FACULTAD DE CIENCIAS SOCIALES Y JURIDICAS DE ELCHE

ESTADÍSTICA EMPRESARIAL

TRABAJO DE FIN DE GRADO
CURSO 2020-2021

TÉCNICAS DE MACHINE LEARNING APLICADAS AL ANÁLISIS DE INTENCIÓN DE COMPRA ONLINE

ALUMNO: ALICIA DIAZ MARTINEZ

TUTOR: JOSE LUIS SAINZ-PARDO AUÑON

ÍNDICE

1.RESUMEN

En la actualidad, se presenta un continuo auge en lo que refiere a la compra online debido a que los clientes encuentran este método más cómodo y a su vez, es más factible localizar entre una gran diversidad de productos lo que concretamente se desea. Es por ello, que los comercios deben actualizar ininterrumpidamente su plataforma digital con ayuda de los datos más importantes sobre su comercio, como puede ser la cantidad de clientes que visitan su página web o los ingresos que generan de manera online.

Todo lo anteriormente mencionado, se refleja en la analítica web, una disciplina que actualmente se encuentra en continuo crecimiento. En ella, se recogen todos los datos de interés para una página web, destacando la función influyente en la misma, de la herramienta Google Analytics.

Muchos datos de interés para la página web de un negocio se obtienen gracias a la analítica web, permitiéndose así conocer el comportamiento del cliente que acceda a la misma.

En el presente informe y mediante la aplicación de diferentes técnicas de Machine Learning, como Componentes Principales y diversas técnicas de clasificación de observaciones, en concreto, Análisis Cluster, K-Nearest Neigbour y Support Vector Machine, cuya finalidad consiste en la simplificación de la base de datos, así como de enriquecer la competencia y la profesionalidad del negocio online en el mercado, se desarrolla un análisis sobre las compras adquiridas por los clientes que inician sesión en cierto negocio web durante un año, con el fin de evaluar que método de los empleados, proporciona un mayor nivel de fiabilidad, mediante la división de los datos en dos, la mitad utilizados como datos de entrenamiento, y el resto de ellos, para tratamiento de prueba, de tal forma que se genere un modelo apto, que permita estudiar futuros datos que recoja la empresa, incluso predecir las futuras compras llevadas a cabo por clientes que acceden a la página web.

ABSTRACT

Nowadays, a continuous boom of online shopping is shown due to the fact that customers find this method more comfortable and at the same time, they find a great diversity of products. For this reason, businesses must continuosly update their digital platform with the help of the most important data about their business, such as the number of customers visiting their website or the revenue they generate online.

Al lof above is reflected in web analytics, a discipline that is in continuous growth. In it, all the data of interest to a website are collected, highlighting the influential role in it, the Google Analytics tool.

Many interesting data for the website of a business are obtained thanks to web analytics, thus allowing to know the behaviour of the customer who accesses it.

In this report and though the application of the different Machine Learning techniques, like Principal Components, and several classification techniques, in particular, Cluster Analysis, K-Nearest Neigbour and Support Vector Machine, whose purpose is to simplify the database, as well as to enrich the competence and professionalism of the online business in the market, an analysis is developed on the purchases made by customers who log in to a certain web business during a year, in order to evaluate which of the methods used, provides a higher level of reliability, by diving the data in two, half of them, for test treatment, in such a way that a suitable model is generated, through which, it is possible to study future data collected by the company, even to predict future purchases made by future customers of the website.

2.INTRODUCCIÓN

“Las compras “online” vuelven a batir récord al crecer un 286% “– La Vanguardia 27/04/2020. [1]

“El consumidor tras el coronavirus: más compras por internet…” – El país 13/06/2020. [2]

Ambas noticias destacan la importancia que ha adquirido el tráfico comercial online en los últimos años, a causa de la aparición del COVID-19 y, por consiguiente, el confinamiento de una elevada cantidad de países, lo que provocó que incluso personas de elevadas edades, que en ocasiones anteriores se veían incapaces de adquirir conocimientos tecnológicos, se hayan visto obligadas a hacer uso de la compraventa online.

Este hecho denota una creciente expansión de las compras por internet, provocando una necesidad por parte de todos los tipos de comercios a comprometerse con la creación, desarrollo o mejoras de páginas web para sus negocios.

Y es que el éxito del comercio online, no se basa únicamente en la belleza aparente de la web comercial, sino en la importancia de generar un completo análisis sobre los clientes que acceden a la web, sus características o la duración que generan en la misma, así como diversos elementos fundamentales que existen en los negocios online.

Cabe destacar que, el logro del conocimiento completo de cierta web comercial, se visualiza desde la perspectiva del Big Data por la masiva cantidad de datos que pueden ser analizados.
Es por ello que, gracias a diferentes técnicas estadísticas, y en concreto, mediante las técnicas de machine learning, es posible alcanzar un conocimiento completo de los clientes propios de un comercio.

La motivación de realizar el presente informe, proviene del interés en la aplicación de técnicas de machine learning en término de negocios online, con el fin de poder adquirir amplios conocimientos sobre aquellos elementos de

estudio trascendentales para cualquier empresa permitiendo cumplir sus objetivos, así como mejorar sus estrategias referentes a sus clientes e incluso las de mercado.

La finalidad del presente estudio se fundamenta en, mediante el uso de técnicas estadísticas, la previsión de que intención poseen los usuarios que inician sesión y perduran un tiempo concreto en las distintas categorías de las que goza la página web de un comercio, de comprar o no hacerlo, al finalizar su visita en la misma. Se tendrán en cuenta ciertas características del inicio de sesión que efectúan los clientes, como podría ser, la duración en la página web o el tipo de página a la que acceden. Además, finalmente se ofrecen propuestas futuras que sean de uso para mejorar las características que contiene el comercio online, y, por lo tanto, influenciar el incremento de los beneficios generados por las compras de los clientes en los próximos años.

3.ESTADO DE LA CUESTIÓN Y MARCO TEÓRICO

3.1. ANALITICA WEB, BIG DATA E INTELIGENCIA ARTIFICIAL

La analítica web se ha convertido en uno de los instrumentos de elevada relevancia para los negocios que han decidido lanzarse al comercio online. Sus inicios se remontan unos años atrás con la aparición y desarrollo del mundo tecnológico, surgiendo así lo que conocemos por Internet.

Tras ello, surgieron los primeros programas informáticos, los cuales, comenzaban a recopilar y trabajar con datos. Con el crecimiento constante de estos nuevos sistemas, el Marketing divisó una gran posibilidad de mejora que aportaría mayor simplicidad en sus estudios dirigidos a los productos en venta, por lo tanto, decidió comprometerse con esta nueva era tecnológica. Las empresas comprendieron que podían generar un sistema donde exponer todos sus productos o servicios, de manera que, estos mismos podrían actualizarse de manera constante.

A pesar de ello, no fue hasta el año 2007, cuando surgió el auge total de la publicidad online, cuando aparecieron diversos softwares que servirían a las

empresas para llevar a cabo ciertas mediciones sobre los datos más relevantes de sus páginas web.

Tras ello, surgió la aparición de cierta disciplina que supondría un gran cambio de mejora para la analítica web, el termino Big Data.

Enrique Martín y Rafael Caballero, definen Big Data como “datos masivos que cumplen tres características, las conocidas 3-V (volumen, velocidad y variedad)” [3]

La existencia del Big Data supuso para la analítica web, por tanto, un aumento de la velocidad en la recogida de los datos, así como la posibilidad de generar estudios de cantidades masivas de datos que, sin esta disciplina, no hubiera sido viable.

Otro de los programas del cual se ha beneficiado la analítica web es la inteligencia artificial. La definición que ofrece Margaret A. Boden sobre la presente disciplina es aquella “que tiene por objeto que los ordenadores hagan la misma clase de cosas que puede hacer la mente.” [4]

Destacando la anterior definición, la inteligencia artificial tiene una gran influencia en la analítica web, debido a la posibilidad que ofrece a las personas que manejan webs, de una mayor facilidad de lograr las estrategias u objetivos marcados de manera más sencilla y eficaz para sus mercados online, ya que debido al que el trabajo es realizado por ordenadores o máquinas y por lo tanto, permiten una evaluación continuada de los datos recogidos sobre los elementos fundamentales de un negocio.

Una vez dados a conocer los acontecimientos más relevantes de la evolución de la Analítica Web, la cual, ha generado una gran necesidad para las empresas o negocios por el hecho de mantener un seguimiento completo sobre sus páginas web comerciales, es de mera importancia conocer que es, y cuál es la finalidad de la Analítica Web.

Según Sergio Maldonado, la Analítica Web es la “disciplina que persigue acciones de monitorización y mejora para la consecución de los objetivos que han fundamentado las inversiones y actividades online de la empresa”.[5]

Además, Ronan Chardonneau expone como definición de Analítica Web que "trata del elemento clave de cualquier estrategia de marketing en línea ya que permite calcular el rendimiento de la inversión realizada." [6]

Por otro lado, Maribel Morales Martínez sostiene que "las organizaciones utilizan la Analítica Web para distintos objetivos, como son: la presentación de informes sobre las cifras de tráfico del sitio web, comprender el comportamiento del cliente/usuario, gestionar campañas de marketing…".[7]

Analizando lo que mencionan estos académicos, es evidente la gran envergadura en la que se halla actualmente el manejo de los datos pertenecientes a las webs comerciales que deseen lograr sus objetivos.

3.2. GOOGLE ANALYTICS

Para medir los datos de un negocio Web, se encuentran diversas soluciones. La herramienta analítica más utilizada es Google Analytics, lanzada por la empresa Google. En el libro de Ronan Chardonneau, se muestra el siguiente concepto sobre Google Analytics "Tipo de solución de la analítica web que permite principalmente, analizar el comportamiento de los visitantes en un sitio web, reconocer los posibles problemas del sitio y ser advertido de ellos, calcular el rendimiento de la inversión de las acciones de marketing web y conocer los elementos que deben corregirse para optimizar continuamente el sitio y la experiencia de los visitantes".[8]

Es por ello que Google Analitics se considera una de las mejores herramientas de estudio de webs, debido a que además de ser gratuita, produce un alto potencial sobre el comportamiento de la audiencia que accede a cierta página web, así como de la propia aplicación. Los informes proporcionados por Google Analytics permiten beneficiarse a los miembros de la misma, del control y mejora de los resultados obtenidos en su localización web. Por lo tanto, los individuos asociados a Google Analytics poseen el privilegio de conseguir el éxito en la mundología del marketing digital o lo que es lo mismo, el conjunto de estrategias que toma un determinado negocio.

3.3 TÉCNICAS MACHINE LEARNING

En el presente informe han sido empleadas diferentes técnicas de machine learning, definido dicho termino por Russo, Claudia, Ramón, Hugo D., Alonso, Nicolás, Cicerchia, Lucas Benjamin , Esnaola, Leonardo, Tessore, Juan Pablo como “el área de la inteligencia artificial que engloba un conjunto de técnicas que hacen posible el aprendizaje automático a través del entrenamiento con grandes volúmenes de datos”.[9]

El estudio emprende con la elaboración de un análisis descriptivo tanto gráfico como numérico de todas las variables de interés pertenecientes a la base de datos para conocer mediante técnicas estadísticas, la existencia de patrones o relaciones entre las características del estudio.

Tras analizar dichos patrones, es conveniente realizar el análisis de componentes principales. Este último, beneficia a la futura aplicación de otras técnicas, ya que como menciona Jorge Galbiati R, “tiene por objeto reducir la dimensionalidad de un problema de múltiples variables, aplicando una sucesión de transformaciones lineales a las variables, de modo que un subconjunto de ellas concentre la mayor parte de la variabilidad contenida en las variables originales”.[10]

Para ejecutar dicho análisis es necesario que las variables sean numéricas y que las mismas, estén correladas entre sí. La primera componente principal, se obtiene maximizando su varianza, teniendo en cuenta la restricción de que la suma de sus pesos ha de ser igual a 1. Dicho problema de maximización se resuelve mediante los multiplicadores de Lagrange, aunque en el presente estudio se resuelve haciendo uso del programa dedicado a la computación estadística RStudio.

$$\text{Maximizar } V(Z1) = u' Vu \text{ s.a restricción } \sum u1\, u1 = 1$$

A continuación, y con el uso de las Componentes Principales extraídas anteriormente, se lleva a cabo el análisis Clúster, el cual permite clasificar en grupos homogéneos las observaciones de cada una de las variables de la base de datos, así se disminuye la dificultad para la elaboración de las futuras técnicas de Machine Learning. El análisis Clúster, se caracteriza según José Luis Vicente

Villardón por ser “una técnica de Análisis Exploratorio de Datos para resolver problemas de clasificación. Su objeto consiste en ordenar objetos en grupos de forma que el grado de asociación/similitud entre miembros del mismo clúster sea más fuerte”. [11]

Dicha técnica, se ha llevado a cabo mediante el algoritmo de K-medias, el cual, radica en la elección inicial de k puntos de la base de datos, e ir asociando el resto de puntos a las k más cercana, y a continuación, volver a calcular la nueva k. La operación se repite hasta alcanzar aquella k, en la que no surja ninguna variación.

Después de agrupar las observaciones en categorías homogéneas, se procede a generar una nueva clasificación de los datos, mediante el algoritmo K-Nearest-Neighbour, definido por José Luis S.A como “el método que, dado un individuo a clasificar, se localiza en los datos de muestra o entrenamiento (cuya clasificación se conoce) los k individuos más cercanos al individuo a clasificar”. [12]

Por tanto, la aplicabilidad de dicho algoritmo, mediante la división de las observaciones de la base de datos en el 50% de ellos, como datos de entrenamiento y el resto, como datos a clasificar, tiene como finalidad alcanzar un nivel de fiabilidad mediante dicha clasificación.

Para incluir las variables categóricas en el análisis, se emplea la técnica Forward, también conocida como la selección de variables hacia delante. Para ello, se han modificado las variables categóricas de manera que únicamente tomen los valores 0 y 1, indicando así, la ausencia de una cualidad. Este tipo de variables adquieren el nombre de variables dummy. Una vez generadas, se incluyen de manera secuencial en la base de datos compuesta por las Componentes Principales, generada anteriormente. A continuación, se prueba la solución del método K-Nearest Neigbour y aquellas dummy que aumenten el porcentaje de acierto del método de clasificación, se mantiene en el análisis.

Por último, se realiza el método de Support Vector Machine. Dicho método es definido por I Barbona y C Beltrán como “método de clasificación supervisada que permite determinar la frontera óptima entre dos grupos que pueden ser linealmente separables o no”. [13]

Dicho método puede aplicarse mediante el uso de diferentes distancias. En el presente estudio se hace uso de dos de ellas, la distancia de Manhattan y la distancia cuadrática. La distancia de Manhattan se define como [14]:

$$d_{L_1}(x, y) = \sum_i |x_i - y_i|$$

Mientras que la distancia cuadrática se calcula de la siguiente manera [15]:

$$d_t^2(\mathbf{x}) = d_t^2(\mathbf{x}) + \ln \left| \mathbf{S}_t \right| - 2 \ln (q_t)$$

Donde, st es la matriz de covarianzas en el grupo t y qt es la probabilidad previa del grupo t o lo que es lo mismo, la división entre el número de observaciones en el grupo t entre el total de observaciones.

El modelo se resuelve generando un algoritmo el cual resuelva un problema de optimización, donde el hiperplano satisfactorio formulado, se efectúa de la siguiente manera:

$$W x - b$$

Y, por tanto, el modelo a resolver es el expuesto a continuación:

Min || w ||k

s.a:

yi(∑ wj*xij + b) ≥ 1, para todo i=1,…,n

yi = {-1,1} para todo i=1,…n

Tras el algoritmo aplicado manualmente, se utiliza la función svm proporcionada por la interfaz Rstudio.

4. OBJETIVOS

El objetivo de este informe, se basa en lograr un algoritmo que permita clasificar los datos de estudio, de tal manera que, sea lo más fiable posible para ofrecer a la empresa un modelo que pueda predecir las futuras decisiones de sus clientes, en el ámbito online.

Para obtener los objetivos marcados, se procede a continuación, a realizar un estudio previo de las características más importantes de las variables que forman la base de datos. Tras ello, se utiliza un método que permita reducir la dimensionalidad de la base de datos, para que, seguidamente, se desarrolle con mayor facilidad la clasificación de las observaciones en grupos homogéneos caracterizados por contener observaciones con las mismas particularidades, y por consiguiente, desarrollar un algoritmo que pueda clasificar los datos de la forma más optima, logrando así, una alta fiabilidad en la clasificación de compras por parte de los clientes en la página web.

5. EXPERIENCIA COMPUTACIONAL

5.1 METODOLOGIA

El presente informe ha sido desarrollado gracias a la utilización del lenguaje de programación R, bajo la interfaz RStudio. Dicho programa se fundamenta según Joseph J. “en la aplicación de aquellas técnicas computacionales estadísticas, tanto numéricas como gráficas, que permiten extender grandes y elaborados estudios de datos de cualquier tipo de ámbito”.[16]

El programa cuenta con cuatro pantallas dedicadas cada una de ellas a diferentes funciones.

Una de ellas, es la que permite crear algoritmos y códigos que a continuación pueden ser ejecutados en la consola, para conocer los resultados de dichas elaboraciones. Además, el programa cuenta con otra pantalla donde se exponen

los gráficos que se generen. Y, por último, la pantalla de la memoria, guarda todos los resultados ejecutados.

RStudio tiene implementadas una elevada cantidad de funciones que pueden facilitar estudios concretos. Para generar el presente estudio, aplicando técnicas de machine learning, han sido de interés alguna de ellas, como es el caso de *prcomp*, función que ofrece un análisis de Componentes Principales, exponiendo como cada variable queda explicada por cada Componente Principal, *dummy_cols,* que permite convertir las variables categóricas en lo que se conoce como variables dummy, es decir, toman valores de 0 y 1, *kmeans* que favorece a la clasificación de las observaciones de una base de datos, según características comunes que compartan con otras de ellas, *knn* función que desarrolla la clasificación de los datos en 0 y 1, y por último la función *svm,* que produce una clasificación predictiva de los datos.

5.2 RECOPILACION DE INFORMACIÓN

El conjunto de datos perteneciente a una empresa, la cual, ha impulsado su negocio al mundo online, contiene un total de 10 variables numéricas y 8 variables categóricas. Dicha base de datos ha sido recogida del repositorio UCI, donde se facilitan una gran variedad de conjuntos de datos para ser tratados [17].

