general*

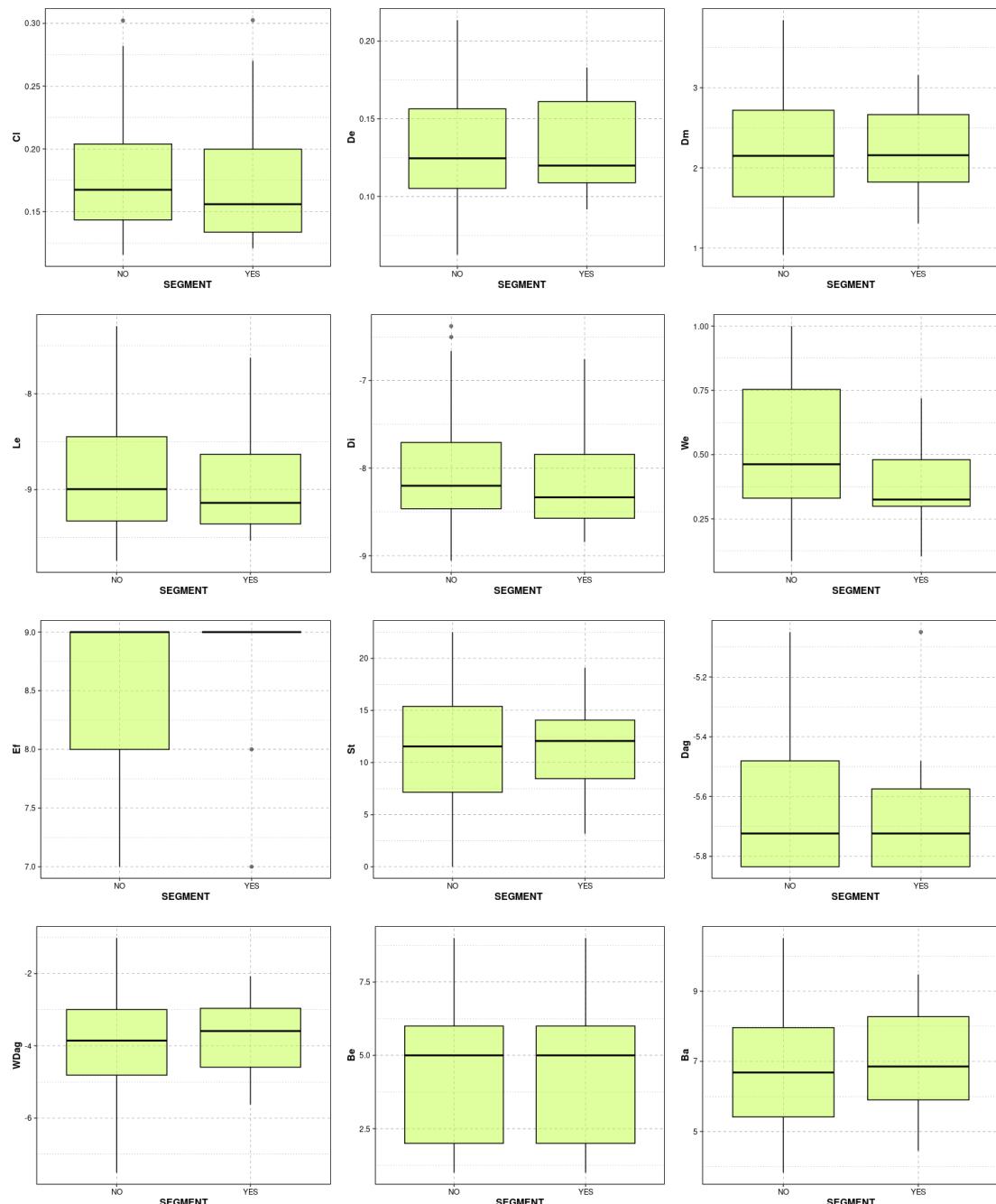


Figura 115: Boxplots de las diferentes medidas de complejidad por grupos de alumnos.

