Experiencias Docentes

Enseñanza del Software Estadístico R a alumnos de Matemáticas.

Teaching R to Mathematical Students

Elena Castilla y Pedro J. Chocano

Revista de Investigación

Volumen XI, Número 1, pp. 057–068, ISSN 2174-0410
Recepción: 28 Sep'20; Aceptación: 21 Oct'20

1 de abril de 2021

Resumen

Habiendo constatado la falta de formación en herramientas estadísticas en algunas titulaciones de matemáticas, así como la creciente importancia del software de programación R, se ha visto la necesidad de impartir un curso de Análisis de Datos con R. Éste se ha desarrollado dentro de la iniciativa "Compumates" de la Facultad de CC. Matemáticas, Universidad Complutense de Madrid, y ha constado de diferentes sesiones, que parten desde la instalación del software y uso de comandos básicos, hasta su aplicación en técnicas de análisis y predicción. Para ello se ha provisto a los alumnos de un manual de aprendizaje. En este trabajo se muestran los resultados que ha tenido la experiencia en el aprendizaje del alumno y la valoración que tiene éste sobre la misma.

Palabras Clave: Aprendizaje, software estadístico, R.

Abstract

The lack of statistical tools in some bachelor's degrees of Mathematics jointly with an increasing importance of the software R leads to the need of giving a course of Data Analysis with R. This course has been developed under the iniciative "Compumates" of the Faculty of Mathematics at Complutense University of Madrid. It consists on different sessions that go from the basic commands to their application to techniques of analysis and prediction. In this paper, it is shown the results obtained by the students and their opinion about the course.

Keywords: Learning, statistical software, R .

1. Contextualización del curso y objetivos

En el curso 2019-2020 se llevó a cabo la II edición de la iniciativa "Compumates". "Compumates" engloba una serie de cursos organizados por la Facultad de CC. Matemáticas de la Universidad Complutense de Madrid (UCM) sobre distintos recursos matemáticos tales como Python o LaTeX, entre otros. Estos cursos están dirigidos a estudiantes de cualquier grado, máster o doctorado, así como a cualquier miembro del personal de la Facultad (PDI, PAS) o persona interesada. La inscripción es totalmente gratuita. En este contexto, en la primera quincena de noviembre de 2019, se desarrolló el curso "Análisis de Datos con R". Nótese que dicho curso surge como una evolución natural del curso "Introducción a R" llevado a cabo en la I edición por los mismos profesores. En el curso desarrollado durante la primera edición se enseñaban los comandos más básicos de R y se daban unas pequeñas pinceladas de su posible aplicación a la estadística. Si bien este primer curso recibió muy buena acogida por los estudiantes, la mayoría de éstos comentaron la necesidad de expandir su duración y temática, centrándose en temas como predicción o representación gráfica y análisis de datos multivariantes. Una de las sugerencias más solicitadas era la de poder disfrutar de un manual dinámico y sencillo para seguir el curso con mayor comodidad y tener más ejemplos al alcance.

R es un lenguaje de programación orientado a la estadística, cuyo uso ha ido ganando protagonismo en los últimos años. Se trata de un software libre en el que los usuarios pueden publicar sus propios paquetes o modificar a su antojo diversos aspectos. Esto último dota a R de gran versatilidad. Por otro lado, la importancia de R en los ámbitos laborales, de investigación y de docencia en nuestros días es un hecho innegable. Así, en 2018, R fue el segundo software estadístico con más artículos en google scholar [8]. En el ámbito laboral R no se queda atrás. En en el año 2015, Rexer Analytics llevó a cabo una encuesta a 1,220 científicos de datos, preguntándoles sobre ciertos aspectos de su trabajo. A la pregunta de qué herramientas usaban para su trabajo (en la que podían seleccionar tantas respuestas como quisieran), R fue la más seleccionada tanto como herramienta primaria como secundaria. En particular, 76 % de los encuestados dijeron usar R y más de un tercio (36 %) identificaron R como herramienta principal [7]. Por otra parte, RStudio es una de las interfaces más utilizadas con R, por su sencillez de uso y su estética (véase Figura 1). Por tanto, la importancia de R, y en particular de su interfaz RStudio, no debe ser ignorada por las universidades.

