Artículo de Actualización

Ventajas de R como herramienta para el Análisis y Visualización de datos en Ciencias Sociales
Advantages of R as a tool for data Analysis and Visualization in Social Sciences

Fernández Lizana, M. I.

Universidad Católica de Temuco, Facultad de Ciencias Sociales y Humanidades, Departamento de Antropología. Chile

RESUMEN
De entre los variados *software* gratuitos actualmente disponibles destaca uno que, por su cada vez mayor aceptación y promoción en los ámbitos académico e investigativo, se ha convertido en un potente referente en lo que respecta a la computación estadística de alto nivel como apoyo a las más diversas disciplinas científicas; nos referimos a R. Como tal R es un lenguaje de programación empleado primordialmente para efectuar análisis estadístico de datos y construcción de gráficos. Actualmente R es considerado la *lengua franca* de la estadística, debido a algunas de sus características que lo sitúan muy por encima de prácticamente todos sus "competidores": R es gratuito y libre, es muy versátil, permite realizar una cantidad insospechable de procedimientos estadísticos y gráficos, permite construir gráficos de calidad inmejorable, etc. Sumándonos a la "corriente R", y junto con esto a la filosofía misma del *software* libre, el presente trabajo tiene como principal objetivo presentar las amplias ventajas de R como herramienta para el análisis y visualización de datos en Ciencias Sociales.
Palabras clave: Software estadístico; Lenguaje de Programación R; Análisis de datos; Visualización de datos; Ciencias Sociales.

ABSTRACT
R is a programming language and free software environment for statistical computing and graphics. Currently, R is considered the *lingua franca* of Statistics due to some of it's characteristics that place it far above from practically all it's "competitors": R is free, it's very versatile, it allows to perform many statistical analysis, it allows to build high quality graphics, etc. Thus, and following the philosophy of *free software*, the main objective of this paper is to present the broad advantages that R has as a tool for data analysis and visualization in Social Sciences.
Keywords: Statistical Software; R Programming Language; Data Analysis; Data Visualization; Social Sciences.

INTRODUCCIÓN

En la actualidad resulta difícil imaginar a un investigador del área de las Ciencias Sociales realizando análisis estadísticos sin la ayuda de un *software*

***Autor Correspondiente: Miguel Ignacio Fernández Lizana**. Universidad Católica de Temuco, Facultad de Ciencias Sociales y Humanidades, Departamento de Antropología. Chile.*
Correo electrónico: miguel.ifl@gmail.com.
Fecha de recepción: Diciembre 2019 Fecha de aceptación: marzo 2020

especializado (Bologna, 2013). En términos generales, un *software* de análisis estadístico hace alusión a todos aquellos programas informáticos que disponen de módulos orientados a la tabulación, gestión, modificación, análisis y representación gráfica de datos. En efecto, el uso de este tipo de programas tiene considerables ventajas con respecto al cálculo manual, ya que permiten reducir el tiempo dedicado al análisis, aumentar su precisión, editar información, realizar representaciones gráficas y obtener salidas para elaborar informes, entre otras funciones. Ahora bien, no está de más reiterar que para lograr un uso adecuado -y por ende efectivo- de este tipo de *software*, se requiere del despliegue de conocimientos estadísticos que permitan efectuar de manera correcta los análisis e interpretaciones que potencialmente deban ser llevados a cabo. Por otro lado, a menudo se hace la distinción entre los *software* de pago (IBM SPSS Statistics, Stata, SAS & JMP, Minitab, Statgraphics, STATISTICA, SYSTAT, SPAD, XLStat, etc.) y los *software* gratuitos (PSPP, SAS University Edition, JASP, Past, InfoStat Versión estudiantil, MaxStat Lite Version, OpenStat, etc.). De entre los variados *software* gratuitos actualmente disponibles destaca uno que, por su cada vez mayor aceptación y promoción en los ámbitos académico e investigativo, se ha convertido en un potente referente en lo que respecta a la computación estadística de alto nivel como apoyo a las más diversas disciplinas científicas; nos referimos a R. Como tal R es un lenguaje de programación empleado primordialmente para efectuar análisis estadístico de datos y construcción de gráficos. Actualmente R es considerado la *lengua franca* de la Estadística (Mizumoto y Plonsky, 2015) debido a algunas de sus características que lo sitúan muy por encima de prácticamente todos sus "competidores": R es gratuito y libre, es muy versátil, permite realizar una cantidad insospechable de procedimientos estadísticos y gráficos, permite construir gráficos de calidad inmejorable, etc.Sumándonos a la "corriente R", y junto con esto a la filosofía misma del *software* libre (Stallman, 2015), el presente trabajo tiene como principal objetivo presentar las amplias ventajas de R como herramienta para el análisis y visualización de datos en Ciencias Sociales. Así, este articulo se divide en cuatro grandes apartados: en el primer apartado, llamado *SPSS: Nuestro viejo conocido y sus limitantes*, se expone por qué dicho *software* sigue siendo popular entre los científicos sociales y cuáles son sus principales limitantes; como su nombre lo indica, en el segundo apartado se presentan las *principales características y potencialidades del lenguaje de programación R*; en el tercer apartado, denominado *ventajas de R como herramienta para el análisis y visualización de datos en Ciencias Sociales*se mencionan diez ventajas que pueden ser tomadas en cuenta al momento de optar por R en tanto *software* de análisis estadístico; con el fin de demostrar cómo opera este lenguaje de programación, en el cuarto apartado se ofrece al lector una*ejemplificación de un uso concreto de R*, esto mediante la presentación de la *aplicación de un algoritmo bayesiano ingenuo para clasificar un conjunto de observaciones del dataset "Titanic"*;finalmente, en las *conclusiones*, junto con una dinámica del tipo "preguntas y respuestas", se vuelve sobre las ideas principales de este trabajo y se invita a los científicos sociales a optar por R en tanto una valiosa herramienta que permite realizar análisis y visualización de datos de forma versátil.