En primer lugar, se cuenta con las diferentes páginas a las que un cliente puede acceder o iniciar sesión. Estas últimas se encuentran clasificadas en tres categorías, de carácter administrativo*,* informativo o páginas relacionadas con el producto que se vende*.* Asimismo, se han recogido los tiempos, los cuales se han considerado los segundos invertidos en el periodo de un año, de cada usuario que ha iniciado sesión en las diferentes páginas web definidas anteriormente, denotadas como duración Administrativa, duración Informativa y duración producto.

Cabe destacar que los valores de las distintas categorías proceden de la URL de la página, siendo esta última actualizada a tiempo real, de manera que, al

cambiar de una página web a otra, se contabiliza el acceso del usuario a cada una de las mismas.

A continuación, se han recopilado ciertos datos de interés mediante Google Analytics.

De las funciones implementadas en dicha plataforma, la base de datos a tratar contiene la Tasa de Rebote, Tasa de salidas y el valor de la página.

El valor de la tasa de rebote consiste en el porcentaje que se genera, cuando un usuario inicia sesión en una única página web y tras ello, no realiza ninguna otra solicitud en esa misma sesión, es decir, no accede a ninguna otra página. Un porcentaje de rebote elevado podría implicar un dato negativo o positivo, ya que depende de la disposición de la página web de estudio. Si esta última, contiene una página inicial desde donde se accede al resto de páginas, por ejemplo, a la página de productos, la página de información del propio comercio o la página donde se ejecuta la transacción sí supondría un dato negativo un alto porcentaje de rebote. El presente porcentaje se calcula dividiendo el número de sesiones iniciadas una de las páginas web, en el caso concreto de estudio, en la página Administrativa, Informativa o la relacionada con el producto, entre el total de accesos realizados a todas las páginas web en un momento determinado.[18]

Por otro lado, la tasa de salidas recoge los valores en los que la visita a una página web, ha sido la última de la sesión de cierto usuario. Dicho porcentaje se calcula realizando la división del último número de página que ha sido el fin de la sesión, entre el total de visitas a la página web. [19]

Las visitas de un usuario a una página web, las transacciones que realicen o no, y como consecuencia, los ingresos que se generen, así como los objetivos específicos del comercio electrónico, permiten realizar el cálculo del valor que obtiene la página web. El valor de la página representa el promedio de una visita a la misma, antes de completar una compra en cierto establecimiento electrónico. Por lo tanto, el objetivo de dicho valor consiste en proporcionar al usuario de Google Analytics, cuál de sus distintas páginas del sitio web, genera una cantidad superior de ingresos. Por lo tanto, el cálculo del valor de la página, consiste en la división entre la suma de los ingresos generados más el valor total

objetivo generado por el propietario de la web entre el número de páginas totales visitadas por el usuario de manera única.[20]

Por otro lado, la variable Día especial expresa lo cercano que se encuentra el inicio de sesión en la página web por parte del cliente, ha un día especial como podría ser Navidades. La importancia de dicha variable, se encuentra en que, si el final de la transacción de un usuario termina siendo una compra, puede deberse a la proximidad de un día especial. El valor de este atributo se determina considerando la dinámica del comercio electrónico como la duración entre la fecha del pedido y la fecha de entrega.

A continuación, se procede a la explicación de las variables categóricas de interés que contiene la base de datos de estudio.

Se recogen los datos sobre el usuario referentes al mes en el que se realizó el inicio de sesión en cierta página web, si el visitante accede a la misma un fin de semana o entre semana, la operación en el sistema, el navegador utilizado, la región desde donde se realiza el inicio de sesión, el tipo de gráfico. De las cuatro últimas variables mencionadas, se excluyen los nombres reales de cada categoría, denotándolas por cantidades numéricas, como seguridad de protección de cada individuo que ha proporcionado los datos.

Como características del cliente “VisitorType”, se destaca, si el mismo, es un visitante de la página web recurrente o si, en cambio, es la primera vez que accede.

Por último, se cuenta con dos variables tipo booleana (TRUE o FALSE).

La primera de ellas denotada como “fin de semana”, informa de si el inicio de sesión se ha realizado entre semana o un fin de semana.

La segunda, denotada por “compra”, la cual aporta la información necesaria de si un cliente realiza una transacción al iniciar sesión. Dicha variable se considera de trato especial, pues en el presente informe se efectúa la predicción de dichos ingresos en función del resto de las características de cada cliente que accede a la página web mencionadas anteriormente.

5.3 RESULTADOS

5.3.1 ANÁLISIS DESCRIPTIVO

La siguiente tabla representa los principales parámetros estadísticos como son la media, varianza o mediana de las variables numéricas de estudio, para conocer las características principales que contienen cada una de las variables numéricas.

Tabla 1. Parámetros I

	Administrativa	***Duración Administrativa***	***Informativa***	***Duración Informativa***
Media	2.32	80.82	0.50	34.47
Varianza	11.03	31250.85	1.61	19810.36
Mínimo	0.00	0.00	0.00	0.00
Mediana	1.00	7.50	0.00	0.00
Máximo	27.00	3398.75	24.00	2549.38

Tabla 2. Parámetros II

	Producto	***Duración Producto***	***Porcentaje Rebote***	***Tasa de Salidas***	***Valor Página***	***Dia Especial***
Media	31.73	1194.75	0.02	0.04	5.89	0.06
Varianza	1978.07	3662130.14	0.00	0.00	344.79	0.04
Mínimo	0.00	0.00	0.00	0.00	0.00	0.00
Mediana	18.00	598.94	0.00	0.03	0.00	0.00
Máximo	705.00	63973.52	0.20	0.20	361.76	1.00

Resaltando el valor de la media y de la mediana en *las tablas 1 y 2*, cabe destacar que de las variables de cada categoría de la página web, así como la duracion del usuario dentro de las mismas, se aprecia la considerable diferencia entre el valor máximo y minimo. Este hecho informa de que estas variables probablemente no sigan una distribucion normal, y es por ello que el parámetro

que expone de mejor manera la tendencia de los valores recogidos sea la mediana.

De las tres páginas que contiene la web, el 50% de los usuarios que inician sesión en la misma, acceden mayoritariamente a la página del producto, con una duración aproximada de 598.94 segundos al año, seguida de la página de carácter Administrativo, con una duración entorno a los 7.5 segundos al año y por último, la de carácter informativo. De la página de la categoria informativa, cabe destacar que la duración ejercicida por el 50% de los clientes, es de 0. Este hecho indica que esta categoria es la menos deseada por los clientes que han iniciado sesion durante un año.

En cuanto al porcentaje de rebote y tasas de salidas, se analiza que ambas variables localizan los valores recogidos entre 0 y 0.2. Que el valor máximo de dichas variables sea 0.2 es un dato excelente sobre la página web de estudio. Esto se debe a que el porcentaje de rebotes y la tasa de salidas tiene un valor infimo y por tanto, los clientes que acceden a la web del comercio estan interesados en la misma. Cabe destacar que la media de ambas variables son prácticamente 0, fomentando lo indícado anteriormente.

En lo que respecta a la variable valor de la página, sus valores toman valores muy extremos. Es por ello que la mediana será el parámetro que mejor indique entorno a que valor medio se encuentra valorada la página web de estudio. De los usuarios que inician sesion, el 50% de ellos considera que el valor adquirido a la página web visitada es de aproximadamente 344.

Por último, los parámetros expuestos para la variable día especial, indican que probablemente dicha variable siga una distribución normal. Esto se debe a que, eliminando el valor más repetido que es 0, el 50% de los datos se encuentra en 0.6 por lo que gráficamente se podrá observar una clara distribucion normal de los datos.

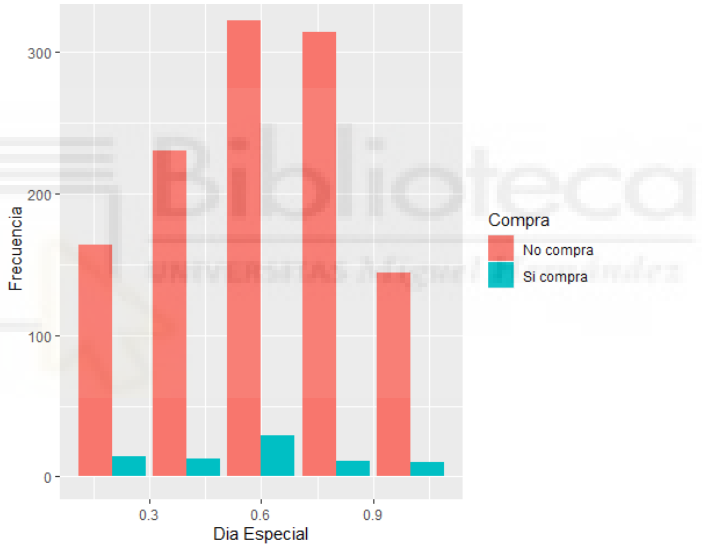
A continuación mediante la interfaz RStudio, se realiza un análisis gráfico de las variables númericas, que permite ampliar la información de dichas variables obtenidas anteriormente con los parámetros fundamentales de cada una de ellas.

ANALISIS GRAFICO DIA ESPECIAL

Cabe destacar, que la variable dia especial parece tener un comportamiento diferente frente al resto de variables, ya que númericarmente se había detectado que la variable posiblemente siguiera una distribucion normal de sus datos. Es por ello, que se procede a realizar un estudio individualizado y teniendo en cuenta como afecta que sea o no un día especial a la decisión de comprar o no que tiene el usuario. Para ello, se han eliminado los datos que tomaban valor 0, para adquirir una mejor visualizacion del comportamiento de dicha variable.

Gráfico 1. Histograma de la decisión de compra según la cercanía de un Día Especial.

Fuente: Elaboración propia.



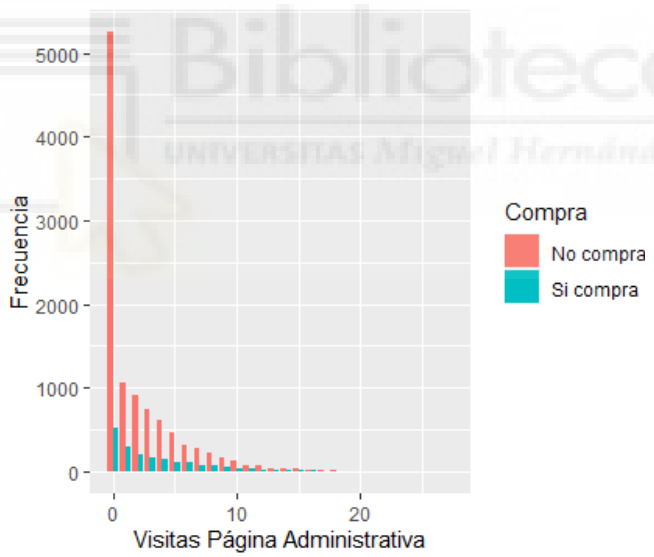
El *gráfico 1*, muestra la influencia que tiene iniciar sesión en la página web un día cercano a una festividad y sí dicho usuario decide comprar o no algún producto. Analizando un resumen de los datos que conforman la variable día especial, se observa como las visitas a la página web son más comunes cuando la fecha de inicio de sesión en la misma no es próxima un día especial. Es por ello, que en la representación gráfica se han obviado todos los inicios de sesión por parte de los clientes que no acceden a la página web en una temporada cercana a un día especial, para analizar de mejor manera el comportamiento de los datos que afectan los inicios de sesión por parte de los clientes cerca de un

día especial. Implica, por tanto, que se representa un conjunto de datos asimétricos, en concreto hacia la derecha, o que pueden no estar distribuidos normalmente. Además, se cuenta con valor atípico en el extremo derecho del eje de abscisas, que podría asumirse como una causa especial o un error en la toma de observaciones

ANÁLISIS GRÁFICO CATEGORIAS DE LA PAGINAS WEB

Tras el estudio de la variable Dia Especial, se han representado mediante un histograma y un gráfico boxplot la frecuencia de acceso de los clientes a las distintas páginas que contiene la web comercial frente a la decisión final que toma el usuario que ha iniciado sesión en las mismas, de ejecutar una transacción o no.

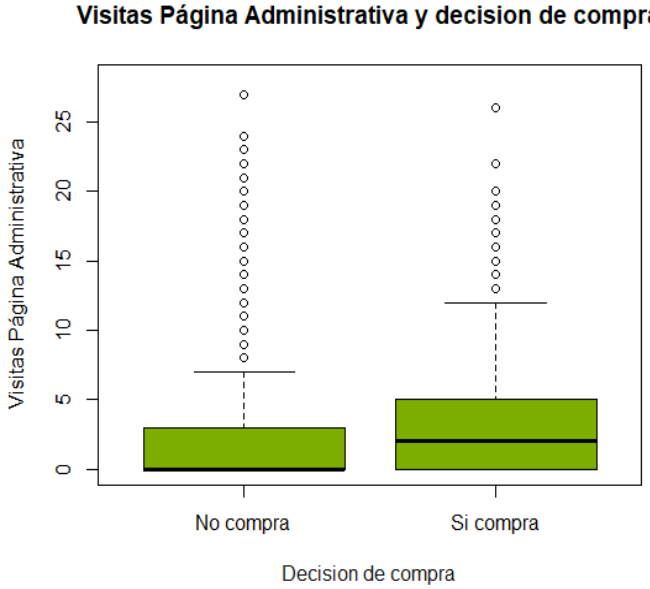
Gráfico 2.1 Histograma Página Administrativa. *Fuente: Elaboración propia.*



Analizando el *gráfico* 2.1 se destaca que la distribución de las visitas que los usuarios realizan a la página administrativa no es una distribución normal, corroborando así lo anteriormente dicho según los parámetros principales de la presente variable. Además se observa cierta asimetría hacia la derecha de los datos, siendo destacable que de los 12.330 usuarios que inician sesion en la web comercial de estudio, 5254 usuarios no acceden a la categoría administrativa.

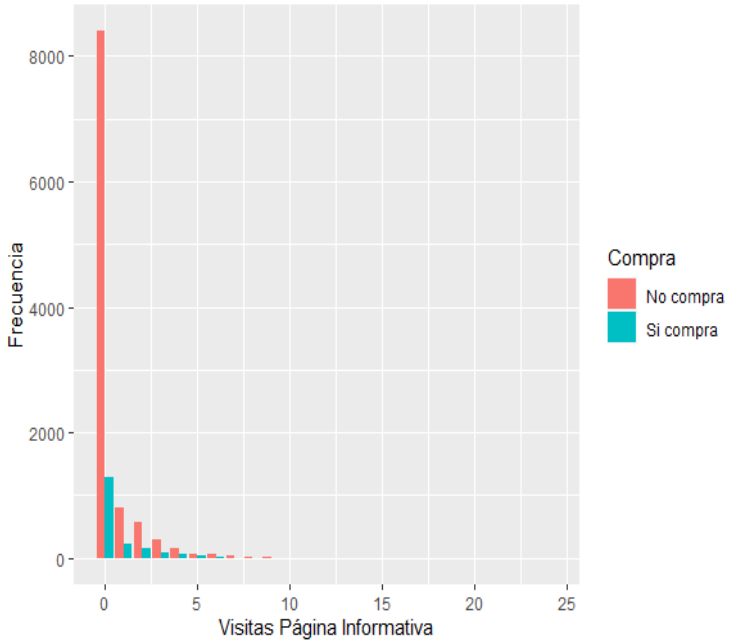
Por otro lado, es de mera importancia destacar la exístencia de una tendencia mayoritaria de que los usuarios no comercializan tras la visita a dicha página .

Gráfico 2.2 BoxPlot Página Administrativa. *Fuente: Elaboración propia.*



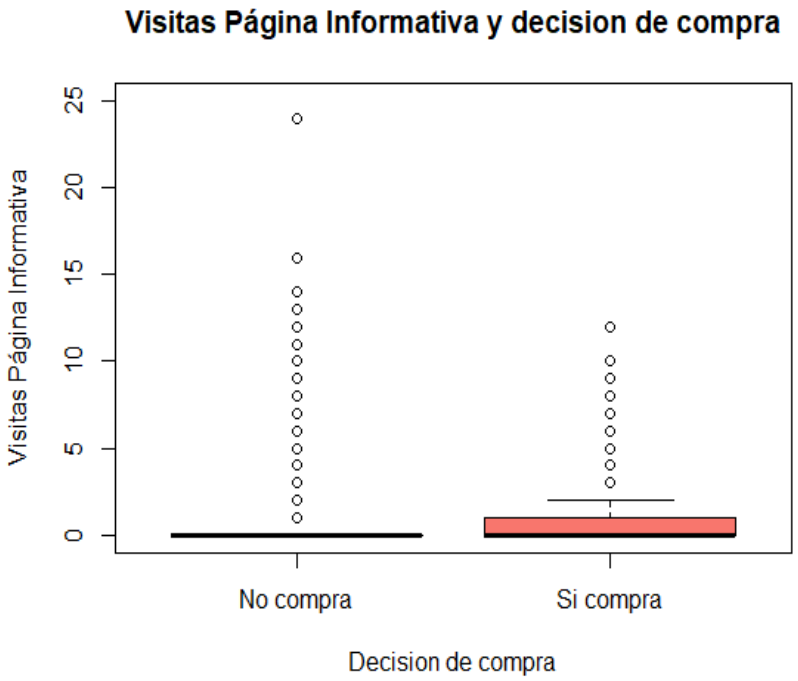
El *gráfico 2.2* expone algunos de los parámetros significativos referentes a la variable de estudio, visitas de la página Administrativa. Los pesos medios entre ambos grupos,es decir, aquellos usuarios que compran tras acceder a la presente página y los que no lo hacen, son similares, ya que ambos se encuentran entorno al valor 0, pero existen diferencias de variabilidad entre ambos, o lo que es lo mismo, se localizan datos con carácteristicas comunes según su naturales, existiendo así una relación entre los usuarios que compran tras la visita a la página administrativa y los que no lo hacen de semejanza y uniformidad. Además, se observa la existencia de valores atípicos en ambos grupos.