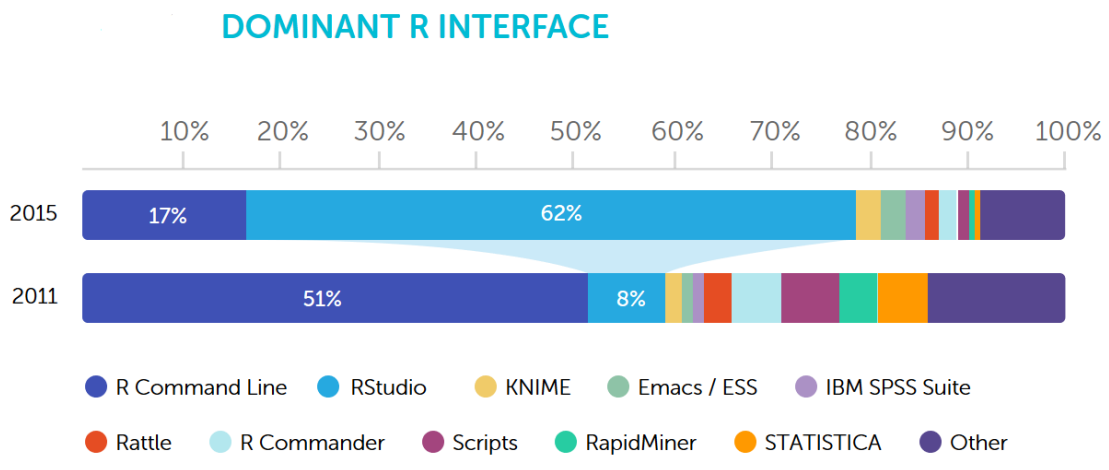

Figura 1. Interfaces de R, comparativa de uso en 2011 y 2015. Imagen obtenida de [7].

Obsérvese que en la Facultad de CC. Matemáticas de la UCM se imparten actualmente 3 grados y 3 dobles grados, así como 4 másteres y 2 programas de doctorado de matemáticas y/o estadística. Los grados (*Matemáticas*, *Matemáticas y Estadística* y *Ingeniería Matemática*) constan

de 4 cursos académicos, mientras que los dobles grados (*Matemáticas y Física*, *Matemáticas e Informática* y *Matemáticas, Estadística y Economía*) constan de 5 cursos académicos. El software R sólo se imparte en los grados de *Matemáticas y Estadística* y el doble grado de *Matemáticas Estadística y Economía*, así como en los másteres aplicados (*Máster en Estadísticas Oficiales*, *Máster en Tratamiento Estadístico Computacional de la Información* y *Máster en Ingeniería Matemática*). Entendemos, por tanto, necesario el ofrecer esta formación al resto de alumnos de grado y máster, así como reforzar la de los alumnos que lo hayan estudiado previamente. Es importante observar que en el sistema educativo español el acceso a la universidad es a partir de los 18 años.

Por otra parte, la Facultad de CC. Matemáticas de la UCM ofrece a los alumnos de cualquier Grado y Doble Grado de la Facultad, además de alumnos de los Másteres en *Ingeniería Matemática* y *Estadísticas Oficiales e Indicadores Sociales y Económicos*, la posibilidad de realizar prácticas curriculares. El único requisito es, para los alumnos de grado o doble grado, el haber cursado al menos la mitad de su plan de estudios. Es decir, este tipo de actividad está dirigida para alumnos de últimos cursos y posgrado. Si bien este tipo de prácticas se consideran de aprendizaje, y las propias empresas contratantes son conscientes de ello, es muy habitual que soliciten conocimiento básico de algún software estadístico. La posibilidad de acreditar un curso de análisis de datos con R sirve de ayuda a aquellos alumnos que quieran realizar prácticas externas y, ya de manera general, a todos ellos que al terminar sus estudios acceden al mundo laboral.

2. Aplicación Práctica

2.1. Recursos Materiales

El curso se realizó en un aula con acceso libre a internet y enchufes para que los alumnos, que tenían que traer su propio ordenador personal, pudieran conectarse. En caso de que algún alumno no dispusiera de ordenador personal, se permitía el trabajo por parejas. Los profesores disponían de una pantalla digital en el aula con la que poder compartir la pantalla de su ordenador, así como una pizarra complementaria. Para seguir el curso se proporciona a los alumnos un manual, que se actualiza constantemente, realizado por los propios profesores. Dicho manual puede encontrarse en la página web de ambos profesores [3].