I. SPSS: NUESTRO VIEJO CONOCIDO Y SUS LIMITANTES

Que el *software* de análisis estadístico empleado por antonomasia en las Ciencias Sociales por muchos años haya sido SPSS no es ningún secreto para

los científicos sociales del mundo. Y que SPSS a día de hoy siga siendo popular entre los académicos, profesores y estudiantes de Ciencias Sociales se debe a múltiples factores, entre los que destacan la relativa facilidad de su uso, el abundante material disponible para aprender a manejar dicho *software* (libros, artículos, tutoriales, videos, cursos etc.), una buena cantidad de procedimientos estadísticos disponibles y el proceso mismo de endoculturación asociado a SPSS en el cual los estudiantes van internalizando la "tradición" de usar SPSS, tradición que, en algún momento de sus vidas como estudiantes, también fue internalizada por los que ahora son sus profesores. En este sentido -en lo que respecta al uso propiamente tal de SPSS- son muchos los científicos sociales (antropólogos socioculturales, sociólogos, psicólogos, etc.) que cada vez que se enfrentan con, por ejemplo, la elaboración de un informe, echan mano al famoso programa de IBM. Ahora bien, el profesor que quiera enseñar Estadística a sus estudiantes, perfectamente podrá hacerlo mediante SPSS (siempre y cuando la universidad en la que esté trabajando tenga a disposición este *software* en sus laboratorios de computación), pero una vez que sus estudiantes marchen hacia sus hogares, ¿podrán éstos tener acceso gratuito a SPSS con el fin de seguir practicando los contenidos del curso en casa o tendrán que conformarse con esperar a una nueva sesión de clases? Esta es una de las grandes limitantes de SPSS; su no gratuidad. Y si a esto sumamos el precio de este *software* (precio que sin lugar a dudas está muy alejado de la realidad presupuestaria de muchos estudiantes) nos encontramos ante un *software* que, si bien es cierto tiene algunas claras ventajas, no logra cumplir con todas las expectativas que nosotros nos imaginamos. Por otro lado, SPSS opera mediante una interfaz gráfica basada en menús que, a pesar de simplificarnos bastante el trabajo, puede en muchos casos provocar malos entendidos o inclusive puede propiciar el mal uso de procedimientos estadísticos; por ejemplo, cualquier persona haciendo unos pocos *clicks* puede realizar análisis multivariantes de distinta naturaleza en SPSS, pero ¿quién nos asegura que la persona que realiza dichos *clicks* es, en efecto, un conocer de los análisis multivariantes que hizo o que (vagamente) pretendió hacer? Y es que, querámoslo reconocer o no, cualquier persona puede sentirse un "experto analista de datos" con sólo hacer unos cuantos *clicks*. Lo anterior refleja una problemática que más allá de ser "anecdótica" constituye un grave falencia que se ve reflejada en la creencia de que el análisis estadístico de los datos se reduce a operaciones de las cuales sólo la computadora debe encargarse. Pero rememorando un famoso dicho de las Ciencias de la Computación, "*garbage in*, *garbageout*" ("basura entra, basura sale"), debemos siempre tener en cuenta que, independientemente del *software* que estemos utilizando, todos los procedimientos estadísticos responden a ciertas lógicas que debemos manejar con soltura con el fin de realizar análisis coherentes y en correspondencia con los supuestos teóricos que los sustentan (y no esperar que "mágicamente" la computadora "razone por nosotros").

II. PRINCIPALES CARACTERÍSTICAS Y POTENCIALIDADES DEL LENGUAJE DE PROGRAMACIÓN R

Sin querer presentar un explícito tutorial de cómo usar R, en este segundo apartado se presentan las principales características y potencialidades de este lenguaje de programación.

2.1. ¿Qué es R, para qué sirve y cuáles son sus orígenes?

R es un entorno y lenguaje de programación empleado primordialmente para efectuar análisis estadístico de datos y construcción de gráficos (R Core Team, 2019). R es un *software* libre que se asienta dentro del proyecto GNU y se distribuye bajo licencia GNU GPL (del inglés *General Public License* o Licencia Pública General) (Elosua, 2011). Dada su calidad de lenguaje, en R se emplean líneas de código. Para ejemplificar rápidamente lo anterior, a continuación se presenta una simple línea de código escrita con el fin de realizar una sencilla matriz de cuatro filas y cuatro columnas:

```
Matriz<-matrix(c(1,2,3,4,5,6,7,8), nrow = 4, ncol=4)

Matriz

     [,1] [,2] [,3] [,4]
[1,]    1    5    1    5
[2,]    2    6    2    6
[3,]    3    7    3    7
[4,]    4    8    4    8
```

Actualmente R es ampliamente usado en áreas como la bioestadística, el *data mining*, la econometría, la visualización de datos, etc. Como ya se ha planteado, R se utiliza primordialmente para efectuar análisis estadístico de datos y construir gráficos, tal como lo hacen muchos *software* de análisis estadístico, pero la potencia de R reside en su calidad de lenguaje. En este sentido, con R se pueden realizar muchas cosas más, como por ejemplo: (1) Analizar y/o editar imágenes. (2) Realizar análisis de sonido. (3) Analizar textos (cuantitativa y cualitativamente). (4) Generar mapas y realizar análisis espacial. (5) Realizar análisis de redes sociales (ARS). (6) Extraer, analizar y visualizar datos de páginas web (como Facebook, Twitter, Instagram, etc). (7) Escribir textos (artículos, libros, etc.) y publicarlos digitalmente. (8) Crear gráficos interactivos y animaciones, etc.[1]

En términos históricos, R es una derivación de los lenguajes de programación "S" y "Scheme" (Salas, 2008; Elosua, 2011). La primera versión de R fue desarrollada en 1993 en el Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda, por Ross Ihaka y Robert Gentleman (Ihaka y Gentleman, 1996). La letra "R" deriva de la primera letra de ambos nombres (**R**oss y **R**obert).