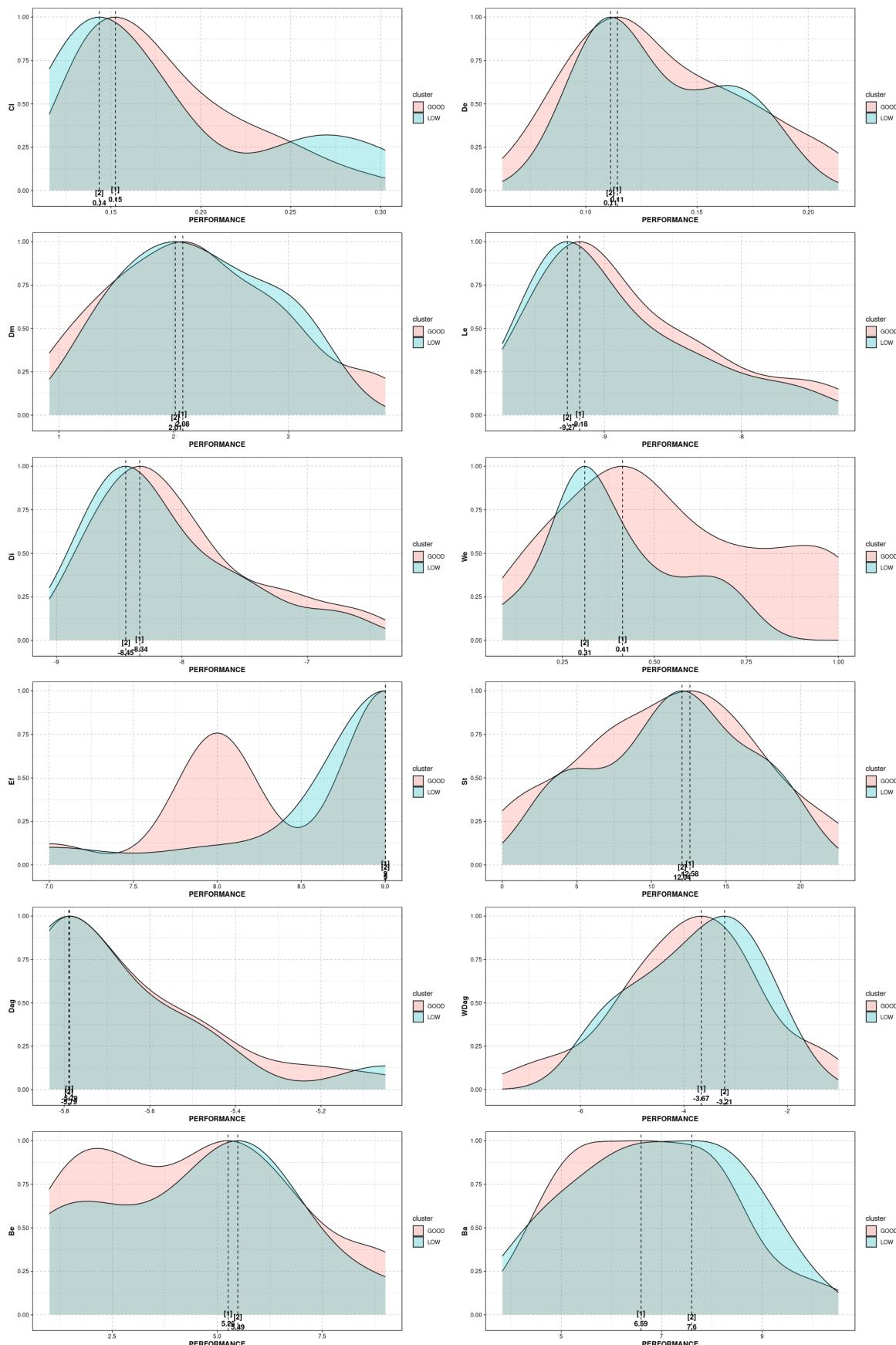


Figura 116: Funciones de densidad de las diferentes medidas de complejidad por grupos de alumnos.

CLASIFICACIÓN DE LOS GRUPOS DE ALUMNOS SEGÚN SU RENDIMIENTO

Para realizar la identificación de los subconjuntos de alumnos definidos en la Sección 12.4.3, se empleará el algoritmo de clasificación de Quinlan C5.0 [Quinlan \(2004\)](#), que no es más que una herramienta de aprendizaje supervisado que genera un árbol de decisión o un conjunto de reglas. Así pues, las distintas categorías en las que se clasificarán a los grupos vendrán dadas en las hojas de dichos árboles y en los consecuentes de las reglas respectivamente. El clasificador estadístico C5.0 se basa en el concepto de entropía, seleccionando primero aquellas características cuyos valores se diferencian más entre diferentes categorías de grupos de alumnos.

Se utilizarán como variables de entrada del clasificador tanto las medidas clásicas de rendimiento presentadas en el Capítulo 8 como las medidas grafo-teóricas (Sección 5.2). Por su parte, como categorías de salida, tendremos tanto una única partición en alumnos “LOW”, que se supone que requieren más esfuerzo para superar el laboratorio, como las cinco particiones que se muestran en la Figura 101.