2.2. Participantes

Atendiendo a el orden de inscripción, se aceptó la matriculación de un total de 32 alumnos, de los cuáles todos tenían una relación directa con la Facultad, siendo 28 alumnos de grado o doble grado, y el resto alumnos de posgrado (máster o doctorado). La edad de los estudiantes se sitúa entre los 18 y los 26 años.

En la Figura 2 podemos ver la distribución de los alumnos por titulación y por curso académico, respectivamente. La mayoría de los alumnos pertenece a alguno de los grados o dobles grados, principalmente de cursos intermedios o finales. Observamos que la mitad de los alumnos provienen del grado en *Matemáticas*, mientras que sólo un 6 % de ellos proviene del citado grado en *Matemáticas y Estadística*. Parece claro que hay un gran interés por parte de aquellos alumnos que no tienen la posibilidad de estudiar R en su titulación, lo cuál refuerza uno de los objetivos planteados con anterioridad. Además, que la mayoría de alumnos sean de cursos elevados puede estar ligado al interés de una futura incorporación al mundo laboral, también tratado en la sección previa.

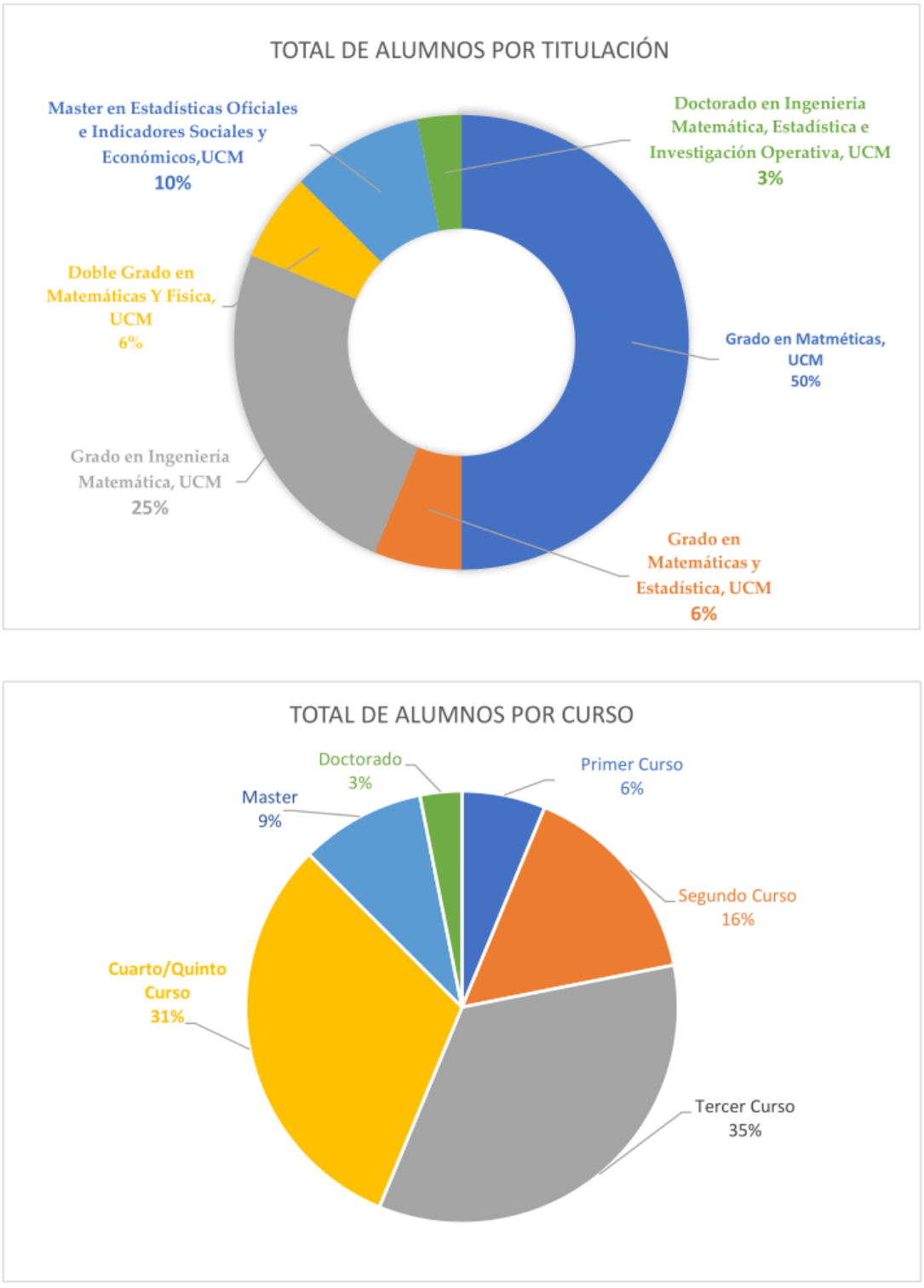

Figura 2. Distribución de alumnos por titulación (arriba) y distribución de alumnos por curso acdémico (abajo).