[1]Inclusive, con R se pueden realizar procedimientos que, en la práctica, están totalmente desvinculados del análisis y visualización de datos, como por ejemplo: (1) Enviar correos electrónicos; (2) Crear partituras y tablaturas para guitarra y; (3) Crear piezas de arte abstracto o, por lo menos, simularlas, etc.

2.2. ¿Por qué usar R? Los *pros* y "*contra*" de R

Demás está decir que R ostenta muchos más *pros* que *contras*(de ahí a que aquí se haya puesto la palabra *contra* "entre comillas" y en singular).

1. PROS:

- Es gratuito.
- Es libre. Cualquier usuario de R puede "desarrollar una nueva aplicación del mismo, puede modificar lo ya existente, puede compartirlo y usarlo libremente" (Ruiz-Ruano y Puga, 2016, p. 76).
- Es multiplataforma: R puede operar en Windows, Macintosh y Unix.
- Es muy versátil y potente.
- Gracias a los "paquetes" se pueden realizar una cantidad insospechable de procedimientos estadísticos y creaciones gráficas. Asimismo, se pueden construir gráficos de calidad inmejorable.
- Desde R fácilmente se pueden leer archivos previamente generados en otros programas, como Excel, SPSS, SAS, Stata, Minitab, etc.

2. CONTRA:

- El hecho que haya que escribir líneas de código hace que al principio sea (más o menos) arduo trabajar con R. Pero esto puede ser mitigado mediante algunas herramientas que sirven para hacer más llevadero el trabajo en R, como por ejemplo la IDE "RStudio" (RStudio Team, 2016)o la interfaz gráfica "R commander (Rcmdr)" (Fox, 2017).

Tomando como base el punto anteriormente expuesto ("contra"), a continuación mencionáremos brevemente una plataforma que puede ser empleada como medio para facilitar y agilizar el trabajo en R. Nos referimos a la ya mencionada IDE RStudio:

2.3. ¿Qué es y para qué sirve RStudio?

En términos generales, RStudio es una IDE (siglas en inglés de *Integrated Development Environment*), es decir, un "Entorno de DesarrolloIntegrado" exclusivamente diseñado para R. En palabras sencillas: RStudio es una herramienta que nos permite interactuar de forma más "amigable" con R. Lo anterior se expresa en cosas concretas, como el auto completado de código o el resaltado de sintaxis.

2.4. ¿Desde qué páginas podemos descargar R y RStudio y cuáles son los pasos a seguir?

Para descargar R nos dirigimos a: https://www.r-project.org/

Para descargar RStudio nos dirigimos a: https://www.rstudio.com/

Una vez descargadas ambas herramientas, debemos primero instalar R y luego RStudio. Es necesario indicar que tener instalado R es obligatorio para poder usar RStudio, de lo contrario no podremos usar esta IDE.

2.5. ¿Qué son y para qué sirven los "paquetes" en R?

En términos generales, un paquete (*package*), es un conjunto de funciones interrelacionadas que permiten, por ejemplo, efectuar análisis estadísticos específicos o gráficos concretos (aunque cabe destacar que no todos los paquetes están orientados a dichos procedimientos). Es decir, los paquetes permiten aumentar las funcionalidades que por defecto trae R (más adelante, en el cuarto apartado del presente trabajo, el lector podrá "ver en acción" algunos de estos paquetes). Algunos paquetes vienen instalados por defecto en R y otros deben ser instalados, lo cual resulta ser bastante sencillo. Los paquetes vienen acompañados con documentación de ayuda, bases de datos y fragmentos de código en tanto ejemplos que permiten poner a prueba las funciones que ostentan los paquetes. Los paquetes son creados, mantenidos, actualizados y puestos a disposición de la gente por personas desinteresadas que ayudan a engrandecer la comunidad de usuarios de R. Cabe destacar que algunos paquetes han sido creados por renombrados estadísticos y profesionales de las más diversas disciplinas científicas (Levshina, 2015). Y dado que actualmente R es la *lengua franca* de la estadística (Mizumoto y Plonsky, 2015), muchos de los nuevos desarrollos de esta ciencia son rápidamente implementados en R. Es por este motivo que no es ninguna novedad que los algoritmos más avanzados estén disponibles como funciones en algunos paquetes de R (Elosua, 2008). Por último es necesario mencionar que si bien es cierto existen muchos paquetes de uso general, algunos de éstos están orientados a ciertas ramas del conocimiento y otros están orientados a ciertos métodos.

III. VENTAJAS DE R COMO HERRAMIENTA PARA EL ANÁLISIS Y VISUALIZACIÓN DE DATOS EN CIENCIAS SOCIALES

Dado que aquí se parte del supuesto que indica que muchos de los lectores no conocen las amplias ventajas de R como herramienta para el análisis y visualización de datos en Ciencias Sociales, a continuación se mencionarán diez ventajas que, a juicio de este autor, pueden ser tomadas en cuenta al momento de optar por R en tanto *software* de análisis estadístico:

1. Si bien es cierto este trabajo no está centrado en el abordaje de los malos entendidos en torno a la estadística y al empleo poco razonado de algunos conceptos y operaciones provenientes de aquella disciplina (Wasserstein y Lazar, 2016; Wasserstein, Schirm y Lazar, 2019), es necesario recalcar aquí que dicha ciencia no debe ser nunca tomada "a la ligera" (Senn, 2008). Y es que, querámoslo reconocer o no, una de las grandes problemáticas en el marco de la enseñanza de la Estadística en las Ciencias Sociales, es que ésta -la Estadística- es enseñada muchas veces como una "caja negra" en nuestras disciplinas, es decir, se les dice a los estudiantes cosas como las siguientes: *introduzcan las variables X, Y y Z en SPSS y esperen que éste les arroje los resultados esperados y si es que dicho software les indica un p valor significativo*($p < 0.05$) *dense por satisfechos*. Con la analogía de la caja negra nos estamos refiriendo a que efectivamente se les plantea a los estudiantes qué procedimientos emplear al estar manipulando determinadas variables, pero no necesariamente se les enseña cómo funcionan en realidad dichos procedimientos, y esta caja negra también se ve potenciada al hacer un uso indiscriminado de *software* estadístico como, precisamente, el ya mencionado SPSS, el cual efectivamente opera como una caja negra en la cual se van introduciendo variables y se espera que aquel *software* arroje

por otro lado los resultados que más nos acomoden, pero no tenemos acceso al conocimiento teórico-algorítmico que subyace a los procedimientos estadísticos realizados por dicho *software*. En este sentido, y debido a que R no necesariamente es un programa en el cual se deban realizar simples *clicks* para obtener resultados, su utilización permite fomentar un uso más razonado, y por ende más adecuado de los procedimientos estadísticos comúnmente empleados en Ciencias Sociales (tanto a nivel básico como a nivel avanzando).

2. La gratuidad en cuanto a la descarga de R permite que no sólo los investigadores en el campo de las Ciencias Sociales tengan al alcance un potente programa, sino que también los docentes a cargo de cursos de Estadística puedan enseñar y promover sin restricción alguna el empleo de un *software* de análisis estadístico de alta calidad.

3. R está en constante evolución. Esto se traduce en que los algoritmos más avanzados estén disponibles como funciones en muchos de los paquetes de R (Elosua, 2008).

4. A diferencia de otros programas, los análisis de datos realizados en R producen salidas concisas "y dejan al usuario la opción de solicitar un mayor nivel de detalle" (Salas, 2008, p. 230) en el caso de que esto se requiera.

5. En términos históricos, la visualización de datos ha desempeñado un rol importante en las ciencias (Rahlf, 2017). Y en este sentido R no se ha quedado atrás, más bien todo lo contrario. Los gráficos realizados en R son de alta calidad; sin dejar de lado los aspectos puramente técnicos, y desde un punto de vista estético-visual, se puedan lograr creaciones gráficas visualmente bastante atractivas mediante algunos paquetes de visualización de datos. A esto debemos sumar que R nos ofrece el más amplio abanico de opciones en lo relacionado con la cantidad de gráficos que se pueden realizar. Muchos de los gráficos que sí se pueden crear en R simplemente no se pueden realizar en otros programas.

6. Existen paquetes de R desarrollados por científicos sociales y pensados para ser utilizados justamente por otros científicos sociales de distintas disciplinas. Prueba de ello la podemos encontrar en las decenas de paquetes orientados al análisis y visualización de datos en Ciencias Sociales y más específicamente en ámbitos como por ejemplo: (1) Encuestas y muestreo. (2) Econometría (regresiones, modelos ARIMA, econometría espacial, etc.). (3) Psicometría (análisis y visualización de escalas tipo Likert, teoría clásica de los test, teoría de respuesta al ítem, etc.). (4) Análisis y visualización de datos categóricos. (5) Análisis multivariante. (6). *Text mining* (minería de textos). (7). Análisis de redes sociales. (8). Inferencia causal. (9). Análisis bayesiano, etc.

7. Tal como se mencionó en el apartado dos, el hecho que haya que escribir líneas de código hace que sea relativamente arduo trabajar con R al principio. En ese mismo apartado se mencionaron dos herramientas que permiten hacer más llevadero el trabajo en R: la IDE "RStudio" (RStudio Team, 2016) y la interfaz gráfica "R commander (Rcmdr)" (Fox, 2017). No obstante, existen otras herramientas creadas con base en R y que están bastante más orientadas al tipo de datos que comúnmente son analizados en Ciencias Sociales. Ejemplosde aquello podemos encontrar en los dos siguientes *software* de descarga gratuita: (1) "IRAMUTEQ"(Interfaz de R para el Análisis Multidimensional de Textos y Cuestionarios) el cual es un *software*desarrollado por el francés Pierre Ratinaud (2014) al alero del

laboratorio LERASS[2]de la Universidad de Tolouse, y (2) "TXM", *software* orientado a la textometría (Heiden, Magué y Pincemin, 2010). También existen otras alternativas, como el paquete "RcmdrPlugin.temis" (Bouchet-Valat y Bastin, 2018) el cual, a diferencia de los dos *software* recién mencionados, permite analizar datos textuales sin "salir" del entorno R, debido a que éste se inserta en la interfaz gráfica "R commander (Rcmdr)" (Fox, 2017) (cabe destacar que "IRAMUTEQ", "TXM" y "RcmdrPlugin.temis" fueron desarrollados por franceses, lo cual, en parte, evidencia la tradición de la *escuela francesa de análisis de datos* y sus vínculos con el análisis de datos categóricos y textuales en el ámbito de las Ciencias Sociales).

8. R puede ser empleado como un *software* para el análisis cualitativo asistido por computador (programas conocidos en inglés como CAQDAS: *Computer Assisted/Aided Qualitative Data Analysis Software*). A pesar de las ostensibles particularidades numéricas de R, éste también puede ser utilizado como una alternativa gratuita de los tradicionales *software* de análisis cualitativo de pago (como ATLAS.ti, NVivo, etc.). A modo de ejemplo, y para más detalles, puede ser consultado el paquete de R denominado "RQDA" (Huang, 2018).