Dado que hay muy pocos registros (77 en cada nivel), el problema de clasificación va a ser difícil, y, con el fin de aumentar las evidencias requeridas por C5.0 para trabajar, el conjunto de datos se agrupará cada tres niveles consecutivos. Así pues, por un lado, agruparemos los niveles 3, 4 y 5 para realizar predicciones al principio de la práctica y, por otro lado, agruparemos los niveles 8, 9 y 10. Así pues, las siguientes secciones se dividirán en dos subapartados, dependiendo de la parte del dataset que estemos utilizando para entrenar el clasificador.

14.1 CLASIFICACIÓN EMPLEANDO LAS MÉTRICAS CLÁSICAS DE RENDIMIENTO

14.2 CLASIFICACIÓN EMPLEANDO LAS MEDIDAS DE COMPLEJIDAD DE PROPÓSITO GENERAL

14.2.1 Clasificación empleando los niveles 3, 4 y 5

El resultado es que C5.0 obtiene un conjunto de reglas capaz de clasificar exactamente los 35 de los 36 casos de grupos “*LOW*” del nivel 3 en adelante, es decir, a partir de un tercio del periodo de tiempo dedicado a la práctica, con un *p* value $p = 3,843e - 16$ estadísticamente muy relevante (Figura 117). Además, las reglas obtenidas (Figura 14.2.1) tienen pleno sentido. Por ejemplo, la regla 7 dice: “Un grupo está en riesgo si está centrado en los problemas con numeración alta ($Be > 6$)”. O la regla 3 que dice “Un grupo está en riesgo si no se ha centrado por igual en todos los problemas ($Ba > 3.366005$) y recorre caminos menos largos en media ($WDag > -4.317488$) teniendo en cuenta que el valor de máxima probabilidad de la métrica $WDag$ es $-4,19$.

		Referencia	
		GOOD	LOW
Predicción	GOOD	186	1
	LOW	0	35

Accuracy : 0.9955
 95 % CI : (0.9752, 0.9999)
 No Inform. Rate : 0.8378
 P-Value : 3.843e-16

Figura 117: Aprendiendo a identificar a los grupos “*LOW*”, es decir, aquellos con una puntuación inferior a una nota 8,1 sobre 10.

Reglas de clasificación para identificar grupos de tipo “LOW”.

Rule 11/1: (17.8, lift 1.5)

Ba > 9.26712
-> class GOOD [0.949]

Rule 11/6: (9.4/1.1, lift 2.3)

De > 0.1153846

Ef > 3

Rule 11/2: (200.7/68.3, lift 1.0)

De > 0.0952381
-> class GOOD [0.658]

Ef <= 4

WDag <= -6.612041

Ba > 3.366005

-> class LOW [0.818]

Rule 11/3: (8.5/0.3, lift 2.5)

WDag > -4.317488
Ba > 3.366005
-> class LOW [0.880]

Rule 11/7: (3.4, lift 2.3)

Be > 6
-> class LOW [0.814]

Rule 11/4: (5, lift 2.4)

Dm <= 1.2
Ef <= 4
St > 1.791759
-> class LOW [0.858]

Rule 11/8: (13.3/2.3, lift 2.2)

De <= 0.0952381
Ba <= 9.26712
-> class LOW [0.783]

Rule 11/5: (8.3/0.6, lift 2.4)

De <= 0.0989011
Ef <= 4
Ba <= 9.26712
-> class LOW [0.850]

Rule 11/9: (19.9/5.4, lift 2.0)

Dm <= 1.222222
Ef <= 3
St > 0.6931472
Ba <= 9.26712
-> class LOW [0.707]

Pero no sólo esto, ¿podríamos dar un gran paso y predecir no sólo equipos “LOW”/“GOOD”, sino también un límite inferior y superior de sus calificaciones finales, y adivinar en cuál de los 5 intervalos de la Figura 101 se encuentran? La respuesta es, felizmente sí.

14.2.2 Clasificación empleando los niveles 8, 9 y 10

14.3 CLASIFICACIÓN EMPLEANDO UNA COMBINACIÓN DE TODAS LAS MÉTRICAS

A continuación, se repetirán los experimentos anteriores, pero ahora empleando las funciones s , p , Cl , De , Dm , Le , Di , We , Ef , St , Dag , $WDag$, Be y Ba .