2.3. Metodología

Se realizaron 6 sesiones de 2 horas de duración cada una. Para el desarrollo de las clases se tuvo en cuenta la formación previa de los estudiantes, la cuál se pidió en el momento de inscripción mediante un formulario. Al tratarse de alumnos de la propia facultad y, por tanto, con buena base matemática (muchos de ellos, además, de cursos superiores), se optó por desarrollar el contenido de manera más profunda explicando con más detalle las ideas intuitivas de los conceptos matemáticos que se aplicaban sin obviar la parte más práctica y aplicada.

Sesión 1: Introducción, Instalación de R, RStudio y Conceptos Básicos

Esta sesión inicial comienza con una pequeña introducción motivadora a R y su importancia en los últimos años. Se enseña a los alumnos el proceso de descarga e instalación del Software. Es importante aquí entender que no todos los alumnos disponen del mismo sistema operativo, y que R dispone de un procedimiento diferente en cada situación. En cualquier caso, toda la información puede encontrarse en la propia página web oficial de R (https://cran.r-project.org/index.html). Para facilitar el uso del Software se aconseja y ayuda a los alumnos a que instalen la interfaz RStudio a pesar de no ser algo estrictamente necesario para poder seguir el curso.

Aunque se trate de un software libre y su instalación no entrañe demasiados problemas, este proceso dura gran parte de la primera sesión. El resto de la primera sesión se emplea para introducir los conceptos más sencillos de R como pueden ser operadores básicos, definición de vectores, matrices, etc. Es importante que el alumno pierda el miedo en esta primera toma de contacto con el Software y vea que R es un entorno de programación intuitivo y sencillo de manejar.

Sesión 2: Condicionales y Operadores Lógicos, Bucles y Funciones

Esta sesión es de un contenido matemático más puro, en el que se enseña al alumno a usar condicionales, operadores lógicos y bucles y a incluirlos dentro de sencillas funciones. Si bien es un poco más difícil de seguir para los estudiantes del máster de *Estadísticas Oficiales* (con una formación generalmente no tan matemática), y es cierto que existen gran cantidad de librerías y paquetes de R, es importante que sean capaces de crear sus propias funciones.

La sesión consta de una primera hora en la que se enseñará el manejo de R para estas cuestiones, seguida de otra hora de trabajo propio, en la que se les plantearán una serie de problemas a resolver. La solución de estos problemas se mostrará al final de la clase. A continuación, vemos un ejemplo:

Ejemplo 1. *Crea una función que tenga como parámetros de entrada una matriz $\boldsymbol{A}$ y un número natural n, y devuelva $\boldsymbol{A}^n$. Resolver el problema usando recursividad.*

Sesión 3: Análisis Gráficos de Datos y Tests Estadísticos

Esta es la primera sesión con contenido estadístico, en la que se enseña a los alumnos la representación de datos con distintos gráficos, e.g., histogramas, diagramas de barras, diagramas de cajas y bigotes, así como comandos más específicos para definir color, leyendas o ejes. Es importante que no sólo aprendan a realizar gráficos, sino también a interpretarlos y elegir el más apropiado para cada situación.

A su vez, se les enseña algunos de los principales tests estadísticos, tales como: el test de independencia χ^2, tests de Normalidad de Kolmogorov-Smirnov y Shapiro-Wilk y test de

Kolmogorov-Smirnov para otras distribuciones, explicándoles el concepto de p-valor.

Sesión 4: Generación de Variables Aleatorias e Introducción al método de Monte Carlo

En esta sesión se enseña la importancia de generación de variables aleatorias y se realiza una pequeña introducción al método de Monte Carlo. Dichos conceptos juegan un papel importantes en labores de investigación. Para ello, es necesario en primer lugar que conozcan el concepto de "semilla" (seed en inglés). Entendemos que la mejor forma de asimilarlo es de manera visual, por lo que se propone el siguiente ejercicio.