9. Es muy fácil compartir código escrito en R, esto permite fácilmente reproducir de forma exacta procedimientos estadísticos o gráficos creados por otras personas. Esto se asocia con lo que se denomina investigación reproducible. Existen algunos libros que tratan dicha temática, y específicamente mediante R (véase p. ej. Stodden, Leisch y Peng, 2014; Gandrud, 2015). Cabe destacar que ésta es una gran ventaja si comparamos a R con otros programas, debido a que muchas veces nos encontramos con informes técnicos, artículos o libros en los cuales se dejan ver procedimientos estadísticos o gráficos realizados con programas de pago que no nos permiten reproducir cómodamente dichos procedimientos. Con R este problema se ve superado, dado que tan sólo basta con tener los datos empleados[3] y el código utilizado para reproducir de forma exacta cualquier procedimiento realizado por cualquier autor de cualquier parte del mundo.

10. Por último, y no menos importante, se debe destacar que existe mucha información disponible para aprender a usar R. Son muchos los libros, artículos, tutoriales, videos, cursos *online* (y presenciales) y seminarios destinados a aquellas personas que quieran iniciarse en R o que quieran profundizar conocimientos en áreas específicas de la Estadística a través de R (en este mismo trabajo se presentan, en la tabla N°1, algunas páginas web junto con breves descripciones y sus links, en las que se puede encontrar abundante material disponible sobre el lenguaje R). Y aparte de los en extremo abundantes materiales en inglés, existe también mucha documentación pensada desde y para el mundo hispanohablante, esto se constituye en un "punto a favor" de R, dado que existen otros lenguajes de programación que, a pesar de ostentar una potencialidad interesante, no cuentan con el suficiente material disponible para aprender a utilizarlos, quedando así bajo la influencia y/o el dominio casi exclusivo de los

[2]*Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales*. URL https://www.lerass.com/

[3] Es más, en muchos casos ni siquiera es necesario tener una matriz de datos como tal, dado que algunos autores optan por incluir los datos a analizar en el mismo código escrito en R.

profesionales especializados en programación o en general en Ciencias de la Computación.

Tabla 1: Páginas web en las que se puede encontrar material disponible sobre el lenguaje de programación R

N°	Página web	Breve descripción	Link
1	**The R Project for Statistical Computing**	Página oficial del *R Project for Statistical Computing*. Desde esta página se descarga el *setup* que nos permitirá instalar R.	**https://www.r-project.org/**
2	**RStudio**	Página oficial de *RStudio*. Desde esta página se descarga el *setup* que nos permitirá instalar la IDE *RStudio*.	**https://www.rstudio.com/**
3	**RPubs**	Página mantenida por los desarrolladores de *RStudio* con el fin de permitirles a los usuarios de R redactar documentos directamente desde *RStudio* -a través de *R Markdown*- para posteriormente ser publicados online en la propia página *RPubs*. Cabe destacar que en *RPubs* nos podemos encontrar con muchas publicaciones en español orientadas a la enseñanza de R. **Nota:***R Markdown* permite "generar documentos que combinen texto, imágenes e instrucciones de R, más los resultados que dichas instrucciones produzcan (estos resultados pueden ser simples valores numéricos, tablas o gráficos)" (Santana y Nieves, 2016).	**https://rpubs.com/**
4	**Bookdown**	Página creada con el fin de permitirles a los usuarios de R redactar libros directamente desde *RStudio* -a través de *R Markdown*- para posteriormente ser publicados online en la propia página *Bookdown*. Cabe destacar que en *Bookdown* nos podemos encontrar con muchos libros online de acceso gratuito orientados a la enseñanza de R.	**https://bookdown.org/**
5	**R-bloggers**	Siendo una de las páginas más populares entre los usuarios de R, *R-bloggers* está constituida por muchas personas que comparten sus escritos sobre tutoriales, noticias relacionadas con el mundo de R, etc.	**https://www.r-bloggers.com/**
6	**Rseek**	Más que una página como tal, *Rseek* es un buscador centrado totalmente en R y en sus materiales asociados.	**http://rseek.org/**
7	**The R Journal**	Revista de acceso libre del *R Project for Statistical Computing*, y en la que se publican artículos pensados tanto para los usuarios como para los desarrolladores de R.	**https://journal.r-project.org/**
8	**The Journal of Statistical Software**	Revista de acceso libre en la que se publican artículos, *reviews* de libros, fragmentos de código y *reviews* de *software* estadístico. Si bien es cierto esta revista no está totalmente centrada en R, la mayoría de sus artículos sí están relacionados con este lenguaje.	**https://www.jstatsoft.org/index**
	Microsoft R Application	Página mantenida por *Microsoft*. Esta página incluye un buscador de paquetes de R (por ejemplo, si en dicho buscador escribimos "*Social Sciences*", éste nos	

9	**Network.**	arrojará una lista con todos los paquetes de R vinculados con procedimientos matemáticos y estadísticos aplicados a las Ciencias Sociales: econometría, psicometría, encuestas y muestreo, análisis y visualización de datos categóricos, análisis multivariante, análisis de redes sociales, inferencia causal, etc.).	**https://mran.microsoft.com/**
10	**CRAN Task View: Statistics for the Social Sciences**	Página mantenida por el sociólogo y estadístico estadounidense John Fox (creador de la anteriormente mencionada interfaz gráfica *R commander (Rcmdr)*. Esta página reúne muchos materiales sobre Estadística aplicada a las Ciencias Sociales a través de R.	**https://cran.r-project.org/web/views/SocialSciences.html**
11	**Data Visualization for Social Science. A practical introduction with R and ggplot2**	Libro *online* de acceso gratuito escrito por el sociólogo estadounidense Kieran Healy. Este libro está orientado al aprendizaje de la visualización de datos en Ciencias Sociales, para ello Healy emplea R y un paquete en específico llamado *ggplot2* (Wickham, 2009), el cual es ampliamente usado en la construcción de gráficos de alta calidad.	**http://socviz.co/**
12	**Text Mining with R**	Libro *online* de acceso gratuito escrito por los *data scientist* estadounidenses Julia Silge y David Robinson. Este libro procura ser una introducción a la minería de textos o *text mining* mediante R.	**http://tidytextmining.com/**
13	**Estadística básica con R**	Libro de descarga gratuita en formato PDF escrito por el estadístico español Guillermo Ayala. Este libro procura ser una introducción al análisis estadístico de datos mediante R.	**http://www.uv.es/ayala/docencia/nmr/nmr13.pdf**
14	**El arte de programar en R: un lenguaje para la estadística**	Libro de descarga gratuita en formato PDF escrito por los mexicanos Julio Santana (informático) y Efraín Farfán (oceanógrafo). Si bien es cierto los autores de este libro no están vinculados con el mundo de las Ciencias Sociales, este libro expone de forma sencilla la potencialidad analítica que R pone a disposición de cualquier científico.	**https://cran.r-project.org/doc/contrib/Santana_El_arte_de_programar_en_R.pdf**

Fuente: Elaboración propia.