Gráfico 3.1 Histograma Página Informativa. *Fuente: Elaboración propia.*



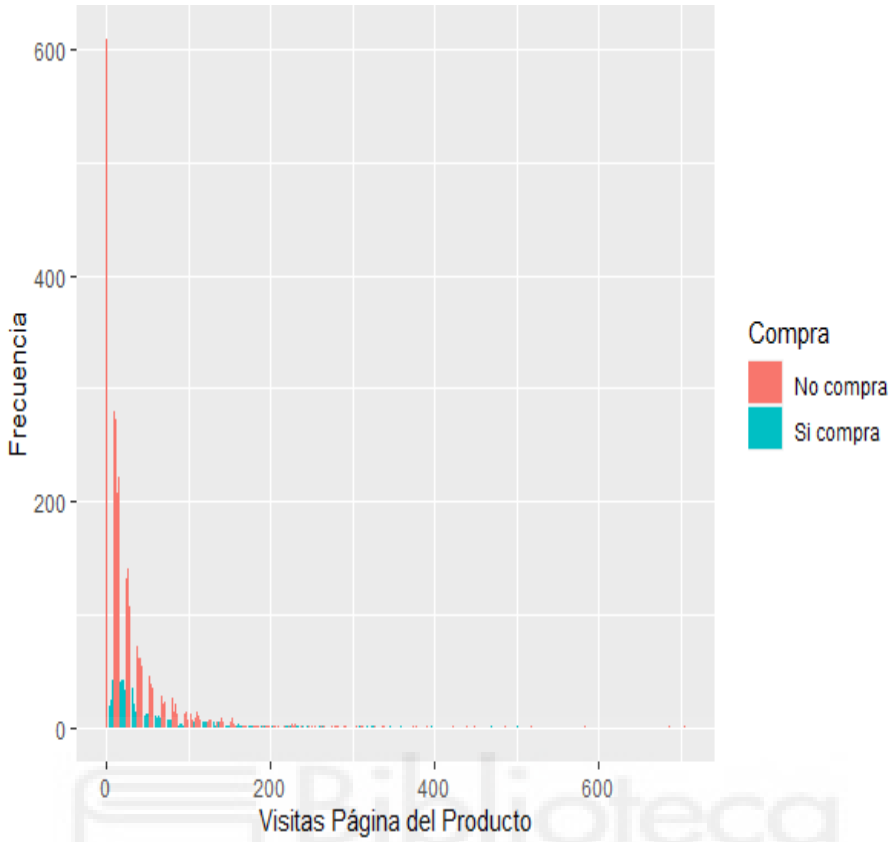
El *gráfico 3.1* reprensenta la frecuencia con la que los usuarios inician sesión en la página informativa de la web comercial. Se observan caracteristicas similares a las analizadas en el *gráfico 1.1*, pues existe una asimetria hacia la derecha de los datos, por lo que se evidencia que el valor que más reincide es 0, siendo un total de 8404 clientes los que no acceden a la página informativa. Además, se observa que de los usuarios que acceden a la presente página, la mayoría de ellos, cierra sesión sin hacer niguna compra comercial.

Gráfico 3.2 BoxPlot Página Informativa. *Fuente: Elaboración propia.*



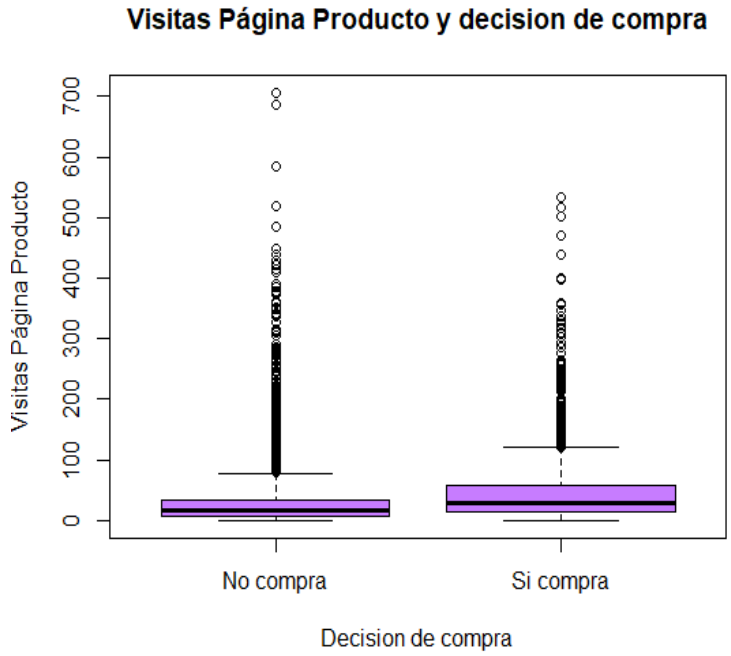
El *gráfico 3.2* expone la existencia de valores atípicos, así como la existencia de que ambas clasificaciones de visitas a la página informativa, contienen datos homogeneos y por lo tanto, existe variabilidad entre ellos. Cabe destacar ambos grupos son similares en media, como se había manifestado anteriormente en la *tabla 1.1* parámetros descriptivos numéricos.

Gráfico 4.1 Histograma PáginaProducto. *Fuente: Elaboración propia.*



La distribución de los datos que representan la frecuencia con la que los visitantes inician sesión en la página relacionada con el producto, no es normal. Además se presenta una asimetria en los datos. Cabe destacar, que a diferencia de las páginas administrativa y informativa, el rango de frecuencias en la que se repiten las visitas a la página relacionada con el producto es inferior, significando que la página más visitada es la relacionada con el producto. Todo lo mencionado anteriormente se presenta en el *gráfico 4.1*.

Gráfico 4.2 Histograma Página Producto. *Fuente: Elaboración propia.*



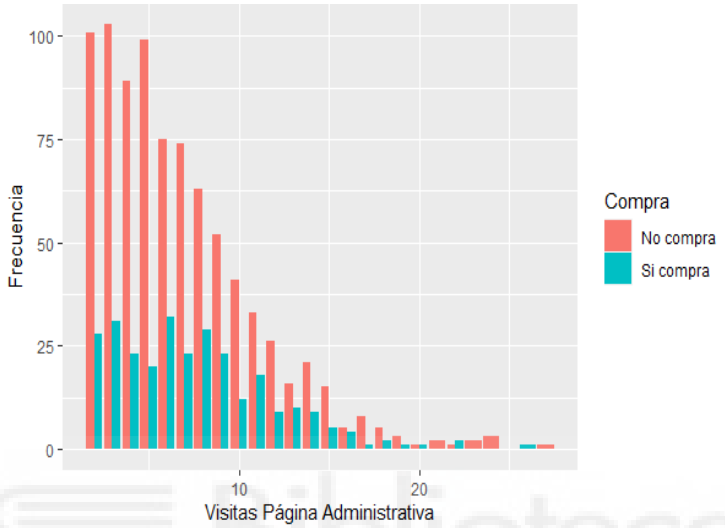
Por otro lado, el *grafico 4.2* expone que, los grupos asociados a la compra de producto, en la presente página, muestra variabilidad inferior a la que aparecia en las paginas administrativas e informativas. También se destaca la existencia de valores atípicos.

ANALISIS GRAFICO WEB/COMPRAS DE LOS CLIENTES INTERESADOS

Debido al comportamiento observado en los gráficos expuestos anteriormente referidos a los datos que conforman las variables página Administrativa, informativa y página relacionada con el producto, el valor que más se repite es 0, lo que puede significar que los clientes no estaban interesados en la página del comercio. Es por ello, que a continuación, se va a realizar el mismo estudio descriptivo, pero únicamente para los clientes que se consideran interesados, es decir, se seleccionan aquellos clientes que acceden a las distintas categorías de la página más de una vez, así como que la duración de tiempo en las mismas páginas sea superior a 10 segundos. Dicho grupo cuenta con un total de 1230 usuarios que inician sesión en la página web en el periodo de un año. Realizando este nuevo conjunto de datos, se podrá apreciar con mayor exactitud que

caráteristicas de los clientes o del inicio de sesión afectan a que finalmente el usuario se decida a llevar a cabo una transaccion comercial en la web o no.

Gráfico 5. Distribución página Administrativa. *Fuente: Elaboración propia.*



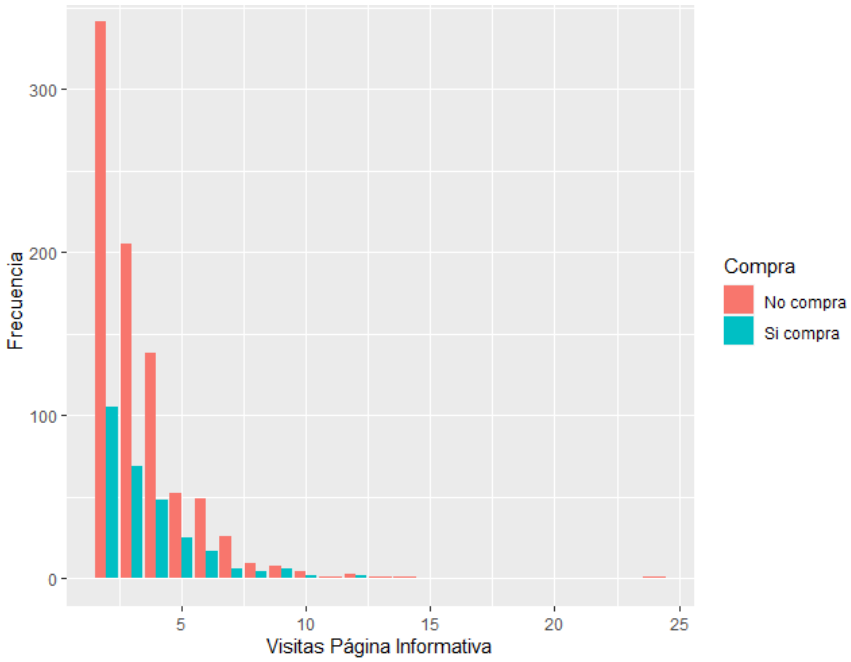
En el *gráfico 5*, se muestra, la distribución y frecuencia con que los clientes que se han considerado interesados en la página comercial de estudio, acceden a la página administrativa y cuales de ellos, terminan realizando una compra en el negocio.

A diferencia del *gráfico 2.1,* se expone con mayor claridad la diferencia entre el grupo de clientes que acceden a la página administrativa que adquieren un artículo del presente negocio y los que no.

En el rango de frecuencia de 40 y 100, se localizan los usuarios que visitan la página administrativa de 0 a 10 veces en la misma sesión, mientras que en el rango de frecuencia de 0 a 18, se encuentran los clientes que realizan desde 10 hasta 27 visitas. Este último rango, se caracteriza por ser menos frecuentado que el primer rango descrito.

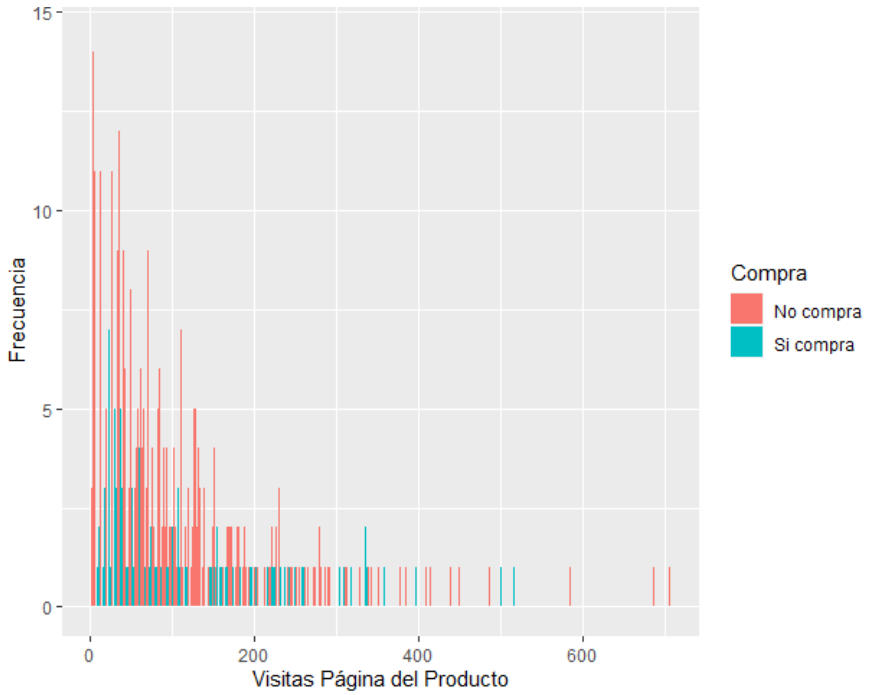
Es de mera importancia destacar que las compras que se realizan tras acceder a la página administrativa, son menos frecuentadas que los clientes que acceden a dicha página y no realizan trasacción final.

Gráfico 6. Distribución página Informativa. *Fuente: Elaboración propia.*



Analizando la comparativa entre *el grafico 6*, que expone los clientes que se han considerado como usuarios interesados en la web, en concreto los que acceden a la página informativa, con el *grafico 3.1*, que muestra todos los usuarios que inician sesion en la página de información, se manifiesta que la diferencia entre ambos gráficos es exiguo ya que la distribución de los gráficos es la misma y se observan de la misma manera los datos en ambos dos.

Gráfico 7. Distribución página relacionada con el producto. *Fuente: Elaboración propia.*



El *gráfico 7* alega la relación existente entre los usuarios que están interesados en la página relacionada con el producto, o lo que es lo mismo, aquellos que acceden al menos 1 vez y perdura en la misma más de 10 segundos, y finalmente quienes de ellos terminan la sesión en la actual página de estudio realizando una compra. Se detectan claros valores atípicos en aquellos usuarios que acceden a la página más de 450 visitas al año. Por otro lado, se manifiesta que la mayoría de los usuarios que acceden a dicha página, frecuentan un rango entre 0 y 200 visitas al año. Además, la mayoría de clientes que acceden a la página relacionada con el producto acaban no comprando nada, pero a diferencia de lo expuesto en las páginas administrativa e informativa, si se detectan más usuarios que compran algún artículo tras la visita en la página web.

ANALISIS GRÁFICOS TASA DE SALIDAS Y % DE REBOTE

Gráfico 8. Histograma Tasa de salidas. *Fuente: Elaboración propia.*

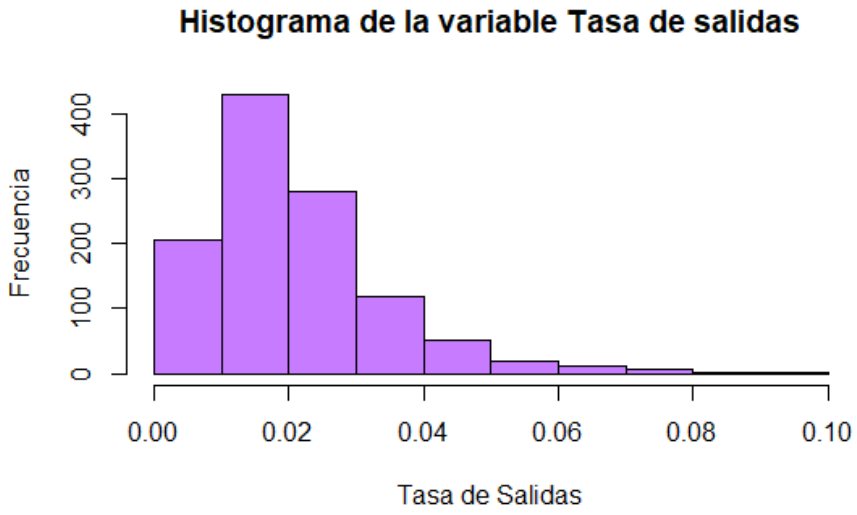


Gráfico 9. Histograma Porcentaje Rebote. *Fuente: Elaboración propia.*

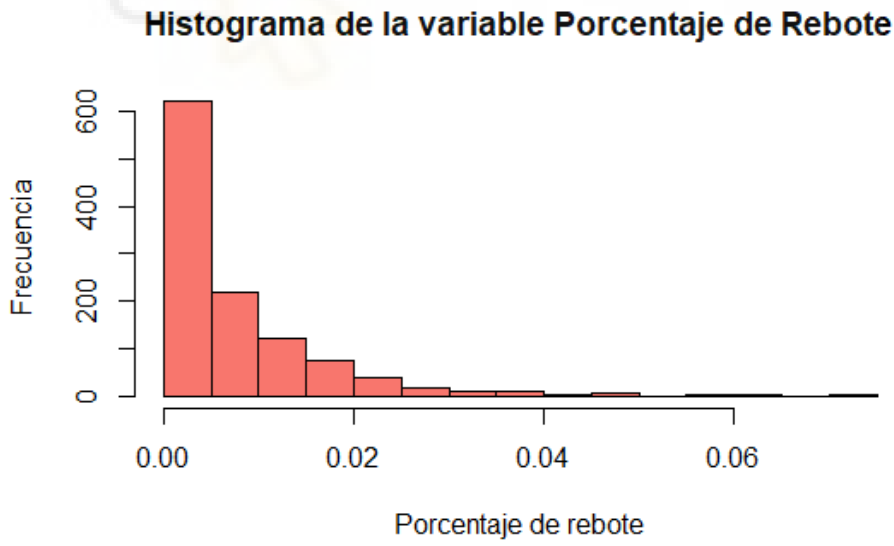
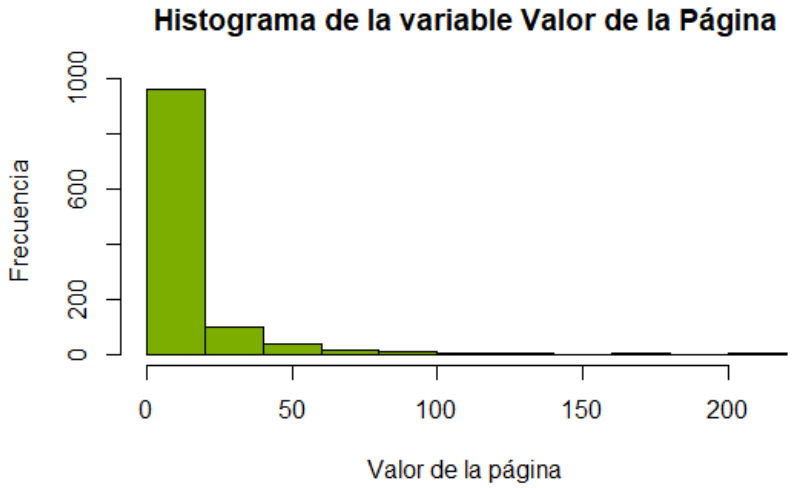


Gráfico 10. Distribución del Valor de la página web. *Fuente: Elaboración propia.*

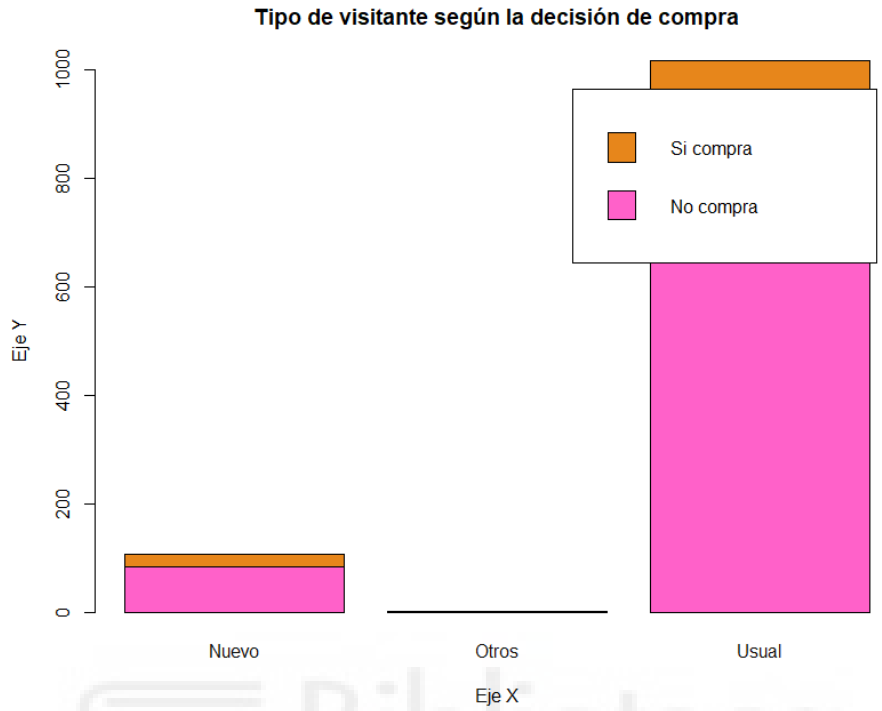


Los *gráficos 8 y 9* exponen los datos recogidos tanto de la tasa de Salidas como el Porcentaje obtenido de la página web de estudio. Ambas variables presentan un rango de sus datos entre 0 y 0.20, debido a que son porcentajes extraidos de la herramienta Google Analytics. Los dos gráficos manifiestan asimetria de los datos a la derecha. Los valores de porcentaje de rebotes, son más cercanos a 0 que a 1. Esto significa un valor positivo, pues de todos los usuarios que acceden a las distintas categorías de la página web, solo un porcentaje muy reducido de los visitantes, han generado el valor 1 página visita. Por otro lado, con lo que respecta a la variable Tasa de salidas, la exposición de la misma significa que de los usuarios que visitan las páginas, pocos de ellos salieron seguidamente del sitio web. Cabe destacar, que en ambos gráficos se analizan ciertos valores atípicos, en el extremo izquierdo de las representaciones.