Ejemplo 2. *Genera 20 observaciones de la variable aleatoria N(1,2). Ejecuta el programa realizado varias veces. ¿Qué observas? Con la ayuda de R busca la función set.seed() y aplícala al principio de tu programa. ¿Qué observas?*

Una manera divertida y muy visual de entender la simulación matemática son los diversos métodos que hay de estimar el número π. A continuación proponemos otro ejercicio que sirve a los alumnos para adentrase en el Método de Monte Carlo sin conceptos muy avanzados.

Ejemplo 3 (Estimación de π)**.** *Dado un cuadrado de lado 2, con vértices en (-1,-1), (-1,1),(1,-1), y (1,1); la probabilidad de que un punto generado aleatoriamente en ese cuadrado caiga dentro de la circunferencia de radio 1 y centro el origen es igual al área del círculo correspondiente entre el área del cuadrado, $\frac{\pi}{4}$. A partir de n puntos, hacer una estimación de π. Usa diferentes valores de n y observa las diferencias.*

Sesión 5: Modelos de Regresión, Análisis de Componentes Principales

En esta sesión, introducimos métodos estadísticos multivariantes de análisis y predicción. En concreto nos centramos en modelos de regresión y análisis de componentes principales. Es importante tener en cuenta para esta sesión que sólo unos pocos alumnos han visto estos conceptos en sus respectivas titulaciones. Así pues, antes de la propia aplicación práctica, es necesario hacer una explicación teórica de los modelos, así como de su importancia en la vida real. Para llevar a cabo dicha tarea se intentan transimitir las ideas intuitivas que subyacen de una manera sencilla y directa.

Sesión 6: Análisis de Datos Topológico y Resolución de dudas.

El análisis de datos topológico es una herramienta matemática que ha ganado gran importancia en los últimos años. En ella se mezclan herramientas de topología, álgebra y cada vez más de estadística y probabilidad [2, 6]. En esta sesión se dan unas pinceladas sobre esta materia y se enseñan los principales paquetes de R que incorporan esta herramienta. Esto se hace de manera superficial y teniendo en cuenta que algunos de los conceptos necesarios no se estudian hasta el último curso del grado en *Matemáticas*.

También se deja espacio para la resolución de dudas surgidas a lo largo de las sesiones, así como para la realización de una encuesta final, la cuál se detallará más adelante.

2.4. Evaluación

Al final de curso se realiza una sencilla prueba de evaluación. Esta prueba tiene dos objetivos principales:

- Motivar al alumnado. Al informar a los alumnos el primer día de clase de que se les realizará una prueba de evaluación, se consigue mantener una mayor atención, incluso en temas que pueden ser de menor interés personal.

- Acreditar. Además del anteriormente comentado interés que puede suscitar el tener una acreditación de R para entrar en el mundo laboral, desde el Decanato de la Facultad de CC. Matemáticas se pide aprobar esta prueba para poder realizar una convalidación de créditos de estudios.

En cualquier caso, y dado el carácter de la actividad, así como la diversidad de alumnos (tanto en titulaciones como en cursos), se le da la opción al alumno que no supere la prueba de resolver un ejercicio, adaptado a su nivel e intereses, para aprobar el curso.

La prueba constó de 11 preguntas y se realizó con el programa *kahoot*, que permite responder cuestiones desde dispositivos móviles o desde el propio ordenador personal (https://kahoot.com). Se puede elegir el tiempo de respuesta que se deja al alumno, las diferentes opciones y se puede adjuntar un gráfico a la pregunta. En la Figura 3 se puede observar un ejemplo de las preguntas del examen. El uso de un programa tan visual fue de gran acogida entre los alumnos, como era de esperar [9, 5]. Los resultados de la prueba fueron muy positivos, tal y como se muestra en la Figura 4.

Figura 3. Ejemplo de pregunta de la Prueba de Evaluación

3. Satisfacción del alumnado y posibles mejoras

Se realizó una encuesta al alumnado para ver su satisfacción con el curso. La encuesta constaba de 6 preguntas con 4 opciones cada una (Apéndice A). Se trata de un cuestionario totalmente anónimo y de carácter voluntario, a la que respondieron un total de 20 alumnos. En las Figura 5 recogemos alguno de los principales resultados. Podemos resumir las conclusiones obtenidas en los siguientes puntos:

- Una gran satisfacción con el curso. A la pregunta *¿cómo valorarías el curso?*, 19 alumnos respondieron que *útil* o *muy útil*, mientras que sólo un alumno respondió que *poco útil*. Además, la mayoría de los alumnos recomendarían el curso a un compañero. Esto está en concordancia con los buenos resultados de satisfacción del curso.