14.3.1 Clasificación empleando los niveles 3, 4 y 5

El resultado es que C5.0 obtiene un conjunto de reglas capaz de clasificar exactamente los 36 casos de grupos “LOW” del nivel 3 en adelante, es decir, a partir de un tercio del periodo de tiempo dedicado a la práctica, con un p value $p < 2,2e - 16$ estadísticamente muy relevante (Figura 118). Entre las reglas obtenidas (Figura 14.3.1), la regla 6 dice: “Un grupo está en riesgo si ha realizado pocas sesiones ($s \leq 21$) y el grafo no está lo suficientemente balanceado ($Ba > 3,366005$)”. O la regla 7 que dice “Un grupo está en riesgo si tiene pocas sesiones $We \leq 0.01620162$ y no se centra por igual en todos los problemas $Ba > 3.366005$ ”.

		Referencia	
		GOOD	LOW
Predicción	GOOD	186	0
	LOW	0	36

Accuracy : 1
 95 % CI : (0.9835, 1)
 No Inform. Rate : 0.8378
 P-Value : < 2.2e-16

Figura 118: Aprendiendo a identificar a los grupos “LOW”, es decir, aquellos con una puntuación inferior a una nota 8,1 sobre 10.

Reglas de clasificación para identificar grupos de tipo "LOW".

Rule 5/1: (20.9, lift 1.5)	We <= 0.01620162
s > 177	Ba > 3.366005
Ba <= 5.651057	-> class LOW [0.918]
-> class GOOD [0.956]	
	Rule 5/8: (19/1.4, lift 2.5)
Rule 5/2: (21.2/0.3, lift 1.5)	s > 177
s > 21	s <= 271
Cl > 0.6111111	Cl <= 0.4866667
WDag > -6.562621	WDag > -6.39693
-> class GOOD [0.945]	Ba > 5.651057
	-> class LOW [0.885]
Rule 5/3: (11, lift 1.4)	
Ba <= 3.366005	Rule 5/9: (9.8/0.6, lift 2.4)
-> class GOOD [0.923]	s <= 105
	Cl <= 0.6111111
Rule 5/4: (41.2/4.6, lift 1.4)	We > 0.1657658
WDag <= -6.678342	-> class LOW [0.868]
-> class GOOD [0.871]	
	Rule 5/10: (19.6/4.3, lift 2.1)
Rule 5/5: (143.1/44.1, lift 1.1)	WDag > -6.678342
Cl <= 0.4866667	WDag <= -6.562621
-> class GOOD [0.689]	-> class LOW [0.755]
Rule 5/6: (12.4, lift 2.6)	Rule 5/11: (26.4/7.8, lift 1.9)
s <= 21	Cl > 0.4866667
Ba > 3.366005	Cl <= 0.6111111
-> class LOW [0.931]	WDag > -6.678342
	Ba > 3.366005
Rule 5/7: (10.3, lift 2.6)	-> class LOW [0.690]

14.3.2 Clasificación empleando los niveles 8, 9 y 10

Parte VI

CONCLUSIONES Y VÍAS FUTURAS

Logros y nuevas vías de investigación.

15

CONCLUSIONES

Este estudio ha demostrado varias hipótesis. La más importante es que los grupos en riesgo pueden detectarse desde el primer tercio del periodo de laboratorio con una gran significancia. La segunda es que esto puede hacerse simplemente observando su comportamiento desde un punto de vista topológico, algo también intuitivo porque todos sabemos lo que hacemos cuando estamos perdidos: deambular por transiciones sin sentido. Y, en tercer lugar, la detección temprana podría deberse probablemente al hecho de que las primeras etapas son cruciales para el resto, y los errores y despistes más importantes se producen muy pronto, justo al principio.

En este trabajo se ha conseguido realizar una predicción de los grupos en riesgo de obtener un peor rendimiento así como predecir en qué intervalo de notas está la calificación de cada uno de los grupos con evidencias (análisis, descarte de outliers, correlaciones).

16

VÍAS FUTURAS

Son múltiples las diversas propuestas que surgen a partir del trabajo aquí realizado. En primer lugar, podríamos tratar de predecir las calificaciones del alumnado de una manera más precisa, reduciendo los tamaños de los intervalos que el clasificador entrenado asignará a cada uno de los grupos. Esto podría hacerse simplemente aumentando el número de clusters en el que se subdividen los grupos de prácticas, aunque podría acarrear una pérdida de precisión por parte del clasificador al reducirse las diferencias entre clusters.

Por otra parte, se podría realizar un análisis del comportamiento de un mismo grupo a lo largo del tiempo con el fin de determinar si el comportamiento del mismo varía durante el desarrollo de las prácticas.