El *gráfico 10*, representa la frecuencia de los valores obtenidos antes de realizar una transacción comercial en las distintas páginas web. El pico del histograma o lo que es lo mismo, el valor más repetido es de 0 euros. Lo que significa que la mayoría de los usuarios que acceden a la presente página web no realizan ninguna compra.

Gráfico 11. Gráfico de barras decision de compra según el tipo de visitante.

Fuente: Elaboración propia.

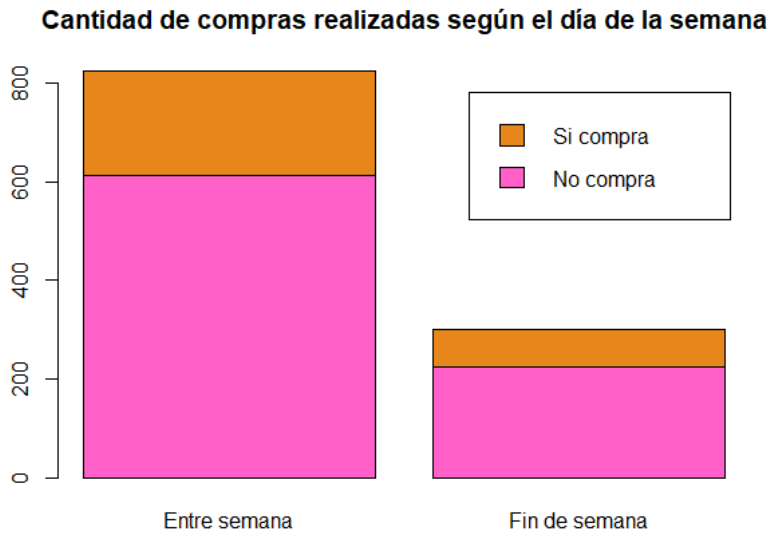


El *gráfico 11* recoge la cantidad de compras por cada tipo de cliente que visita la página. Se obtiene que los visitantes que son nuevos accediendo a la página web, aproximadamente, unos 100 de ellos, generan ingresos en la web, tomando la decisión de comprar algún producto. Por otro lado, en cuanto a los usuarios habituales en la tienda online, se expone que, 200 de los mismos, se deciden a compras cierto producto en el inicio de sesión en la web.

Por último, cabe destacar que, de los visitantes de la página de otras categorías distintas a las mencionadas anteriormente, son prácticamente inexistentes sus ingresos en la misma.

Gráfico 12. Gráfico de barras decision de compra según el dia de la semana.

Fuente: Elaboración propia.

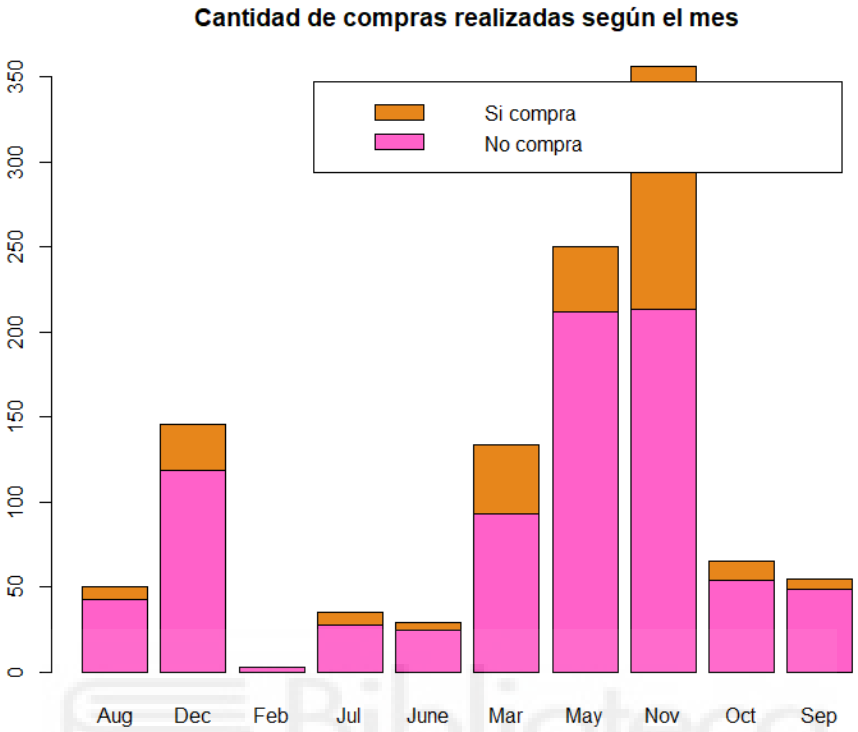


El *gráfico 12* expone la cantidad de compra de aquellos usuarios que inician sesión, teniendo en cuenta si el acceso del cliente a la página web, ha tenido lugar un día entre semana o, por lo contrario, fin de semana.

Se examina que, de los 1123 inicios de sesión registrados en la base de datos, 200 clientes adquieren un producto antes de finalizar dicha sesión en un día entre semana, mientras que un total de aproximadamente 90 clientes, tramitan dicha compra un fin de semana.

Gráfico 13. Gráfico de barras decision de compra según el mes.

Fuente: Elaboración propia.



A continuación, se analiza la dependencia existente entre las compras que se generan en la web online según el mes en el que el cliente inicia la sesión en la misma.

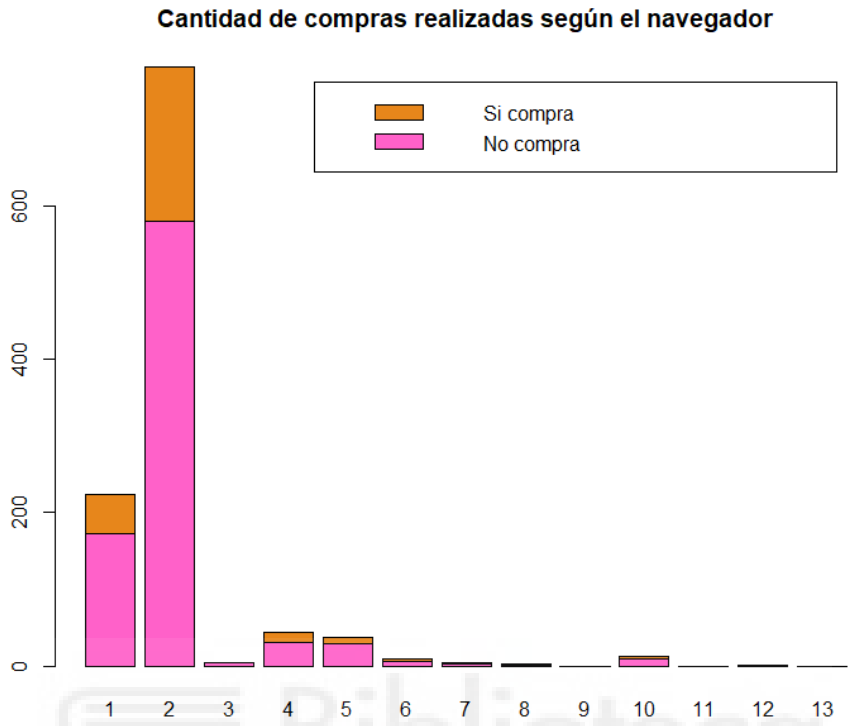
En el *grafico 13* se muestra que el mes donde mayor cantidad de compras se realizan es en noviembre con un total de 143 compras en un año, seguido de marzo y a continuación mayo. En el mes de febrero, cabe destacar que ningún cliente ha comprado ningún producto tras el acceso a la tienda online.

Por otro lado, es primordial destacar que el mes de noviembre, también cuenta con el mayor número de clientes que no han comprado tras el inicio de sesión.

Estos datos, podrían deberse a cualquier oferta que ofrezca el comercio en el mes de noviembre, logrando que la entrada de clientes en su web, sea superior a la del resto de meses del año de estudio.

Gráfico 14. Gráfico de barras cantidad de compras según el tipo de ordenador.

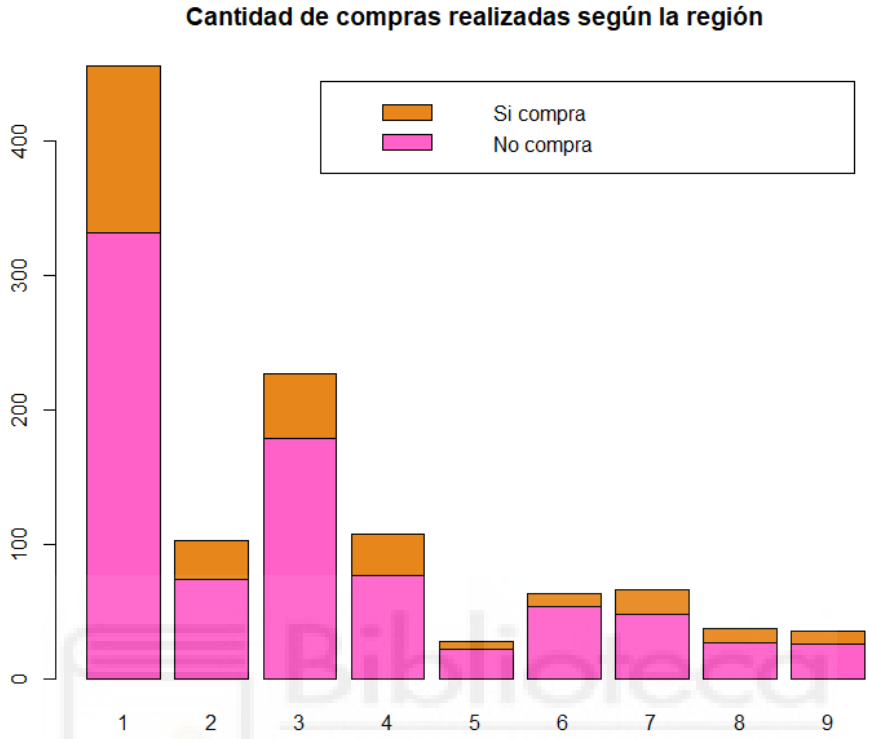
Fuente: Elaboración propia.



En el *gráfico 14* se analiza la influencia del navegador desde donde se inicia la sesión sobre las compras en la web comercial. El navegador 2 de acceso es el que genera más compras en la página del negocio. Por otro lado, se analizan ciertos navegadores desde los cuales prácticamente ningún cliente inicia sesión. Estos son el navegador 9, 11, 12 y 13.

Gráfico 15. Gráfico de barras cantidad de compras según la región.

Fuente: Elaboración propia.



En el presente *gráfico 15,* se evidencian aquellas regiones desde donde inician sesión los clientes, que generan mayores cuantías de compras, así como de modo general, la cantidad total de usuarios de cada región que acceden al comercio.

Analizando el gráfico, se presenta que la región uno, es aquella desde la que más usuarios acceden en la página web. Además, también es la región desde la que más compras se generan.

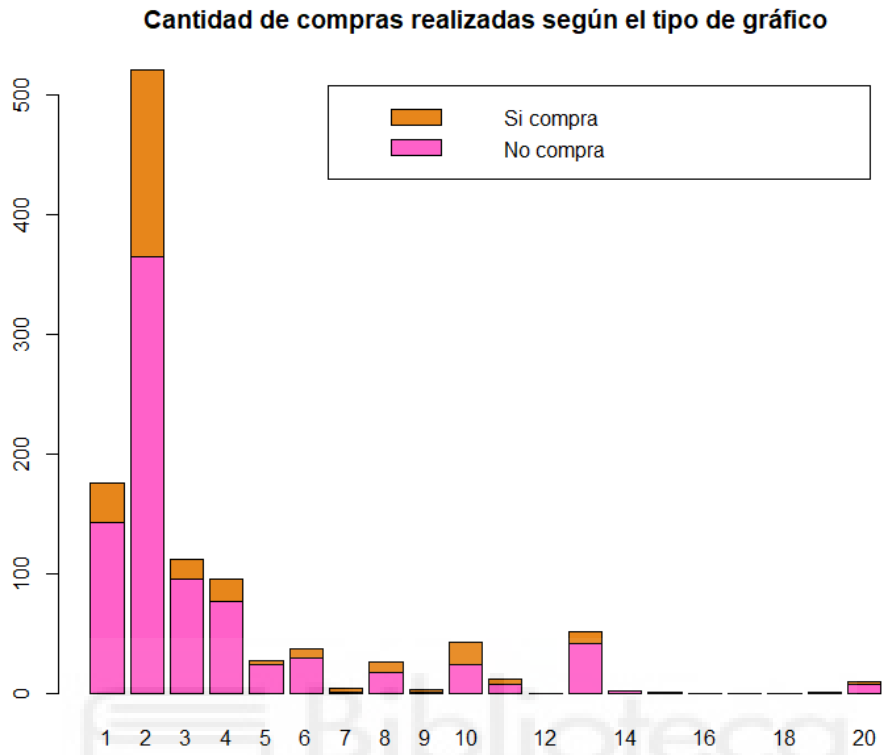
A continuación de la región uno, es destacable la cantidad de accesos desde la región tres, así como el total de usuarios que inician sesión desde la misma.

De todas las regiones, la que menos compras genera es la cinco. Esto podría deberse a que excepto la región uno y tres, el resto de regiones sean zonas económicamente dañadas y por ello, la cantidad de compras sea inferior.

Otro motivo podría ser, la falta de un desarrollo tecnológico, que impida o limite a las personas pertenecientes a esa región, acceder a webs comerciales.

Gráfico 16. Gráfico de barras cantidad de compras según el tipo de gráfico.

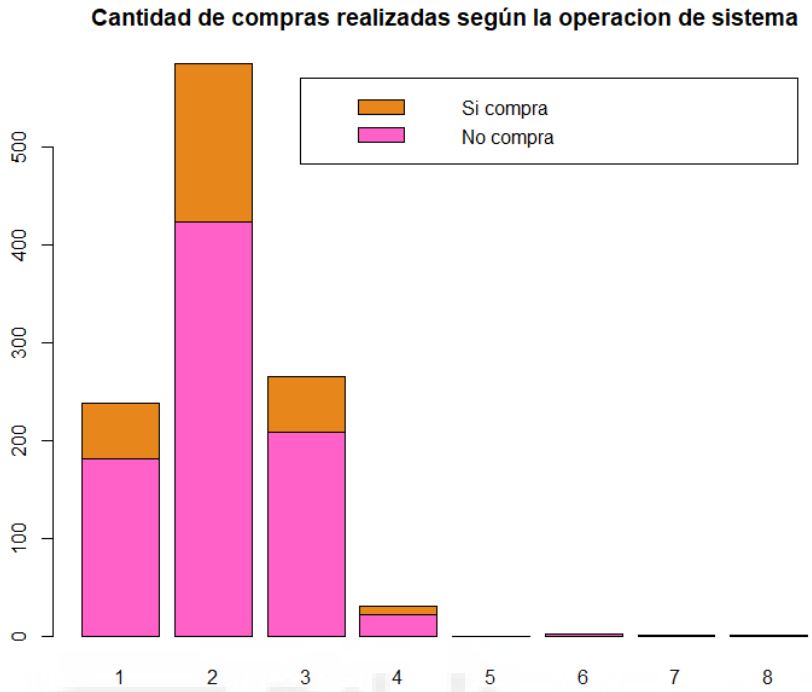
Fuente: Elaboración propia.



A continuación, se representa las compras adquiridas por los clientes según el tipo de gráfico utilizado al acceder a la web. En el *grafico 16*, se expone como el gráfico 2, es el más usado por los clientes, generando los mismos un total de 847 compras. Por otro lado, cabe destacar, la inutilización para el inicio de sesión en la página web de los gráficos 12, 14, 15, 16, 17, 18 y 19.

Gráfico 17. Gráfico de barras cantidad de compras según la operación del sistema.