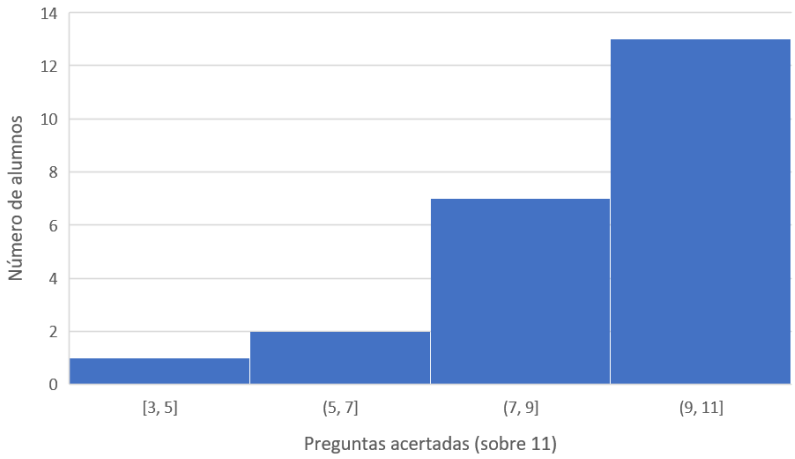

Figura 4. Resultados de la Prueba de Evaluación

- El ritmo del curso resultó algo rápido a algunos alumnos, mientras que a la mayoría les resultó adecuado. Sólo dos alumnos consideraron que había tenido un ritmo lento. Esto es explicable con la gran cantidad de materia en sólo 6 sesiones. Sería importante, por tanto, barajar la posibilidad de alargar un poco la duración del curso. También se plantea la opción de dar tutorías personales a aquellos alumnos que sientan que el ritmo es demasiado elevado.

- Si bien el tema que más había gustado varía mucho entre los alumnos, la mitad de ellos eligieron *Análisis gráfico y tests estadísticos*, lo que lleva a pensar sobre la eficacia de los sistemas de visualización para estudiantes con poco conocimiento de la materia [4].

- Más de la mitad de los alumnos dieron una valoración *muy buena* al material didáctico proporcionado. Esto, acompañado de las valoraciones personales de los alumnos a lo largo del curso, refuerza la importancia de dar al alumno una guía que ayude en esa labor de auto-aprendizaje.

4. Conclusiones

En el curso 2019-2020 se impartió el curso "Análisis de Datos con R" dentro de la inciativa "Compumates", de la Facultad de CC. Matemáticas, UCM. Este curso, ofrecido de manera gratuita, surgió con dos objetivos principales:

1. Proporcionar a todos los alumnos de la Facultad la opción de formarse en esta materia, con independencia de la titulación al que pertenezcan o al curso académico en el que se encuentren.

2. Ofrecer a los alumnos una formación básica que les permita acceder con más facilidad a un mercado de trabajo que cada vez valora más este tipo de conocimientos.

Como hemos visto a lo largo del artículo, el curso tuvo una gran acogida entre los estudiantes, sobre todo entre aquellos con titulaciones más teóricas (como el grado en *Matemáticas* y el doble grado en *Matemáticas y Física*). El curso constó de 6 sesiones de 2 horas cada una, en las