IV. EJEMPLIFICACIÓN DE UN USO CONCRETO DE R: APLICACIÓN DE UN ALGORITMO BAYESIANO INGENUO PARA CLASIFICAR UN CONJUNTO DE OBSERVACIONES DEL *DATASET* "TITANIC"

Antes de entrar de lleno al ejemplo aquí propuesto, y sin entrar en mayores detalles técnicos, a continuación se explica en términos generales qué es un clasificador bayesiano ingenuo. En *machine learning*[4], el concepto "clasificador bayesiano ingenuo"(*naive bayes classifier*)alude a un clasificador basado en el teorema de Bayes. El término "ingenuo" le es dado a dicho clasificador debido a

[4] Conocido en español como "aprendizaje automático", el *machine learning* es una rama de las Ciencias de la Computación que emplea inteligencia artificial para identificar patrones en conjuntos de datos grandes y complejos. A su vez, los algoritmos de *machine learning* se dividen en dos grandes grupos: los algoritmos supervisados y los no supervisados. Sobre la aplicación del *machine learning* en el ámbito de las Ciencias Sociales se puede consultar el libro *Big Data and Social Science: A Practical Guide to Methods and Tools* (Foster, Ghani, Jarmin, Kreuter y Lane, 2016).

que éste asume la independencia de las variables predictoras (Zhang, 2016; Harzevili y Alizadeh, 2018). Generalmente en la literatura especializada este algoritmo es descrito como uno de los más sencillos pero, a su vez, uno de los más efectivos clasificadores (Zhang, 2016; Feki-Sahnoun, Njah, Hamza, Barraj, Mahfoudi, Rebai y Bel Hassen, 2018; Khanna y Sharma, 2018; Shiri y Alizadeh, 2018). Tal como lo indica el título de este apartado aquí se utiliza el *dataset* "Titanic", el cual incluye variables categóricas y numéricas vinculadas con la tragedia del transatlántico británico RMS Titanic, hundido en 1912. Variables como sexo, edad, clase, supervivencia o no supervivencia del pasajero, etc. (Dawson, 1995). A modo de justificación es preciso indicar que la elección de este conjunto de datos se hizo con fines netamente ilustrativos. Sin embargo, cabe destacar que el *dataset* "Titanic" es ampliamente reconocido y popular en la comunidad de científicos de datos (*data scientist*) debido a que éste comúnmente es utilizado con fines educativos y con fines "probatorios", esto último en el sentido de que a menudo este *dataset* es sometido a diversos análisis mediante una serie de algoritmos con la finalidad de poner a prueba la potencia y/o efectividad de dichos algoritmos (Friendly, Symanzik y Onder, 2019). Por otro lado, y si bien es cierto el afán de este apartado no es mostrar en detalle cómo analizar y visualizar datos con R (lo cual podría ser tratado en otras publicaciones), a juicio de este autor el ejemplo que a continuación se presenta puede servir para (1) evidenciar la versatilidad y potencia de R y (2) despertar el interés de cualquier investigador social que vislumbre la utilidad de este lenguaje de programación. Pues bien, para implementar el algoritmo señalado se utilizó la función *naiveBayes*incluida en un paquete de R llamado *e1071* (Meyer, Dimitriadou, Hornik, Weingessel y Leisch, 2019). Aclarado lo anterior, a continuación se presenta el ejemplo aquí propuesto: nuestra labor es determinar la supervivencia o la no supervivencia de los pasajeros del Titanic, a partir de las variables sexo, edad y clase. A continuación se presenta el código escrito en R y posteriormente los resultados:

```
# Cargamos el paquete 'e1071' y el dataset 'Titanic'.
library(e1071)
data("Titanic")
# El dataset 'Titanic' lo guardamos como 'data frame'.
Titanic_df <- as.data.frame(Titanic)
# Creamos nuevos datos a partir del data frame elaborado.
rep_secuencia <- rep.int(seq_len(nrow(Titanic_df)),
Titanic_df$Freq)
# Creamos un dataset a partir de los nuevos datos elaborados.
Titanic_dataset <- Titanic_df[rep_secuencia,]
# Ya no necesitamos las frecuencias.
Titanic_dataset$Freq = NULL
# Ajustamos el clasificador bayesiano ingenuo.
Modelo_Naive_Bayes <- naiveBayes(Survived ~., data =
Titanic_dataset)
# ¿Qué nos indica el modelo? Pedimos a R el resumen del modelo.
Modelo_Naive_Bayes
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)
```