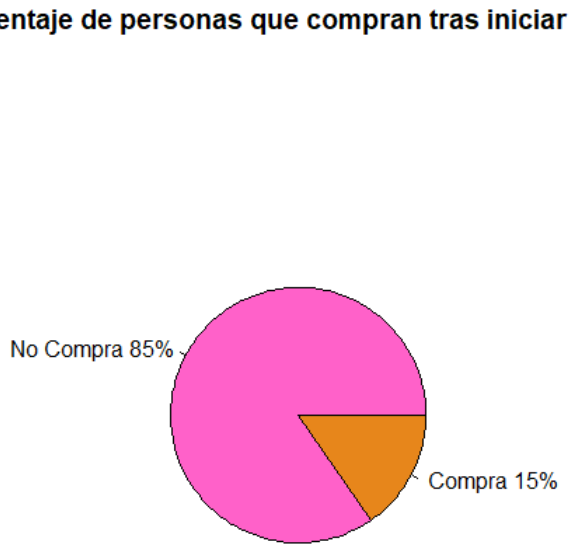
Fuente: Elaboración propia.



El *grafico 17,* el cual representa la decisión del cliente frente a la operación de sistema utilizada, se analiza que el sistema operativo 2, continuado del 3, 1 y 4, son los más usados, mientras que el resto de sistemas prácticamente no son utilizados. Esto puede deberse a que el incio de sesión desde los ordenadores con los sistemas 5, 6, 7 y 8, sean más antiguos y puedan generar algun tipo de incompatibilidad con la página web de estudio y por tanto, que los clientes no adquieran ningun bien de la misma.

Gráfico 18. Gráfico circular % de clientes que compran o no.

Fuente: Elaboración propia.



De las variables categóricas de estudio, es trascendental el estudio individualizado de la variable denominada compra. El interés mostrado en dicha variable, se debe a que el objetivo principal del presente informe, es la predicción de las compras generadas a través de la web del negocio. Es necesario recordar, que, a mitad del análisis descriptivo, se tomó la decisión de desarrollar el estudio únicamente para aquellos clientes que inician sesión más de una vez en algún de las categorías de la web, así como, que estén en las mismas una duración superior a 10 segundos. En el *gráfico 18,* se presentan los porcentajes de compra realizados al finalizar un año. El 15% de los clientes, adquieren un producto al iniciar sesión, mientras que el 85% deciden retirarse del comercio online, sin generar ninguna compra.

5.3.2 COMPONENTES PRINCIPALES

Tras el análisis numérico y gráfico, el cual, ha proporcionado información de las variables que exponen las características de los usuarios interesados en la web comercial, es beneficioso reducir su dimensionalidad, para facilitar la futura construcción de un modelo que pueda predecir a un nivel fiable, las futuras compras que adquiera la empresa.

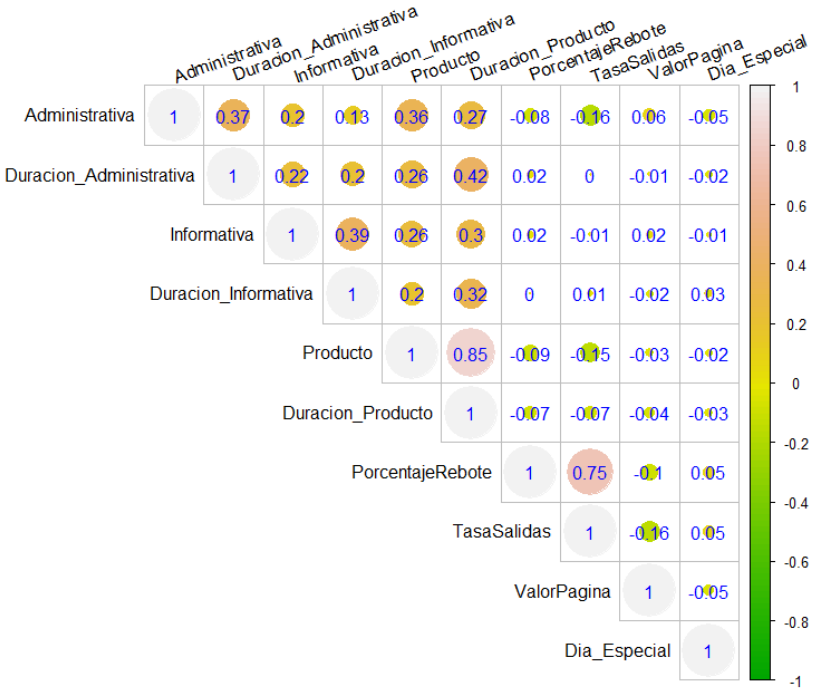
En primer lugar, dicho análisis exige que las variables han de ser numéricas, por ello, se tratan únicamente las 18 variables numéricas de la base de datos.

A continuación, se estudia la correlación existente entre las mismas, debido a que el análisis de Componentes Principales se realiza de aquellas variables numéricas que no están correlacionadas entre sí, es decir, las variables que no proporcionan información sobre otras variables.[21]

GRAFICO DE CORRELACIONES

Gráfico 19. Gráfico de correlaciones entre variables numéricas.

Fuente: Elaboración propia.



Existe altas correlaciones directamente proporcionales entre las variables producto y duración del producto, así como del porcentaje de rebote y la tasa de salidas. Por otro lado, el resto de las variables presentan correlaciones cercanas a 0, es por ello por lo que son necesarias para el análisis posterior de Componentes Principales. Por lo tanto, se ha decidido eliminar las variables duración del producto y tasa de salidas, ya que de ambas se adquiere información a partir de las variables producto y porcentaje de rebote.

Una vez seleccionadas las variables, se procede a realizar dos pruebas que proporcionan el óptimo uso del análisis de Componentes Principales. Son las denominadas pruebas de KMO (Kaiser, Meyer y Olkin) y de Bartlett.

El test de KMO, relaciona los coeficientes de correlación de las variables, por lo que, si el test proporciona un valor cercano a 1, el uso de Componentes Principales para reducir la dimensionalidad de los datos es correcta.

El test de KMO, de los clientes interesados en la página web de estudio, es de 0.67.

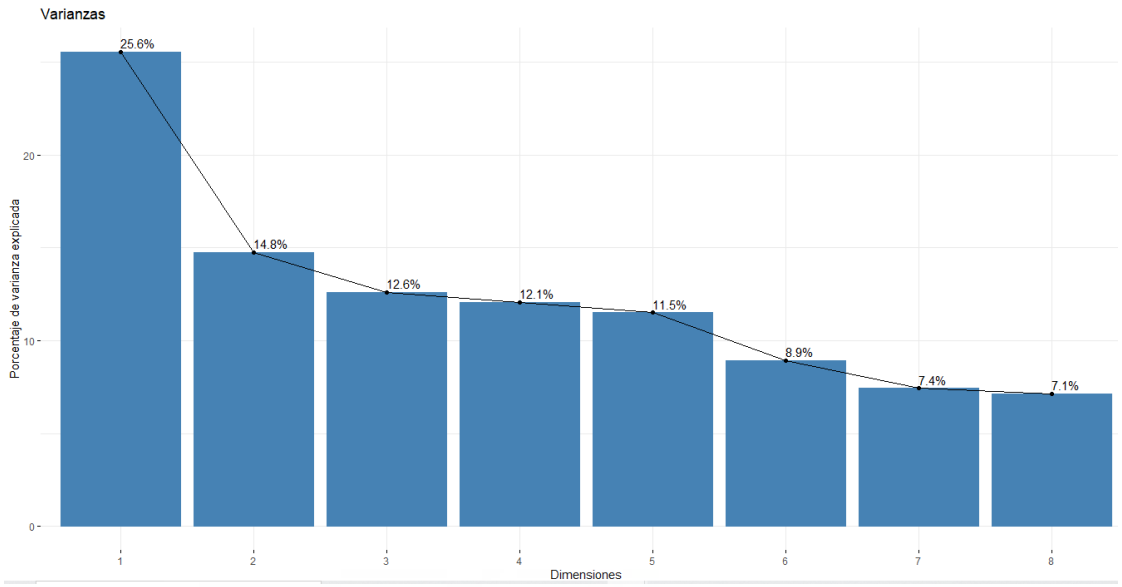
Por otro lado, el test de Bartlett, ofrece un valor que prueba si las variables vienen de poblaciones con varianzas iguales, o lo que es lo mismo, existencia de homocedasticidad en los datos. En el presente caso de estudio, se obtiene un valor inferior a 0.05 lo que quiere decir que la matriz de correlaciones no es significativamente distinta de la matriz de identidad, entonces se concluye con que ambos test corroboran la adecuación del uso de Componentes Principales.

En primer lugar, para evitar problemas derivados de las diferencias de escalas existentes en las variables de estudio, se tipifican sus valores siendo ahora el máximo valor igual a 1.

El criterio seguido para seleccionar el número de componentes principales que se han de seleccionar, se basa en el número de variables nuevas que acumulen más de un 75% de varianza.[22]

Gráfico 20. Gráfico procentaje de varianza acumulada de Componentes Principales.

Fuente: Elaboración propia.



Analizando el *gráfico 20*, se logra alcanzar un porcentaje acumulado de varianza superior a 75% con 5 Componentes Principales. Es por ello que estas últimas podrían ser las nuevas variables de estudio, reduciendo así la dimensionalidad de la base de datos compuesta únicamente por los clientes considerados interesados. Pero al estudiar las variables que quedan explicadas por las Componentes Principales, se advierte la imposible interpretación de la Componente 3. Por lo tanto, reduciendo a 4 Componentes Principales y aplicando la rotación Varimax, la cual consigue rotando las mismas, que queden correlacionadas con alguna de las antiguas variables, todas las nuevas variables pueden ser interpretadas, a pesar de que el porcentaje de varianza explicada se reduce a 64.98%.

Tabla 3. Matriz de correlaciones 5 CP. *Fuente: Elaboración propia.*

	PC1	PC2	PC3	PC4	PC5
Administrativa	0.66	-0.28	-0.34	0.12	0.13
Duración Administrativa	0.65	-0.01	-0.32	-0.01	0.26
Informativa	0.64	0.26	0.41	-0.17	-0.06
Duración Informativa	0.57	0.36	0.48	-0.10	-0.17
Producto	0.66	-0.11	-0.20	0.16	-0.18
Tasa Rebote	-0.09	0.63	-0.23	-0.37	0.57
Valor Página	0.03	-0.55	0.54	-0.03	0.62
Dia Especial	-0.06	0.44	0.10	0.86	0.20

Tabla 4. Matriz de correlaciones 4 CP. *Fuente: Elaboración propia.*

	PC1	PC2	PC3	PC4
Administrativa	0.80	-0.11	0.02	-0.05
Duración Administrativa	0.69	0.13	0.17	-0.06
Informativa	0.21	-0.01	0.79	-0.04
Duración Informativa	0.11	0.01	0.82	0.06
Producto	0.69	-0.07	0.18	0.06
Tasa Rebote	-0.17	0.73	0.14	-0.10
Valor Página	-0.12	-0.73	0.13	-0.15
Dia Especial	-0.05	0.04	0.03	0.98

Gráfico 21. Variables frente CP1 Y CP2.

Fuente: Elaboración propia.

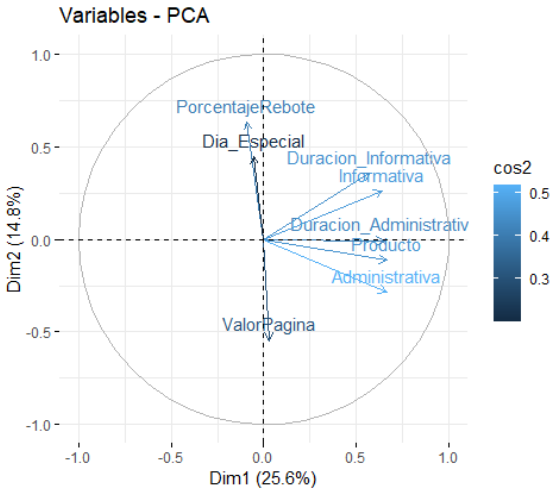


Gráfico 22. Variables frente CP1 y CP3

Fuente: Elaboración propia.

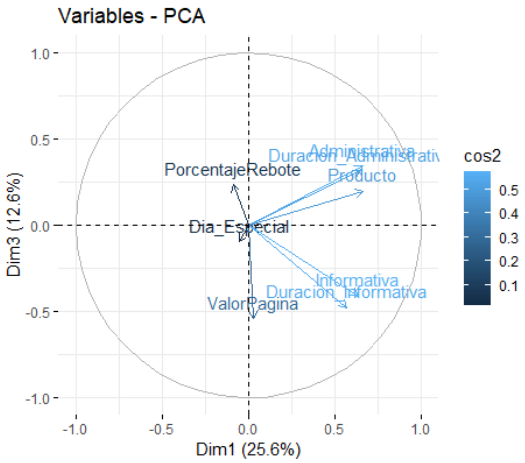


Gráfico 23. Variables frente CP1 y CP4

Fuente: Elaboración propia.

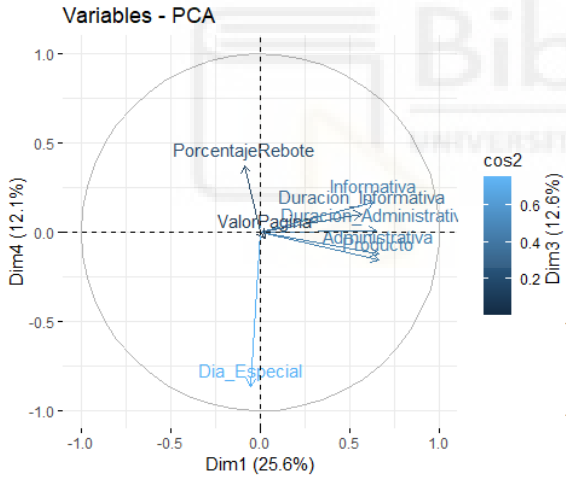


Gráfico 24. Variables frente CP2 y CP3

Fuente: Elaboración propia.

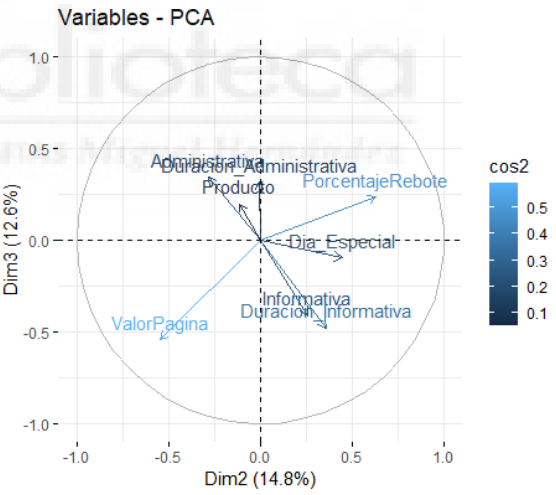


Gráfico 25. Variables frente CP2 y CP4

Fuente: Elaboración propia.

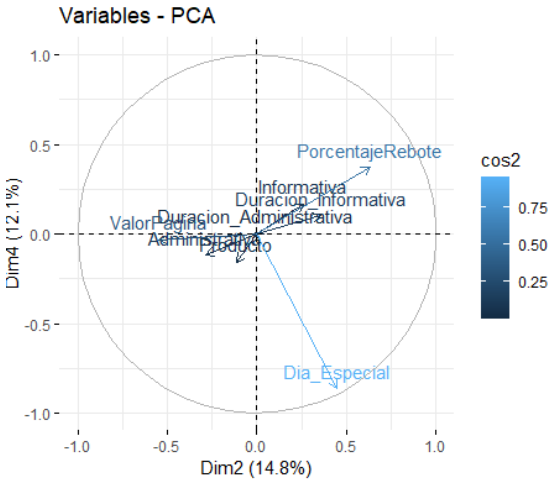
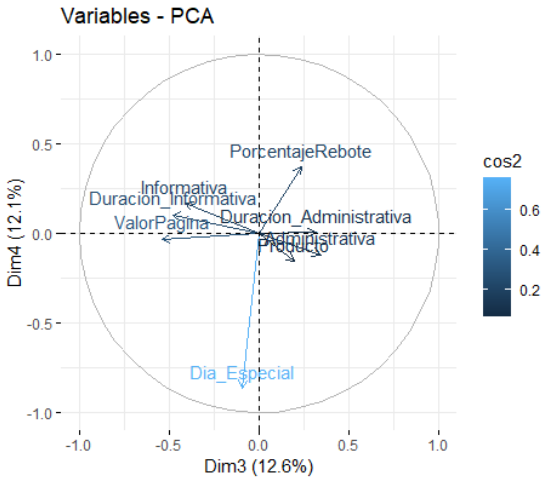


Gráfico 26. Variables frente CP3 y CP4

Fuente: Elaboración propia.



A la vista de los *gráficos 21, 22, 23, 24, 25 y 26*, donde se expone la influencia de las variables originales frente a las nuevas Componentes Principales, se deja ver las relacciones existentes entre las mismas, con el objetivo de eliminar aquellas variables que poseen grandes semejanzas en futuros estudios, y confirmar aquellas variables antiguas mejor explicadas por cada una de las Componentes Principales.

En efecto, se puede afirmar que:

- La Componente Principal 1, está asociada a aquellos valores relacionados con la página administrativa, así como los clientes que acceden a la página del producto, por lo que mide los accesos a la página administrativa y del producto.
- La Componente Principal 2, reúne los datos obtenidos mediante la herramienta Google Analitycs, entonces es de interés nombrarla como la valoración web del comercio.
- La Componente Principal 3, recoge los valores asociados a la página Informativa, por tanto, se nombra a continuación, como Página Informativa.
- Por último, la Componente Principal 4, únicamente queda explicada por la variable Día Especial, por tanto, se proporciona el mismo nombre a la nueva variable.

Gráfico 27. Observaciones frente CP1 Y CP2

Fuente: Elaboración propia.

Gráfico 28. Observaciones frente CP1 y CP3

Fuente: Elaboración propia.

Gráfico 29. Observaciones frente CP1 Y CP4

Fuente: Elaboración propia.