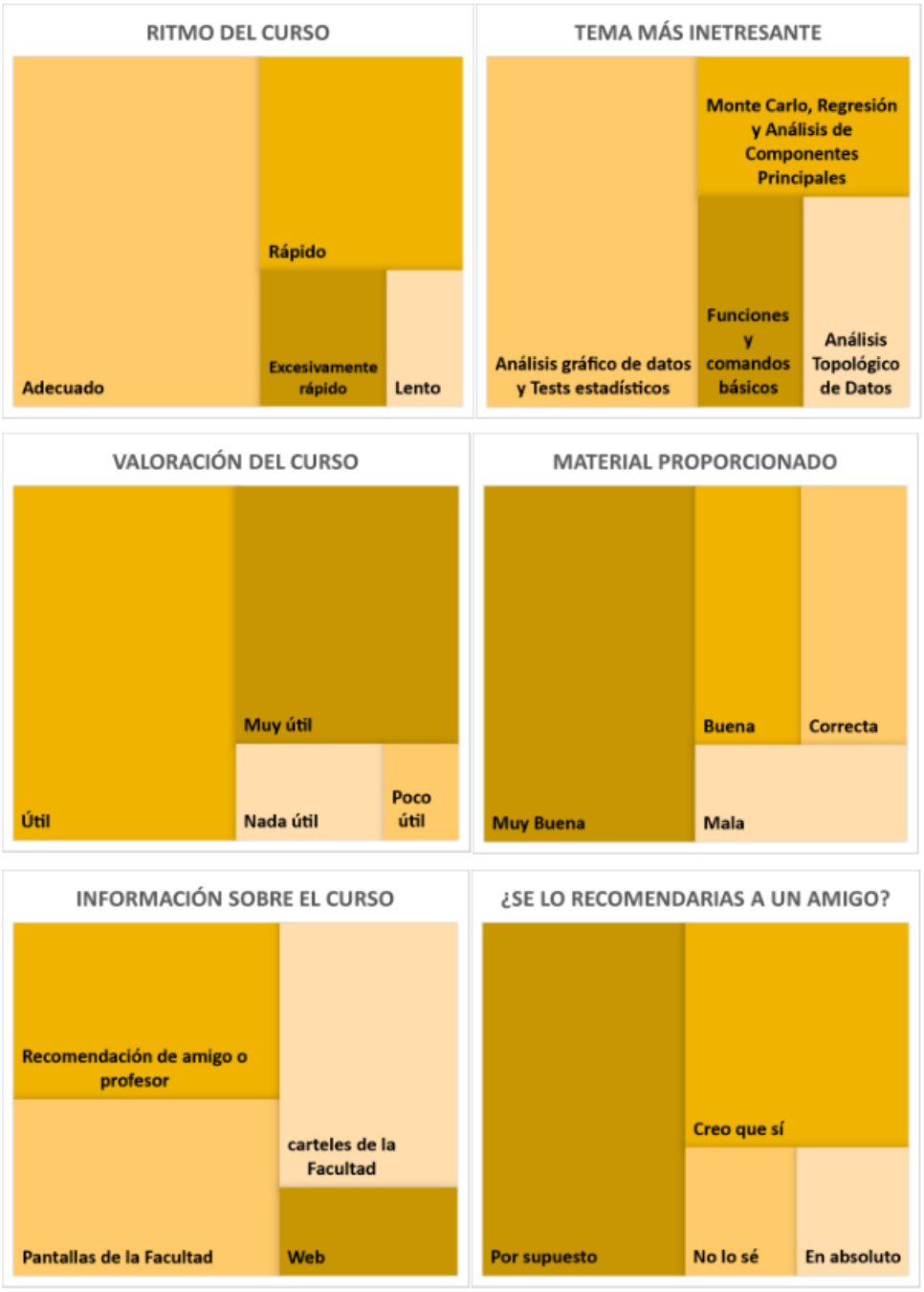

Figura 5. Resultados de la Encuesta de Satisfacción

que se enseñó desde la instalación del Software y ejecución de comandos básicos, hasta la aplicación de diversas técnicas estadísticas de análisis y predicción. Dada la naturaleza matemática de todos los alumnos, se dieron algunas pinceladas del análisis de datos topológico, área en pleno crecimiento en los últimos años. Mediante una prueba de evaluación, los alumnos acreditaron con grandes resultados su aprovechamiento del curso. Gracias a los resultados obtenidos de una encuesta final, voluntaria y anónima, vemos gran satisfacción entre los alumnos, que valoran muy positivamente que se les proporcione un manual para seguir el curso. Si bien todos los temas vistos son del agrado de los alumnos, el tema que recibe una valoración más alta es el de análisis gráfico de datos. Sin embargo, parece que el curso resulta rápido o excesivamente rápido para algunos estudiantes.