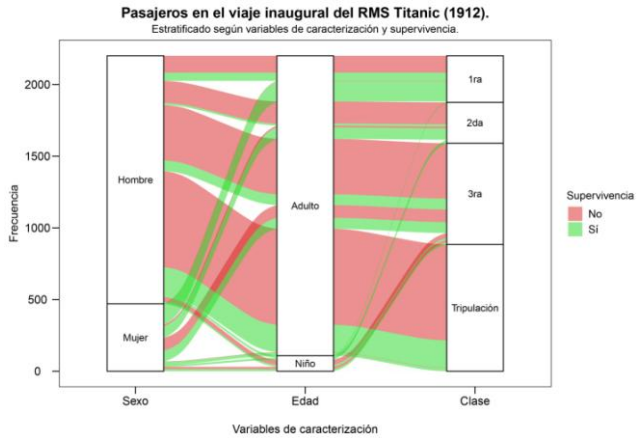
```
A-priori probabilities:
Y
      No      Yes
0.676965 0.323035

Conditional probabilities:
     Class
Y            1st        2nd        3rd       Crew
  No  0.08187919 0.11208054 0.35436242 0.45167785
  Yes 0.28551336 0.16596343 0.25035162 0.29817159

     Sex
Y           Male     Female
  No  0.91543624 0.08456376
  Yes 0.51617440 0.48382560

     Age
Y          Child      Adult
  No  0.03489933 0.96510067
  Yes 0.08016878 0.91983122

# Hacemos unapredicción a partir de los datos del dataset.
MNB_Predic <- predict(Modelo_Naive_Bayes, Titanic_dataset)
# A continuación creamos una matriz de confusión[5]
# para determinar las observaciones
# adecuadamente clasificadas.
table(MNB_Predic, Titanic_dataset$Survived)
MNB_PredicNo  Yes
           No  1364  362
           Yes  126  349
```

Finalmente, podemos realizar un gráfico/diagrama del tipo "alluvial"para visualizar el procedimiento anteriormente realizado. Para esto empleamos los paquetes *ggplot2* (Wickham, 2009) y *ggthemes* (Jeffrey, 2019), en conjunto con el -en este caso- esencial paquete *ggalluvial* (Brunson, 2018). A continuación se presenta el código escrito en R y posteriormente el gráfico:

```
Titanic <- read.csv("C:/.../Titanic.csv", sep=";")
attach(Titanic)

library(ggplot2)
library(ggalluvial)
library(ggthemes)

colores <- c("#D92121", "#21D921")

ggplot(data = Titanic,
       aes(axis1 = Sexo, axis2 = Edad, axis3 = Clase,
           y = Frecuencia)) +
  scale_x_discrete(limits = c("Sexo", "Edad", "Clase"), expand =
c(.1, .05)) +
  xlab("\nVariables de caracterización") +
  geom_alluvium(aes(fill = Supervivencia)) +
  geom_stratum() + geom_text(stat = "stratum", label.strata = TRUE)
+
  theme_par() +
  ggtitle("Pasajeros en el viaje inaugural del RMS Titanic
(1912).",
```

[5]En Ciencias de la Computación, particularmente en inteligencia artificial, una matriz de confusión permite caracterizar el desempeño de un algoritmo.

```
          "Estratificado según variables de caracterización y
supervivencia.") +
  theme(plot.title = element_text(color="black", size=16),
      axis.title.x = element_text(color="black", size=13),
      axis.title.y = element_text(color="black", size=13))+
  theme(axis.text.x = element_text(color="black",size=13),
        axis.text.y = element_text(color="black", size=13)) +
  theme(plot.title = element_text(hjust=0.5)) +
  theme(plot.subtitle = element_text(hjust=0.5)) +
  scale_fill_manual (values = colores)
```



Fuente: Elaboración propia.

Gráfico 1.

CONCLUSIONES

En su artículo *¿Existe vida más allá del SPSS?* Paula Elosua (2008) sostiene lo siguiente en torno a R: "¿Existe algo mejor? Libre, gratuito, asequible, accesible y siempre a la vanguardia" (Elosua, 2008, p. 652). Las palabras de esta autora pueden resumir bastante bien lo que constituye el lenguaje de programación R. Sin embargo, y con justo derecho, alguien podría plantear las siguientes preguntas: ¿Por qué debería usar R, si mis análisis estadísticos y creaciones gráficas las puedo realizar perfectamente con programas que ya conozco y los cuales son mucho más fáciles de usar? y ¿por qué debería aprender a programar en R, si para calcular cosas tan sencillas como una media aritmética no es necesaria tanta "parafernalia técnica"?En este sentido, las cuatro siguientes preguntas y sus respuestas asociadas podrían esclarecer aún más este tópico:

1. ¿Es necesario ser "programador" para trabajar con R o tener conocimientos y experiencia previa en programación? La respuesta es NO: no es necesario ser programador ni tener conocimientos y experiencia previa en programación para trabajar con R. Con toda seguridad, un programador experto o una persona que cuente con experiencia previa en programación podrá aprender mucho más rápido a trabajar con R, dado que algunos lenguajes de programación comparten ciertas características, pero aquellas personas que no cuenten con dichos conocimientos perfectamente pueden aprender a usar R.

2. ¿Es necesario tener conocimientos avanzados de Estadística para trabajar con R? Otra vez, la respuesta es NO: si bien es cierto es necesario poseer conocimientos de Estadística para trabajar con cualquier *software* orientado a dicha materia, no es menos cierto que, por ejemplo, para realizar gráficos de barra o calcular promedios, no hace falta ser un "experimentado estadístico".

3. ¿Deberé aprender de memoria "todas" las funciones de R y sus argumentos asociados o, dicho en otras palabras, tendré que memorizar todas esas líneas de código? Aunque la respuesta a esta pregunta sea absolutamente obvia, no está de más aclararla: NO, al igual que en cualquier lenguaje de programación, con R no es necesario memorizar miles y miles de líneas de código; ya sea porque es humanamente imposible o porque resulta ser un completo sinsentido. Debemos recordar que el uso de los *scripts*[6]en R, hacen que el trabajo sea mucho más llevadero de lo que uno podría pensar al principio.

4. Está bien, pero, ¿qué tan difícil es usar R? Por supuesto, la respuesta a esta pregunta dependerá de múltiples factores. Pero, si quisiéramos entregar una respuesta rápida y sencilla (y probablemente evidente), diríamos que sólo basta desplegar constancia y trabajo, tanto como para aprender, como para usar R. Sin embargo, una vez asimilada la sintaxis de R (¡asimilada, no memorizada!), el trabajo se torna más fácil, lo cual se traduce en una mayor rapidez a la hora de escribir código.