Por último, nótese que en este estudio se ha caracterizado el comportamiento de aquellos grupos con dificultades para resolver los problemas que se le proponen pero, ¿podemos ayudarles a recuperar un buen ritmo de aprendizaje y un progreso adecuado a lo que el profesorado de la asignatura espera? Así pues, otra vía de investigación sería proponer una serie de mejoras a los grupos que actualmente se encuentran en riesgo y ver si las mismas consiguen cambiar su comportamiento y les ayudan a progresar adecuadamente.

A

TABLAS DESCRIPTIVAS

Cuadro 24: 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 25: 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 26: Número y tipo de las sesiones de trabajo.

BIBLIOGRAFÍA

- Armenta, R. E. (2015). La fórmula de cayley. https://libroweb.alfaomega.com.mx/book/682/free/ovas_statics/cap11/lecturas/La_formula_de_Cayley.pdf.
- Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(1), e1230.
- Bogarín Vega, A., Cerezo Menéndez, R., Romero, C., et al. (2018). Discovering learning processes using inductive miner: A case study with learning management systems (lmss). *Psi-cothema*.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). *Introduction to Algorithms*. MIT Press.
- dos Santos Garcia, C., Meincheim, A., Junior, E. R. F., Dallagassa, M. R., Sato, D. M. V., Carvalho, D. R., Santos, E. A. P., & Scalabrin, E. E. (2019). Process mining techniques and applications—a systematic mapping study. *Expert Systems with Applications*, 133, 260–295.
- Elmoazen, R., Saqr, M., Khalil, M., & Wasson, B. (2023). Learning analytics in virtual laboratories: a systematic literature review of empirical research. *Smart Learning Environments*, 10(1), 1–20.
- Günther, C. W., & Aalst, W. M. V. D. (2007). *Fuzzy Mining—Adaptive Process Simplification Based on Multi-Perspective Metrics*, vol. 5. Springer.
- Günther, C. W., & Rozinat, A. (2012). Disco: Discover your processes. *BPM (Demos)*, 940(1), 40–44.
- Juhaňák, L., Zounek, J., & Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior*, 92, 496–506.
- Keller, F. S. (1968). Good-bye, teacher... *Journal of applied behavior analysis*, 1(1), 79.
- Mayorga, H. S. A., & García, N. R. (2015). Process mining: development, applications and critical factors. *Cuadernos de Administración*, 28(50), 137–157.

- Mondal, B., & De, P. D. K. (2017). An overview applications of graph theory in real field. *International Journal of Scientific Research in Computer Science, Engineering, and Information Technology*, 2(5), 751–759. Assistant Teacher, Bajitpur Primary School, Bajitpur, Purapara, Murshidabad, West Bengal, India. Professor of Mathematics, School of Science, Netaji Subhas Open University, DD-26, Sector-I, Salt Lake, Kolkata, West Bengal, India.
- Mukala, P., Buijs, J., Leemans, M., & van der Aalst, W. (2015). Learning analytics on coursera event data: A process mining approach. In *5th International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2015)*, (pp. 18–32). CEUR-WS. org.
- Quinlan, J. R. (2004). Data mining tools see5 and c5.0. <http://www.rulequest.com/see5-info.html>.
- Sedrakyan, G., De Weerdt, J., & Snoeck, M. (2016). Process-mining enabled feedback: “tell me what i did wrong” vs. “tell me how to do it right”. *Computers in human behavior*, 57, 352–376.
- van der Aalst, W. M. (2016). *Process mining: data science in action*. Springer.
- Vidal, L. C. (2016). A virtual laboratory for multiagent systems: Joining efficacy, learning analytics and student satisfaction. *International Symposium on Computers in Education (SIIE)*. IEEE, (pp. 1–6).
- Vidal, L. C., & Sánchez-Carrascosa, F. (2022). Larva: Learning analytics recollection and visualization agents. *2022 International Symposium on Computers in Education (SIIE)*. IEEE, (pp. 1–6).
- Vidal, L. C., & Sánchez-Carrascosa, F. (2023). Larva, a multiagent system to collect learning analytics and provide immediate feedback to students during laboratory assignments. *IE Comunicaciones. Revista Iberoamericana de Informática Educativa*, 37, 33–43.
- Vidal, L. C., Sánchez-Carrascosa, F., Martínez, M. I. R., & Olivares, J. F. (2023). In heaven as on earth: The performance of students is as good as it is the digraph that describes their behavior. *Submitted to the XXV International Symposium on Computers in Education (SIIE)*.
- Williamson, D. P. (2019). Bridging continuous and discrete optimization. Sep 30, 2019. Lecture 8.