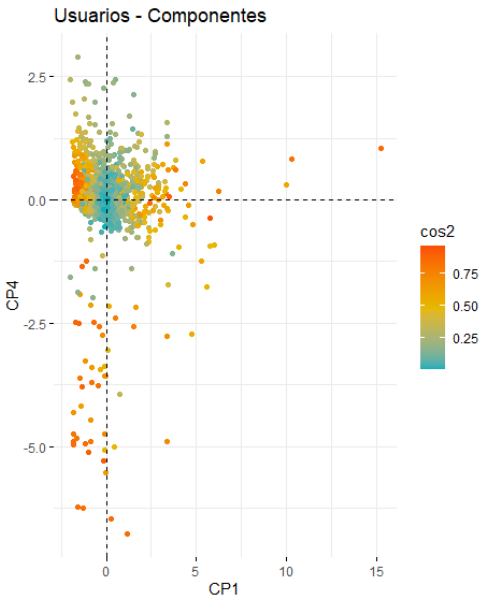


Gráfico 30. Observaciones frente CP2 Y CP3

Fuente: Elaboración propia.

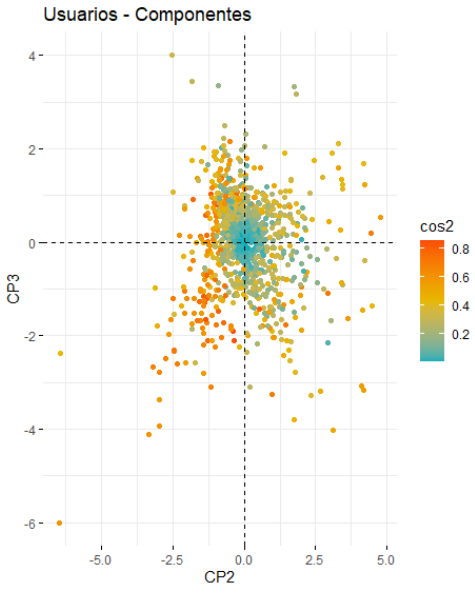


Gráfico 31. Observaciones frente CP2 Y CP4

Fuente: Elaboración propia.

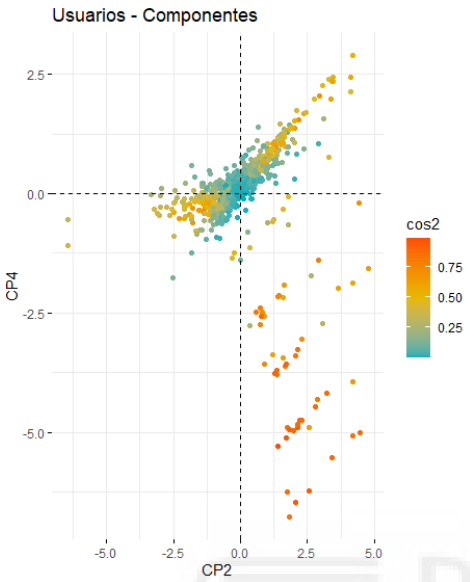
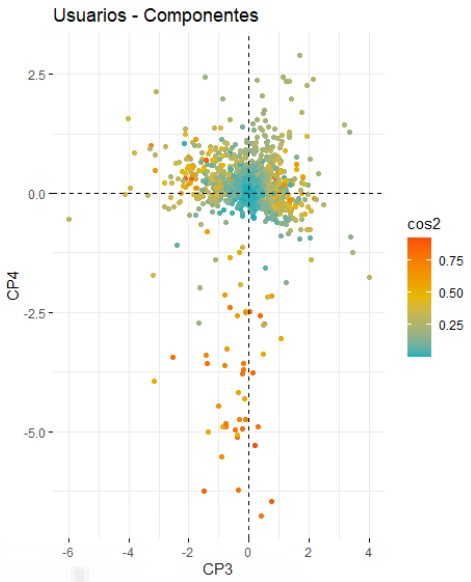


Gráfico 32. Observaciones frente CP3 Y CP4

Fuente: Elaboración propia.



En los *gráficos 27, 28, 29, 30, 31 Y 32* se ha representado la relación existente entre las observaciones de la base de datos de estudio frente a las nuevas variables para adoptar de una manera amplia, el comportamiento de los clientes. En los gráficos aparece una leyenda, la cual agrupa a los individuos por colores según su relación con las CP. Los grupos de individuos localizados en el punto 0,0 de los ejes, son aquellos que no proporcionan mucha información ante las componentes. Por otro lado, aquellos usuarios ubicados en los extremos de cada eje, son los que tienen mayor correlación con las nuevas variables.

En el caso del *gráfico 27*, se localiza una pequeña cantidad de individuos, situados en el cuadrante derecho del gráfico, lo que significa que estos clientes son aquellos que más acceso realizan a la página administrativa y del producto.

Por otro lado, en el cuadrante izquierdo superior, se encuentran aquellos individuos que mejor valoran la página web.

5.3.4 TÉCNICA FORWARD

Ya queda conocida la nueva base de datos para continuar con el estudio de manera más sencilla y simplificada, pero hasta el momento, no se había hecho uso de las variables discretas de la base de datos de los clientes interesados, debido a que la técnica de Componentes Principales no es aplicable a este tipo de variables, por lo que ha sido empleada la técnica forward, o selección de variables hacia delante.

Dicha técnica ha sido aplicada tanto para la nueva base de datos de 5 Componentes Principales, así como para la de 4.

Tras la práctica anteriormente explicada, las variables dummy que fomentan la clasificación, son el gráfico tipo 6 y la región número 3, en la base de datos de 5 Componentes Principales y, por otro lado, las variables Operación del Sistema número 3, el Navegador 3 y la región 8, para la base de datos de 4 Componentes Principales.

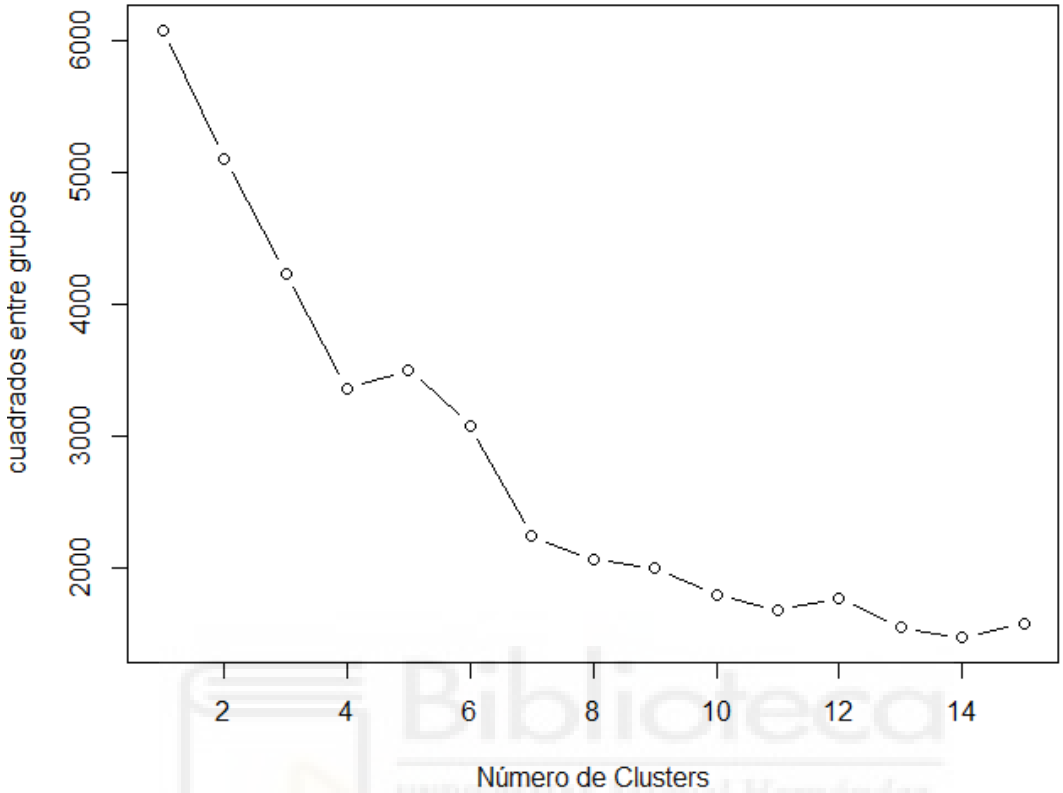
2.3.5 ANÁLISIS CLUSTER

Una vez reducida la base de datos, se aplica la técnica de clasificación no supervisada, denominada, Análisis Clúster o análisis de conglomerados.

Para llevar a cabo dicha clasificación, se escoge la base de datos de 4 Componentes Principales, así como las variables discretas Operación del Sistema número 3, el Navegador 3 y la región 8, ya que como ha sido mencionado anteriormente, es la que ha permitido interpretar las nuevas variables, a diferencia de la base de datos de 5 Componentes Principales.

En primer lugar, se he reproducido el gráfico del método Elbow, caracterizado por indicar el posible valor de k puntos iniciales para desarrollar el método.

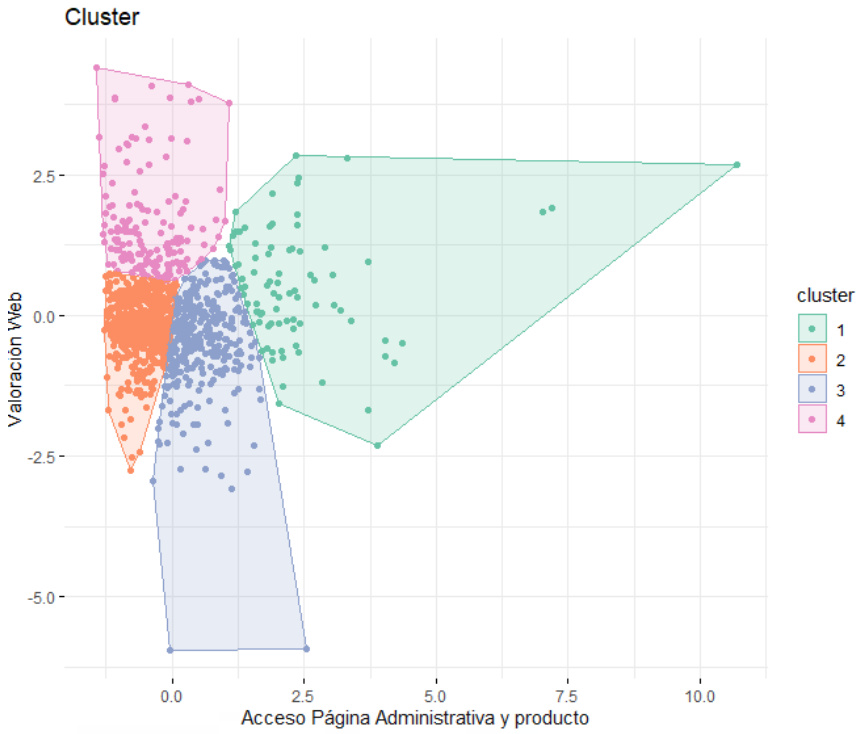
Gráfico 33. Gráfico Elbow. *Fuente: Elaboración propia.*



El *grafico 33* expone lo que se podría denominar como el codo, o lo que es lo mismo, donde la secuencia de los puntos, obtiene el primer desvío a la baja. El valor previsto para iniciar el algoritmo de Kmedias es de 4 puntos como centroides iniciales de las observaciones de la base de datos.

A continuación, y haciendo uso del valor k=4 se inicia el algoritmo de las kmedias, obteniendo una clasificación de las observaciones en 4 grupos según sus características comunes.

Gráfico 34. Clúster según las CP1 Y CP2. *Fuente: Elaboración propia.*



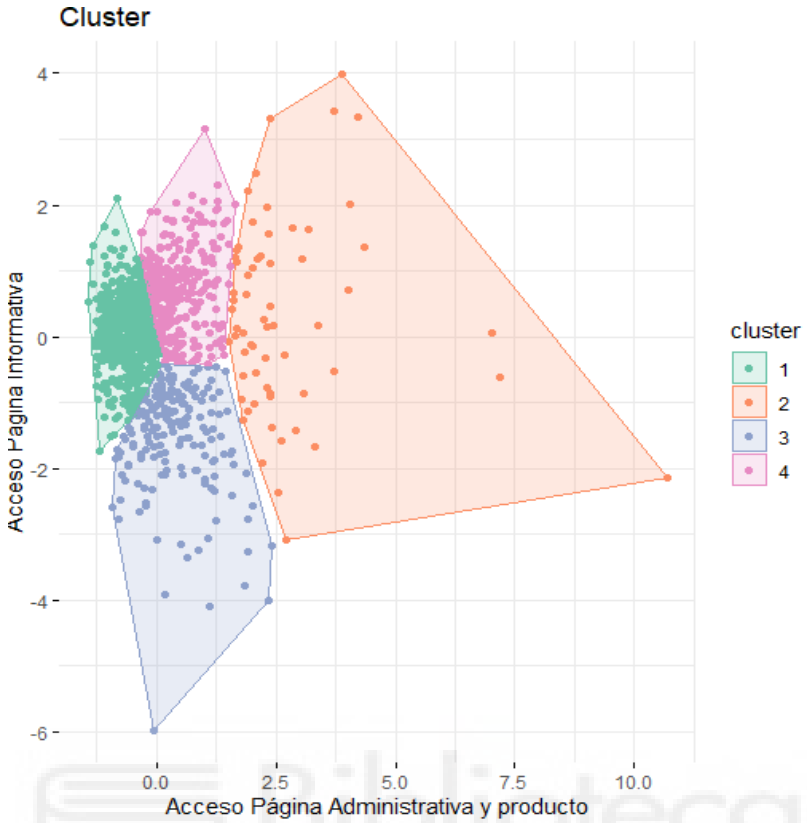
El *gráfico 34* manifiesta la clasificación de las observaciones registradas de cada individuo que inicia sesión en la web clasificados en 4 grupos distinguidos entre ellos, según las características que cada observación comparte con otras observaciones y se encuentran representadas en relación a las Componentes Principales 1 y 2, o lo que es lo mismo, las variables Accesos a la Página Administrativa y del producto y la valoración de la Web.

En lo que se refiere a los usuarios que acceden a la página administrativa y del producto más veces y se mantienen en la misma categoría una duración elevada de tiempo, se localizan los individuos clasificados en el clúster 1, como es el caso de los individuos 869 y 19.

Por otro lado, los grupos 2 y 3 se encuentran clasificados en torno a la coordenada 0,0 por lo que no quedan muy especificadas sus visitas a la página administrativa y del producto, así como la valoración que dan a la web.

En cuanto a la valoración de la web, los individuos pertenecientes al grupo 4, como es el caso de los individuos 325 y 38, son aquellos que proporcionan una valoración baja a la página web, y, por otro lado, adquieren un porcentaje de rebote bajo, generando entonces un dato positivo para la empresa.

Gráfico 35. Clúster según las CP1 Y CP3. *Fuente: Elaboración propia.*



En el *gráfico 35* se analizan los diferentes clústeres establecidos según las nuevas variables página administrativa y del producto y página informativa. Los individuos agrupados en el clúster 2, como por ejemplo el 352 y 948, se caracterizan por acceder un cuantioso número de veces a la página administrativa y del producto, mientras que los individuos como el 5 y 456, clasificados en el grupo 3, no realizan muchos accesos a la página informativa, pues dicho clúster se encuentra localizado en torno a valores negativos.

En cuento a los grupos, 1 y 4, se analiza que respecto a las presentes variables realizan accesos entorno a la media, por lo que no son caracterizados por grandes entradas a ambas categorías de la web comercial.

Gráfico 36. Clúster según las CP1 Y CP4. *Fuente: Elaboración propia.*

En el *gráfico 36*, se forman 2 grupos situados cerca del 0,0, significando que los individuos clasificados en dichos clústeres, no inician sesión en la página web un día cercano a una festividad y acceden un numero común de veces a la categoría administrativa y del producto.

Los individuos como 171 y 273, están clasificados en el grupo 4 en el cuadrante negativo, lo que denota que los clientes de dicho clúster, no acceden al negocio online un día cercano a una fecha denotada como festiva.

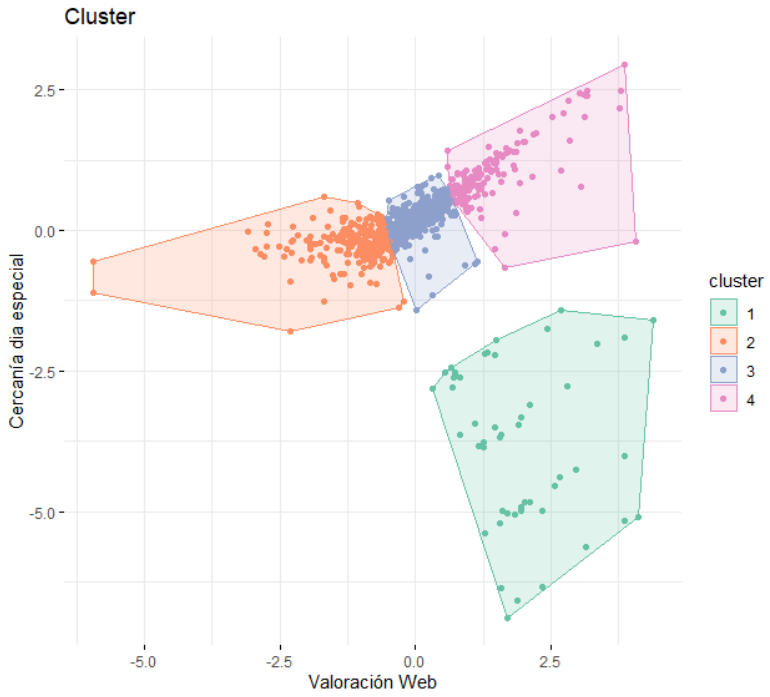
En cuanto al grupo 1, como ya ha sido mencionado en gráficos anteriores, son aquellos individuos que acceden repetidamente y mantienen su estancia en la página administrativa y del producto, pues el grupo 1 queda localizado en el cuadrante positivo.

Gráfico 37. Clúster según las CP2 Y CP3. *Fuente: Elaboración propia.*

El *gráfico 37*, muestra como los individuos clasificados en el grupo 3, como por ejemplo los individuos, 3 y 239 son los que determinan una valoración de la web baja y un valor de tasa de rebote bajo. Por otro lado, el grupo 4, formado por individuos como el 103 y 241, son aquellos que no acceden en muchas ocasiones a la página informativa y, además, valoran la web con un valor bajo y tienen un porcentaje de rebote bajo.

En cuanto a los individuos que forman los clústeres 1 y 2, se localizan en la coordenada, por lo que, como se ha mencionado anteriormente, se caracterizan por dar una valoración neutra a la tienda online, así como que los accesos a la categoría informativa no son ni excesivos ni reducidos.