Tras la realización del curso y las conclusiones obtenidas de él, parece claro lo positivo de volver a realizar esta actividad en el futuro. Ahora bien, planteamos los siguientes puntos a desarrollar o mejorar en esa futura práctica:

- Pese a que el curso estaba dirigido a cualquier estudiante, con independencia de su facultad o universidad, en esta edición sólo hubo alumnos de la propia facultad. Si bien eso tiene sus ventajas, como asegurar unas bases de conocimiento comunes entre todos los alumnos, evita en gran medida la transversalidad de conocimiento [1]. Si la actividad sigue siendo abierta en un futuro, seria necesario mejorar su publicidad fuera del centro.
- El ritmo del curso resultó generalmente adecuado, pero unos pocos alumnos sintieron que éste era excesivamente rápido. Podría ser interesante ampliar el número de sesiones, o incluso reducir el número de alumnos en el curso, para poder hacer un seguimiento más personalizado. También se plantea la inclusión de horas de tutoría, para aquellos alumnos a los que les cueste más seguir el ritmo.
- La proporción de un manual didáctico surgió como respuesta a las sugerencias recibidas en una primera edición. La respuesta ha sido muy positiva, por lo que se ve necesaria la continua actualización y revisión del mismo.
- Dado que uno de los objetivos principales del curso es ayudar a los alumnos en su formación para la búsqueda de trabajo, sería interesante hacer un seguimiento posterior, en el que los propios alumnos valoraran si el curso les ha ayudado en su búsqueda de trabajo o qué mejoras proponen a posteriori.

A. Encuesta de Satisfacción

1. ¿Cómo valoras el curso?
   a) Muy útil
   b) Útil
   c) Poco útil
   d) Nada útil
2. El ritmo del curso ha sido...
   a) Excesivamente rápido
   b) Rápido
   c) Adecuado
   d) Lento
3. ¿Qué tema te ha interesado más?
   a) Funciones y comandos básicos (bucles, condicionales...)
   b) Montecarlo, Modelos de regresión y análisis de componentes principales
   c) Análisis gráfico de datos y Tests estadísticos
   d) Análisis topológico de datos
4. ¿Qué valoración das a las notas y material proporcionado?
   a) Muy buena
   b) Buena
   c) Correcta
   d) Mala
5. ¿Se lo recomendarías a un amig@?
   a) Por supuesto
   b) Creo que sí
   c) No lo sé
   d) En absoluto
6. ¿Cómo te enteraste de la existencia de este curso?
   a) Carteles pegados por la facultad
   b) Página web
   c) Pantallas de la facultad
   d) Recomendación de un amigo o profesor

Referencias

[1] Blanco Sandía, M. D. L. Á, Corchuelo Martínez-Azua, B., Corrales Dios, N., & López Rey, M. J. *Ventajas de la interdisciplinariedad en el aprendizaje: experiencias innovadoras en la educación superior*. XI Jornadas Internacionales de Innovación Universitaria , 2014.

[2] Carlsson, G. *Topology and data*. Bulletin of the American Mathematical Society, 2009, vol. 46, no 2, p. 255-308.

[3] Castilla , E., & Chocano, P.J. *Análisis de Datos con R*, 2020.
https://sites.google.com/view/elenacastillagonzlez/home/teaching/teaching
https://sites.google.com/view/pedrojchocanofeito/home/teaching-and-notes

[4] Compañ-Rosique, P., Satorre-Cuerda, R., Llorens-Largo, F. & Molina-Carmona, R. *Enseñando a programar: un camino directo para desarrollar el pensamiento computacional*. Revista de Educación a Distancia, 2015.

[5] Fernández-Vega, I., Jiménez, J. & Quirós, L. M. *Uso de la app Kahoot para cuantificar el grado de atención del alumno en la asignatura de Anatomía Patológica en Medicina y evaluación de la experiencia*. Educación Médica, 2020.

[6] Ghrist, R. *Barcodes: the persistent topology of data*. Bulletin of the American Mathematical Society, 2008, vol. 45, no 1, p. 61-75.

[7] Rexer , K., Gearan, P., & Allen, H. *Data science survey*. Rexer Analytics. Winchester, Massachusetts, 2015.

[8] Robert, A., *Muenchen. The popularity of data analysis software*, 2019.
http://r4stats.com/articles/popularity/

[9] Sempere, F. *Kahoot como herramienta de autoevaluación en la universidad.* IN-RED 2018. IV Congreso Nacional de Innovación Educativa y Docencia en Red. Editorial Universitat Politècnica de València, 2018.

Sobre el/los autor/es:
Nombre: Elena Castilla González
Correo electrónico: elecasti@ucm.es
Institución: Becaria FPU en el Departamento de Estadística e Investigacaión Operativa I, Universidad Complutense de Madrid.

Nombre: Pedro J. Chocano Feito
Correo electrónico: pedrocho@ucm.es
Institución: Becario FPI en el Departamento de Álgebra, Geometría y Topología, Universidad Complutense de Madrid.