Ahora bien, y tomando en cuenta todos los argumentos expuestos en los apartados anteriores, aquí se hace una invitación a los científicos sociales a optar por R en tanto una valiosa herramienta que permite realizar análisis y visualización de datos de forma versátil, potente, replicable, y mejor aún, de manera totalmente gratuita y libre.

REFERENCIAS BIBLIOGRÁFICAS

Bologna, E. (2013). *Estadística para Psicología y Educación* (3raedición ampliada). Córdoba: Brujas.

Bouchet-Valat, M. & Bastin, G. (2018). Rcmdr Plugin.temis: Graphical Integrated Text Mining Solution. R package version 0.7.10. URL https://cran.r-project.org/web/packages/RcmdrPlugin.temis/index.html

Brunson, J. (2018). ggalluvial: Alluvial Diagrams in 'ggplot2'. R package version 0.9.1. URL https://CRAN.R-project.org/package=ggalluvial

Dawson, R. (1995). The 'Unusual Episode' Data Revisited. *Journal of Statistics Education*, *3*(3), 1-9.

Elosua, P. (2009). ¿Existe vida más allá del SPSS? Descubre R. *Psicothema*, *21*, 652-655.

Elosua, P. (2011). *Introducción al entorno R*. Bilbao: Euskal Herriko Unibertsitateko Argitalpen Zerbitzua.

Feki-Sahnoun, W., Njah, H., Hamza, A., Barraj, N., Mahfoudi, M., Rebai, A. & Bel Hassen, M. (2018). Using general linear model, Bayesian Networks and Naive Bayes classifier for prediction of *Karenia selliformis* occurrences and blooms. *Ecological Informatics*, *43*, 12-23.

Foster, I., Ghani, R., Jarmin, R., Kreuter, F. & Lane, J. (2016). *Big Data and Social Science: A Practical Guide to Methods and Tools*(1st Ed). Boca Raton, FL: Chapman & Hall/CRC Press.

Fox, J. (2017). *Using the R Commander: A Point-and-Click Interface for R*. Boca Raton FL: Chapman & Hall/CRC Press.

[6] Un *script* es simplemente un documento o archivo que contiene un conjunto de líneas de código.

Friendly, M., Symanzik, J. & Onder, O. (2019). Visualising the Titanic disaster. *Significance*, *16*(1), 14-19.

Gandrud, C. (2015). *Reproducible Research with R and R Studio* (2nd Ed). Boca Raton, FL: CRC Press.

Harzevili, N. & Alizadeh, S. (2018). Mixture of latent multinomial naive Bayes classifier. *Applied Soft Computing*, *69*, 516-527.

Heiden, S., Magué, J-P. & Pincemin, B. (2010). TXM: Une plateforme logicielle open-source pour la textométrie - conception et développement. En S, Bolasco, I, Chiari. & L, Giuliano (Eds.), *Proceedingsof 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010, Vol. 2* (pp. 1021-1032). Roma: Edizioni Universitarie di Lettere Economia Diritto.

Huang, R. (2018). RQDA: R-based Qualitative Data Analysis. R package version 0.3-1. URL http://rqda.r-forge.r-project.org.

Ihaka, R. & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*, 299-314.

Jeffrey, A. (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.1.0. URL https://CRAN.R-project.org/package=ggthemes

Khanna, D. & Sharma, A. (2018). Kernel-Based Naive Bayes Classifier for Medical Predictions. En V, Bhateja., C, Coelo., S, Chandra. & P, Kumar (Eds.), *Intelligent Engineering Informatics. Proceedings of the 6th International Conference on FICTA*(pp. 91-101). Singapore:Springer Nature.

Levshina, N. (2015). *How to do Linguistics with R. Data exploration and statistical analysis*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-1. https://CRAN.R-project.org/package=e1071

Mizumoto, A. & Plonsky, L. (2015). R as a Lingua Franca: Advantages of Using R for Quantitative Research in Applied Linguistics. *AppliedLinguistics*, *37*(2), 284-291.

Rahlf, T. (2017). *Data Visualisation with R, 100 Examples*. Cham: Springer Nature.

Ratinaud, P. (2014). Iramuteq. Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. Un logiciel libre construit avec des logiciels libres. Version 0.7 alpha 2. URL http://www.iramuteq.org/

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

RStudio Team. (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL http://www.rstudio.com/

Ruiz-Ruano, A. & Puga, J. (2016). R como entorno para el análisis estadístico en evaluación psicológica. *Papeles del Psicólogo*, *37*(1), 74-79.

Salas, C. (2008). ¿Por qué comprar un programa estadístico si existe R? *Ecología Austral, 18,* 223-231.

Santana, A. & Nieves, C. (2016). *Generación de documentos con R Markdown*. Las Palmas de Gran Canaria: Departamento de Matemáticas de la Universidad de Las Palmas de Gran Canaria (ULPGC).

Senn, S. (2008). *Statistical Issues in Drug Development*(2nd Ed). West Sussex: Wiley.

Stallman, R. (2015). *Free Software, Free Society: Selected Essays of Richard M. Stallman*(3rd Ed). Boston, MA: Free Software Foundation.

Stodden, V., Leisch, F. & Peng, R. (2014). *Implementing Reproducible Research*. Boca Raton, FL: CRC Press.

Wasserstein, R. & Lazar, N. (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70,* 129-133.

Wasserstein, R., Schirm, A. & Lazar, N. (2019). Moving to a World Beyond "$p < 0.05$". *The American Statistician*, *73*(sup1), 1-19.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Zhang Z. (2016). Naïve Bayes classification in R. *Annals of Translational Medicine*, *4*(12), 1-5.