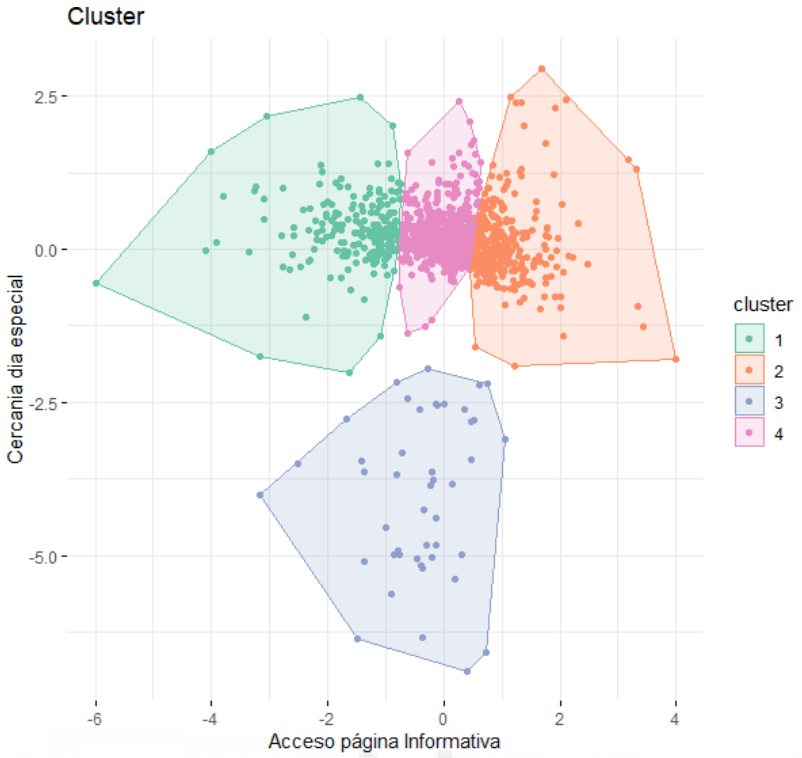
Gráfico 38. Clúster según las CP2 Y CP4. *Fuente: Elaboración propia.*



El *gráfico 38* demuestra que los grupos 2, 3 y 4, son los formados por los individuos que valoran la web e inician sesión de manera neutra, ya que se localizan entorno al valor 0,0.

Por otro lado, el conjunto de individuos clasificados en el grupo 1, como el 138 y 273, son aquellos que se caracterizan por dar una valoración de la web negativa, pero con una baja tasa de rebote, dato de interés para la empresa, pues significa que sus clientes no acceden y se marchan mostrando un bajo interés por la web. A su vez el grupo 1 también define que los individuos pertenecientes al mismo, aseguran que no acceden a la web, días antes de una festividad.

Gráfico 39. Clúster según las CP3 Y CP4. *Fuente: Elaboración propia.*



En el *gráfico 39,* se analiza como los individuos del grupo 3, como el 352 y 384, son aquellos que acceden a la web un día no cercano a cierta festividad, debido a que el clúster se localiza en torno a los valores negativos del eje y en el gráfico.

Por otro lado, el resto de grupos se localizan cercanos al valor 0, por lo que su acceso a la página informativa debería de ser de un valor considerado neutro y en cuanto a la cercanía a un día especial, los valores de dichos individuos se acercan a 0, por tanto, no acceden a la web un día cercano a lo que se ha determinado como festivo.

Tras la visualización los *gráficos 34, 35, 36, 37, 38 y 39* las características que engloba cada clúster son expuestas en *la tabla 5.*

Tabla 5. Características de los clústeres. *Fuente: Elaboración propia.*

GRUPOS	CARACTERÍSTICAS
Clúster 1	Acceso a la página administrativa y del producto.
	Valoración web y tasa de rebote baja
Clúster 2	Acceso a la página administrativa y del producto
Clúster 3	Poca frecuencia en acceder a la página informativa.
	Valoración y tasa de rebote baja
Clúster 4	Inicio de sesión lejano a una festividad
	Valoración y tasa de rebote baja

5.3.5 K NEAREST NEIGHBORS

A continuación, se aplica la segunda técnica de clasificación, el método K-Neighbors o el método de los vecinos más cercanos. En este algoritmo, van a ser empleadas las dos bases de datos nuevas, extraídas gracias a el método de Componentes Principales y la técnica Forward:

- Datos1 = 4 Componentes, Operación del Sistema 3, Navegador 3 y Región 8.
- Datos2 = 5 Componentes, Gráfico 6 y Región 3.

El desarrollo del presente método se basa en la clasificación mediante la división de la base de datos en dos, el 50% de los datos, serán usados para entrenamiento y el resto, serán los datos a clasificar. Por lo tanto, la clasificación de un individuo localizado en los datos de entrenamiento, será clasificado en los k individuos más cercanos a él. El valor de k será aquel que proporcione un error inferior al resto.

En primer lugar, se aplica el algoritmo a la base de datos definida como Datos 1.

Se realiza la prueba con grupos de k igual a 3, 5, 7 ,9 y 11. El grupo seleccionado, es aquel que menor porcentaje de error genere.

Tabla 6. Tabla de confusión método KNN datos 1. *Fuente: Elaboración propia.*

P/K7	0	1
0	369	17
1	141	35

Tabla 7. Porcentajes de aciertos con datos 1. *Fuente: Elaboración propia.*

% / Grupos	3	5	7	9	11
Acierto	69.39	69.75	71.89	70.46	70.81

El método KNN, para la base de datos 1, ofrece un porcentaje de acierto de 71.89%, con k=7.

A continuación, se desarrolla el mismo algoritmo, para la base de datos 2, con la finalidad de conocer la capacidad de predicción del algoritmo con los datos no incluidos en la fase de entrenamiento.

Se vuelve a realiza la prueba con grupos de k igual a 3, 5, 7 ,9 y 11.

Tabla 8. Tabla de confusión método KNN datos 2. *Fuente: Elaboración propia.*

P/K7	0	1
0	371	15
1	142	34

Tabla 9. Porcentajes de aciertos con datos 2. *Fuente: Elaboración propia.*

% / Grupos	3	5	7	9	11
Acierto	69.21	71.17	71.70	72.06	71.35

Para dicha base de datos, se logra superar el nivel de fiabilidad aplicando el método KNN a un porcentaje de 72.06% de k igual a 9.

5.3.6 SUPPORT VECTOR MACHINE

Finalmente, se efectúa otro método de clasificación conocido como Support Vector Machine, que como se ha mencionado anteriormente se desarrolla mediante la función svm, implementada en el programa RStudio y el método manual. La diferencia esencial de ambas maneras, tiene su fundamento en las distancias empleadas en cada una de ellas.

5.3.6.1 SUPPORT VECTOR MACHINE MANUAL

El método se efectúa mediante un algoritmo que resuelva un problema de optimización lineal, aplicando la distancia de Manhattan. La variable objetivo, Compra, adquiere los valores -1, cuando el cliente no realiza ninguna compra y 1 cuando si lleva a cabo una transacción en la web comercial.

A continuación, se efectúa el cálculo del valor support vector machine para los datos 1, mediante la suma de (wj+ - vj-) xij + b >= 0, o lo que es lo mismo, los pesos multiplicados por el valor de las x correspondientes más la pendiente. Si el valor obtenido es inferior a 0, la observación se clasifica en el grupo 0 y en el caso contrario, la observación se asocia al grupo 1.

Del presente método se adquiere un acierto de 67.59%, el cual, proviene de que de las 1123 clientes que han iniciado sesión, 170 han sido clasificados correctamente en el grupo de decisión de compra y 589 en los usuarios que no compran, mientras que 360 usuarios han sido clasificados de forma errónea. Cabe destacar, el método Support Vector Machine, alcanza una fiabilidad de predicción inferior al método KNN empleado anteriormente.

Se repite el procedimiento, para el conjunto de datos 2. En este caso se logra aumentar el nivel fiabilidad de clasificación a 76.05%, pues el algoritmo consigue clasificar correctamente a 854, y solo comete error al clasificar 269.

Tabla 10. Tabla de confusión método SVM manual datos 1. *Fuente: Elaboración propia.*

P/K7	0	1
-1	589	250
1	114	170

Tabla 11. Tabla de confusión método SVM manual datos 2. *Fuente: Elaboración propia.*

P/K7	0	1
-1	707	132
1	137	147

5.3.6.2 SUPPORT VECTOR MACHINE FUNCIÓN SVM RSTUDIO

En este caso, el método Support Vector Machine, se desarrolla haciendo uso de la función SVM propuesta por el programa Rstudio, donde posiblemente aplique la distancia cuadrática. dividiendo la base de datos, el 50% en datos de entrenamiento y el resto de observaciones, para datos de prueba, permitiéndose así conocer la fiabilidad de predicción que proporciona dicho modelo.

Elaborado el modelo SVM para ambas bases de datos, siendo la función discriminante de la base de datos 1 y 2 las expuestas a continuación:

- Modelo1 = svm (Variable Compra ~ CP1 + CP2 + CP3 + CP4 + Variable Operación del sistema 3 + Navegador 3 + Región 8)
- Modelo2 = svm (Variable Compra ~ CP1 + CP2 + CP3 + CP4 + CP5 + Gráfico 6 + Región 3)

A continuación, se genera la predicción de los datos de prueba mediante los datos de entrenamiento, y se compara la clasificación obtenida frente los valores reales que toma la variable objetivo obteniendo un total de 386 observaciones en las que la predicción ha sido correcta, y un total de 176 usuarios que se han clasificado de forma errónea según su decisión de compra. Finalmente, este método proporciona, una fiabilidad del mismo de 84,33%, siendo de todos los métodos utilizados, el que mejor resultado proporciona.

Un dato relevante observado al aplicar la función SVM es que, a diferencia del resto de procedimientos empleados, ha clasificado a todos los clientes pertenecientes a ambos conjuntos de datos de prueba en 0, es decir, que no generan ninguna compra tras acceder a la página web. Aunque, ha logrado alcanzar un nivel de fiabilidad superior en comparación al resto de métodos clasificadores utilizados anteriormente, el resultado de la función SVM, no aporta información relativa de las variables, la cual, podamos emplear de ninguna manera o de la que podamos inducir alguna conclusión.

Tabla 12. Tabla de confusión método SVM función datos 1. *Fuente: Elaboración propia.*

P/K7	0	1
0	386	0
1	176	0

Tabla 13. Tabla de confusión método SVM función datos 2. *Fuente: Elaboración propia.*

P/K7	0	1
0	386	0
1	176	0

6. ANÁLISIS Y DISCUSIÓN

Una vez obtenidos los resultados de todos los procedimientos implementados en el presente proyecto, es de interés desarrollar comparativas entre los métodos empleados fomentando así, posibles mejoras para futuros estudios que desee llevar a cabo el negocio.

En primer lugar, gracias al análisis de componentes principales, se han detectado algunas variables, que, por su alta correlación entre ellas, pueden ser excluidas en futuros estudios. Este caso, sucede concretamente entre las variables duración de la estancia en la página del producto y la categoría web de producto, así como tasa de salidas y porcentaje de rebote.

Se ha tomado la decisión de continuar el análisis con dos nuevas bases de datos, una formada por 5 Componentes Principales, y otra con únicamente 4. Se ha decidido así debido a que, aunque con mayor número de componentes principales se alcanza mayor porcentaje de varianza explicada, con la selección de 5 Componentes Principales, ha sido difícil interpretarlas.

A pesar de ello, se han mantenido ambas nuevas bases de datos, con el fin, de comparar cual, de ellas, genera un modelo que prediga las decisiones de compras realizadas por parte de los usuarios con un menor porcentaje de error.

A continuación, se han incorporado a ambas nuevas bases de datos, aquellas variables dummy, que hacía incrementar la fiabilidad del modelo de clasificación a continuación explicado.

Del análisis clúster se ha obtenido aquellas características comunes que comparten ciertos grupos de observaciones. Mayoritariamente, en el grupo 1 han quedado clasificados aquellos individuos que presentan varios inicios de sesión en la página administrativa y del producto, así como también valoran la página vez con un valor bajo y obtienen una tasa de rebote baja. Por otro lado, el grupo dos de observaciones es el caracterizado por los clientes que acceden a la página administrativa y del producto únicamente. El clúster 3, está compuesto de los usuarios que no tienen muchos accesos al cabo de un año en la página informativa, y, además, proporcionan una valoración y una tasa de rebote

pequeña. Por último, el grupo 4, se forma de los individuos que dan una tasa de rebote y valoración de la página baja y también, no inician sesión en la web un día cercano a una festividad.

Tras el análisis Clúster, se ha hecho uso de dos métodos de clasificación, K-Nearest-Neighbors y Support Vector Machine, para analizar de que algoritmo, se obtiene un mayor nivel de fiabilidad para futuras predicciones de la empresa web, aplicando ambas nuevas bases de datos:

- Datos1 = 4 Componentes, Operación del Sistema 3, Navegador 3 y Región 8.
- Datos2 = 5 Componentes, Gráfico 6 y Región 3.

Tras desarrollar los dos algoritmos que clasifica a partir de la división de la base de datos, la decisión de compra que posee en usuario que inicia sesión en el negocio comercial en 0 si no compra y 1 si sí lo hace, para la base de datos 1, los resultados obtenidos han sido los que se exponen en la siguiente tabla:

Tabla 14. Métodos de clasificación base de datos 1. *Fuente: Elaboración propia.*

KNN	SVM Manual	SVM Función
71.89% con k=7	67.59%	84.33%

Se ha repetido la aplicación de los mismos modelos de clasificación, pero ahora con la base de datos 2, obteniendo los siguientes niveles de fiabilidad:

Tabla 15. Métodos de clasificación base de datos 2. *Fuente: Elaboración propia.*

KNN	SVM Manual	SVM Función
72.06% con k=9	76.05%	84.33%

En las *tablas 14 y 15,* se revelan las diferencias en los resultados obtenidos en cada método de clasificación supervisada al utilizar las distintas y nuevas bases de datos definidas anteriormente. La base de datos 2, ofrece niveles de fiabilidad superiores a los generados con la base de datos 1. En cuanto a los resultados

obtenidos con el método de clasificación Support Vector Machine, desarrollado mediante la función implementada en la interfaz RStudio, es de interés denotar que, en ambas bases de datos, se obtenga el mismo nivel de fiabilidad. Este hecho puede deberse a que la clasificación ofrecida por el método en ambas dos, clasifica a las observaciones consideradas de entreno, en 0, o lo que es lo mismo, en que su decisión de compra como usuario es no realizar ninguna transacción tras iniciar sesión en la web.

Definitivamente, el método que mejor clasifica los datos de entrenamiento, de manera predictiva, debido al mayor nivel de fiabilidad que ofrece, es la función Support Vector Machine implementada en la interfaz RStudio, aun clasificando a los individuos de entrenamiento en 0. Sin embargo, para la empresa sería más conveniente desarrollar uno de los algoritmos de clasificación, que aporte la información de las variables necesarias para inducir a conclusiones sobre las mismas. Es por ello, que es adecuado hacer uso del método Support Vector Machine implementado manualmente, ya que es el segundo algoritmo que mejor nivel de fiabilidad nos ofrece.

7. CONCLUSIONES Y PROPUESTAS

Gracias al desarrollo del presente trabajo, se han denotado ciertos aspectos que podrían resultar beneficiosos en el sentido de producir simplicidad tanto en futuras recogidas de datos como en el mismo análisis mediante técnicas de machine learning.

En primer lugar, se debe aludir a que determinadas variables podrían ser obviadas en futuras recogidas de datos de manera que tanto el análisis como la recolecta de datos sería más asequible. Los *gráficos 21, 22, 23, 24, 25 y 26* enuncian que las variables duración informativa y página informativa presentan una evidente relación entre ellas, en consecuencia, en próximas investigaciones que desee llevar a cabo el negocio, se podría mantener únicamente la variable página informativa que recoge las veces que acceden los clientes a dicha página durante un periodo de un año, debido a que presenta un mayor coeficiente. Lo mismo ocurre entre las variables página administrativa, duración página administrativa y página del producto, por esta razón podrían conservarse para un estudio futuro, las variables página administrativa y página del producto. En relación al resto de características de la web comercial, deberían mantenerse, destacando la función de la variable valor de la página, pues no presenta ninguna similitud con otras variables.

Posteriormente, con el análisis clúster, los individuos se han clasificado en 4 grupos según las siguientes características expuestas en los *gráficos 34, 35, 36, 37, 38 y 39:*

- Grupo 1: Individuos que inician sesión en la página administrativa y del producto y valoran la web y generan una tasa de rebote reducida.
- Grupo 2: Individuos que acceden a la página administrativa y del producto.
- Grupo 3: Individuos que no acceden mucho a la página informativa y proporcionan una valoración y tasa de rebote pequeña.
- Grupo 4: Individuos que dan una tasa y valoración a la web baja, y también, inician sesión online un día lejano a una festividad.

Asimismo, sería conveniente mantener en el análisis el método clúster, teniendo en cuenta que ha facilitado una agrupación de los 1123 clientes interesados en la web, según sus características afines.

Por último, de los distintos métodos de clasificación de los clientes, según su decisión de comprar tras acceder a las páginas web del comercio, considerando el uso de dos bases distintas de datos, destaca el porcentaje de fiabilidad del algoritmo Support Vector Machine empleado con la función proporcionada en RStudio Las ventajas y desventajas de cada método de clasificación empleado en el presente informe son las que se muestran en la *tabla 16.* [23], [24]

Tabla 16. Ventajas y desventajas métodos de clasificación. *Fuente: Elaboración propia.*

	Ventajas	Desventajas
Método K-Nearest Neighbors	No es necesario suponer conceptos clave de dicho modelo	Variación de los valores de k, hasta alcanzar el que ofrezca mejor clasificación.
Método Support Vector Machine manual	Conocimiento en el desarrollo de los cálculos empleados	Resultados menos fiables que los obtenidos con la función implementada en RStudio
Método Support Vector Machine	Ofrece la clasificación con menor porcentaje de error	Clasifica a todos los clientes en 0, es decir, no comprarían en la página web

Podría ser atrayente, el desarrollo en futuros estudios de la empresa de distintos algoritmos, con el fin detectar aquel que proporcione mayor fiabilidad de predicción, como por ejemplo aplicando la regresión logarítmica.

Asimismo, podría ser de interés extender el análisis con el estudio de que según qué características posea el cliente, cuales acceden más a cada página web, o compra un tipo concreto de producto que venda la empresa online.

8. BIBLIOGRAFIA

[1] El consumidor tras el coronavirus: más compras por Internet y menos ropa | Sociedad | EL PAÍS (elpais.com)

[2] Las compras “online” vuelven a batir récord al crecer un 286% (lavanguardia.com)

[3] E.Martín y R.Caballero. Las bases de Big Data, *Google Académico*,2020.

[4] M.A.Boden. Inteligencia Artificial, *Google Académico*, 2017.

[5] M.Sergio. Analítica Web: medir para triunfar, *Google Académico*, 2010.

[6] R.Chardonneau. Google Analytics: analice el tráfico de su sitio web para mejorar los resultados, *Google Académico*, 2014.

[7] M.M.Martínez. Analítica Web para empresas: Arte, ingenio y anticipación, *Google Académico*, 2010.

[8] R.Chardonneau. Google Analytics: analice el tráfico de su sitio web para mejorar los resultados, *Google Académico*, 2014.

[9] Russo, Claudia, Ramón, Hugo D., Alonso, Nicolás, Cicerchia, Lucas Benjamin , Esnaola, Leonardo, Tessore, Juan Pablo, Tratamiento masivo de datos utilizando técnicas de machine learning, VIII Workshop de Investigadores en Ciencias de la Computación, 2016.

[10] Galbiati R. Jorge, Componentes Principales, Google Académico, 2010.

[11] Luis Vicente Villardón. José, Introducción al Análisis Clúster, Universidad de Salamanca, 2007.

[12] S.A.José Luis, Técnicas Estadística en Análisis de Mercados k-Nearest Neigbour, 2019.

[13] I. Barbona, Beltran. Método de clasificación supervisada, support vector machine: una aplicación a la clasificación automática de textos, Universidad Nacional del Rosario, 2016.

[14] C. Ávila, Jorge, Nuevo algoritmo de protección de distancia basado en el reconocimiento de patrones de onda viajera, Repositorio académico digital UANL, 2004.

[15] Funciones de distancia y discriminante para Análisis discriminante, Minitab.

[16] J. Allaire. Joshep, RStudio, Wikipedia, 2021.

[17] Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.

[18] Porcentaje de rebote, Ayuda de Analytics.

[19] Porcentaje de salidas frente a porcentaje de rebote, Ayuda de Analytics.

[20] Calcular el valor de página, Ayuda de Analytics.

[21] U.P. Keiry Margarita, Componentes Principales y Análisis Factorial, RPubs, 2020. RPubs - Componentes Principales y Análisis Factorial

[22] Gil Martínez, Cristina, Análisis de Componentes Principales, RPubs, 2018. RPubs - Análisis de componentes principales (PCA)

[23] Cifuentes Ramos. Felipe, Clasificación automática de Tweets utilizando K-NN y K-Means como algoritmos de clasificación automática, aplicando TF-IDF y TF-RFL para las ponderaciones, Pontificia Universidad Católica de Valparaíso, 2016.

[24] Moreno Hojas. Ignacio, Maquinas de soporte de vectores, Stat Developer.

9. ANEXOS

A continuación, se representa el código fuente de RStudio empleado para desarrollar y obtener los resultados expuestos en el presente informe.

```
# Introducción de los datos

library(readr)

library(tidyverse)

library(DT)

library(kableExtra)

library(fastDummies)

library(cowplot)

library(ggplot2)

library(dplyr)

library(corrplot)

library(grDevices)

library(Hmisc)

library(pander)

library(class)


datos <-
read_csv("C:/Users/34660/Desktop/online_shoppers_intention.
csv")

summary(datos)

############## TRATAMIENTO DE DATOS ######################
```

```
# Renombrar variables

names(datos)[which(names(datos) == "Administrative")] <-
"Administrativa"

names(datos)[which(names(datos) ==
"Administrative_Duration")] <- "Duracion_Administrativa"

names(datos)[which(names(datos) == "Informational")] <-
"Informativa"

names(datos)[which(names(datos) ==
"Informational_Duration")] <- "Duracion_Informativa"

names(datos)[which(names(datos) == "ProductRelated")] <-
"Producto"

names(datos)[which(names(datos) ==
"ProductRelated_Duration")] <- "Duracion_Producto"

names(datos)[which(names(datos) == "BounceRates")] <-
"PorcentajeRebote"

names(datos)[which(names(datos) == "ExitRates")] <-
"TasaSalidas"

names(datos)[which(names(datos) == "PageValues")] <-
"ValorPagina"

names(datos)[which(names(datos) == "SpecialDay")] <-
"Dia_Especial"

names(datos)[which(names(datos) == "Month")] <- "Mes"

names(datos)[which(names(datos) == "OperatingSystems")] <-
"Operacion_Sistema"

names(datos)[which(names(datos) == "Browser")] <-
"Navegador"

names(datos)[which(names(datos) == "TrafficType")] <-
"Grafico"
```

```
names(datos)[which(names(datos) == "VisitorType")] <-
"Visitor"

names(datos)[which(names(datos) == "Weekend")] <-
"Fin_Semana"

names(datos)[which(names(datos) == "Revenue")] <- "Compra"


#Características principales de los datos

datos$Mes<-factor(datos$Mes)

datos$Operacion_Sistema<-factor(datos$Operacion_Sistema)

datos$Navegador<-factor(datos$Navegador)

datos$Region<-factor(datos$Region)

datos$Grafico<-factor(datos$Grafico)

datos$Visitor<-factor(datos$Visitor,labels = c("Nuevo",
"Otros", "Usual"))

datos$Fin_Semana <-factor(datos$Fin_Semana,labels =
c("Entre semana", "Fin de semana"))

datos$Compra <-factor(datos$Compra,labels = c("No compra",
"Si compra"))


##################3 ANALISIS DESCRIPTIVO ##############

summary(datos)    #Resumen datos


#Tabla descriptivos de las variables numericas.

datosnumericos1<-
data.frame(Administrativa,Duracion_Administrativa,Informati
va,Duracion_Informativa)
```

```
tabla1<-
data.frame(c("Media","Varianza","Minimo","Mediana","Maximo"
))

for(i in 1:ncol(datosnumericos1)){

  media <- mean(datosnumericos1[,i])

  varianza <- var(datosnumericos1[,i])

  perc <- quantile(datosnumericos1[,i],probs=c(0,0.5,1))

  vector <- c(media,varianza,perc)

  tabla1 <- cbind(tabla1,vector)}

colnames(tabla1)<-c("Variables",colnames(datosnumericos1))

kable(tabla1,format='markdown',digits=2,padding=0)


datosnumericos2<-
data.frame(Producto,Duracion_Producto,PorcentajeRebote,Tasa
Salidas,ValorPagina,Dia_Especial)

tabla2<-
data.frame(c("Media","Varianza","Minimo","Mediana","Maximo"
))

for(i in 1:ncol(datosnumericos2)){

  media <- mean(datosnumericos2[,i])

  varianza <- var(datosnumericos2[,i])

  perc <- quantile(datosnumericos2[,i],probs=c(0,0.5,1))

  vector <- c(media,varianza,perc)

  tabla2 <- cbind(tabla2,vector)}

colnames(tabla2)<-c("Variables",colnames(datosnumericos2))

kable(tabla2,format='markdown',digits=2,padding=0)
```

```
#Histograma Categoria pagina web y compra.

ggplot(datos, aes(Administrativa,fill=Compra)) + labs(x =
"Visitas Página Administrativa",y = "Frecuencia") +
geom_bar(position="dodge")

ggplot(datos, aes(Informativa, fill=Compra)) + labs(x =
"Visitas Página Informativa",y = "Frecuencia") +
geom_bar(position="dodge")

ggplot(datos, aes(Producto, fill=Compra)) + labs(x =
"Visitas Página del Producto",y = "Frecuencia") +
geom_bar(position="dodge")

boxplot(datos$Administrativa ~ datos$Compra, col =
"#7CAE00",

        main = "Visitas Página Administrativa y decision de
compra",xlab = "Decision de compra", ylab = "Visitas Página
Administrativa",ylim = c(0, 28))

boxplot(datos$Informativa ~ datos$Compra, col = "#F8766D",

        main = "Visitas Página Informativa y decision de
compra", xlab = "Decision de compra", ylab = "Visitas
Página Informativa",ylim = c(0, 25))

boxplot(datos$Producto ~ datos$Compra, col = "#C77CFF",

        main = "Visitas Página Producto y decision de
compra", xlab = "Decision de compra", ylab = "Visitas
Página Producto",ylim = c(0, 706))

# Histograma Dia Especial

Dia_Especial_Sin0<-filter(datos, Dia_Especial>0)  #Base de
datos
```

```
ggplot(Dia_Especial_Sin0, aes(Dia_Especial, fill=Compra)) +
labs(x = "Dia Especial",y = "Frecuencia") +
geom_bar(position="dodge")

#Histograma tasa salida

hist(datos$TasaSalidas, main="Histograma de la variable
Tasa de salidas",

     col="#C77CFF",xlab="Tasa de
Salidas",ylab="Frecuencia")

#Histograma Valor Página

hist(datos$ValorPagina, main="Histograma de la variable
Valor de la Página",

     col="#7CAE00",xlab="Valor de la
página",ylab="Frecuencia")

#Histograma Procentaje rebote

hist(datos$PorcentajeRebote, main="Histograma de la
variable Porcentaje de Rebote",

     col="#F8766D",xlab="Porcentaje de
rebote",ylab="Frecuencia")

#Estudio de usuarios más interesados

datosinteresados<-datos[datos$Duracion_Administrativa>10 &
datos$Administrativa>1 &

                        datos$Duracion_Informativa>10 &
datos$Informativa>1 &

                        datos$Duracion_Producto>10 &
datos$Producto>1,]

#Histograma Categoria pagina web y compra datos
interesados.

attach(datosinteresados)
```

```
ggplot(datosinteresados, aes(Administrativa,fill=Compra)) +
labs(x = "Visitas Página Administrativa",y = "Frecuencia")
+ geom_bar(position="dodge")

ggplot(datosinteresados, aes(Informativa, fill=Compra)) +
labs(x = "Visitas Página Informativa",y = "Frecuencia") +
geom_bar(position="dodge")

ggplot(datosinteresados, aes(Producto, fill=Compra)) +
labs(x = "Visitas Página del Producto",y = "Frecuencia") +
geom_bar(position="dodge")


boxplot(datos$Administrativa ~ datos$Compra, col =
"#7CAE00",
        main = "Visitas Página Administrativa y decision de
compra",xlab = "Decision de compra", ylab = "Visitas Página
Administrativa",ylim = c(0, 28))

boxplot(datos$Informativa ~ datos$Compra, col = "#F8766D",
        main = "Visitas Página Informativa y decision de
compra", xlab = "Decision de compra", ylab = "Visitas
Página Informativa",ylim = c(0, 25))

boxplot(datos$Producto ~ datos$Compra, col = "#C77CFF",
        main = "Visitas Página Producto y decision de
compra", xlab = "Decision de compra", ylab = "Visitas
Página Producto",ylim = c(0, 706))

# Histograma Dia Especial

Dia_Especial_Sin0<-filter(datos, Dia_Especial>0)  #Base de
datos
```

```
ggplot(Dia_Especial_Sin0, aes(Dia_Especial, fill=Compra)) +
labs(x = "Dia Especial",y = "Frecuencia") +
geom_bar(position="dodge")

#Histograma tasa salida

hist(datosinteresados$TasaSalidas, main="Histograma de la
variable Tasa de salidas",

     col="#C77CFF",xlab="Tasa de
Salidas",ylab="Frecuencia")

#Histograma Valor Página

hist(datosinteresados$ValorPagina, main="Histograma de la
variable Valor de la Página",

     col="#7CAE00",xlab="Valor de la
página",ylab="Frecuencia")

#Histograma Procentaje rebote

hist(datosinteresados$PorcentajeRebote, main="Histograma de
la variable Porcentaje de Rebote",

     col="#F8766D",xlab="Porcentaje de
rebote",ylab="Frecuencia")


#VARIABLES CATEGORICAS

#Tipo de visitor frente compra

attach(datosinteresados)

visitorcompra <- table(Compra, Visitor);

barplot(visitorcompra, main="Tipo de visitante según la
decisión de compra", xlab="Eje X", ylab="Eje Y",
col=c("#FF61C9","#E7861B"))

#Tipo de dia de la semana frente compra

finsemanacompra <- table(Compra,Fin_Semana);
```

```
barplot(finsemanacompra, main="Cantidad de compras
realizadas según el día de la semana",
        legend = rownames(finsemanacompra),
col=c("#FF61C9","#E7861B"))

#Mes frente compra

mescompra <- table(Compra,Mes);

barplot(mescompra, main="Cantidad de compras realizadas
según el mes",
        legend = rownames(mescompra),
col=c("#FF61C9","#E7861B"))

#Tipo de navegador frente compra

navegadorcompra <- table(Compra,Navegador);

barplot(navegadorcompra, main="Cantidad de compras
realizadas según el navegador",
        legend = rownames(navegadorcompra),
col=c("#FF61C9","#E7861B"))

#Region cliente frente compra

regioncompra <- table(Compra,Region);

barplot(regioncompra, main="Cantidad de compras realizadas
según la región",
        legend = rownames(regioncompra),
col=c("#FF61C9","#E7861B"))

#Tipo de grafico frente compra

graficocompra <- table(Compra,Grafico);

barplot(graficocompra, main="Cantidad de compras realizadas
según el tipo de gráfico",
        legend = rownames(graficocompra),
col=c("#FF61C9","#E7861B"))
```

```
#Operacion del sistema frente compra

operacioncompra <- table(Compra,Operacion_Sistema);

barplot(operacioncompra, main="Cantidad de compras
realizadas según la operacion de sistema",

        legend = rownames(operacioncompra),
col=c("#FF61C9","#E7861B"))

#Estudio variable y

diag<- c(length(Compra[Compra=="No compra"]),
length(Compra[Compra=="Si compra"]) )

lbls <- c("No Compra", " Compra")

pct <- round(diag/sum(diag)*100)

lbls <- paste(lbls, pct)

lbls <- paste(lbls,"%",sep="")

pie(diag,labels =lbls, main="Porcentaje de personas que
compran tras iniciar sesión",col=c("#FF61C9","#E7861B"))


##### COMPONENTES PRINCIPALES (VARIBALES NUMERICAS)#######

datosnumericos<-
data.frame(Administrativa,Duracion_Administrativa,Informati
va,Duracion_Informativa,
Producto,Duracion_Producto,PorcentajeRebote,TasaSalidas,Val
orPagina,Dia_Especial)

#Correlacciones

attach(datosnumericos)

Mat_R<-rcorr(as.matrix(datosnumericos))

corrplot(Mat_R$r,

         p.mat = Mat_R$r,
```

```
         type="upper",
         tl.col="black",
         tl.srt = 20,
         pch.col = "blue",
         insig = "p-value",
         sig.level = -1,
         col = terrain.colors(100))


attach(datosinteresados)
datosnumericoscp<-
data.frame(Administrativa,Duracion_Administrativa,Informati
va,Duracion_Informativa,
Producto,PorcentajeRebote,ValorPagina,Dia_Especial)


#PRUEBA ADECUACION COMPONENTES PRINCIPALES
#KMO
library(psych)
KMO(cor(datosnumericoscp))
#BARLET
bartlett.test((datosnumericoscp))


# ANALISIS EN CP DE LAS VARIABLES NUMERICAS
#Tipificamos los datoscp
G<-cov.wt(datosnumericoscp,method='ML')$center
V<-cov.wt(datosnumericoscp,method='ML')$cov
escala<-diag(V)
```

```
Z<-scale(datosnumericoscp,center=G,scale=sqrt(escala))

Z<-as.matrix(Z)

#Componentes principales

CP<-prcomp(Z, scale=TRUE)

summary(CP)

#Gráfico con % de varianza explicada.

library("factoextra")

fviz_screeplot(CP, main = "Varianzas",addlabels = TRUE,
xlab="Dimensiones",ylab="Porcentaje de varianza explicada")

### Sacamos coordenadas de 5cp (porque ofrece mejor
clasificacion)

coordenadas<-predict(CP)[,1:5] #Extraccion de las CP

### Sacamos coordenadas de 4cp(porque permite describir las
variables)

coordenadas1<-predict(CP)[,1:4]

#Grafico relacion variables y CP

fviz_pca_var(CP, col.var = "cos2", axes=c(1,2),repel=FALSE)

fviz_pca_var(CP, col.var = "cos2", axes=c(1,3),repel=FALSE)

fviz_pca_var(CP, col.var = "cos2", axes=c(1,4),repel=FALSE)

fviz_pca_var(CP, col.var = "cos2", axes=c(2,3),repel=FALSE)

fviz_pca_var(CP, col.var = "cos2", axes=c(2,4),repel=FALSE)

fviz_pca_var(CP, col.var = "cos2", axes=c(3,4),repel=FALSE)

#Grafico relacion observaciones y CP

fviz_pca_ind(CP,col.ind =
"cos2",geom="point",title='Usuarios -
Componentes',axes=c(1,2),
```

```
            gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"),
            repel = FALSE, xlab='CP1',ylab='CP2')
fviz_pca_ind(CP,col.ind =
"cos2",geom="point",title='Usuarios -
Componentes',axes=c(1,3),
            gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"),
            repel = FALSE, xlab='CP1',ylab='CP3')
fviz_pca_ind(CP,col.ind =
"cos2",geom="point",title='Usuarios -
Componentes',axes=c(1,4),
            gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"),
            repel = FALSE, xlab='CP1',ylab='CP4')
fviz_pca_ind(CP,col.ind =
"cos2",geom="point",title='Usuarios -
Componentes',axes=c(2,3),
            gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"),
            repel = FALSE, xlab='CP2',ylab='CP3')
fviz_pca_ind(CP,col.ind =
"cos2",geom="point",title='Usuarios -
Componentes',axes=c(2,4),
            gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"),
            repel = FALSE, xlab='CP2',ylab='CP4')
```

```
fviz_pca_ind(CP,col.ind =
"cos2",geom="point",title='Usuarios -
Componentes',axes=c(3,4),
             gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"),
             repel = FALSE, xlab='CP3',ylab='CP4')
#Rotacion
CP <- principal(Z, nfactors = 5,rotate="none")
CP
CP1 <- principal(Z, nfactors = 4,rotate="varimax")
CP1


MÉTODOS APLICADOS A LA BASE DE DATOS 4 CP + VAR.CATEGORICAS
# VARIABLES DUMMIES
datosdummy<-dummy_cols(datosinteresados,  select_columns =
c("Mes",
"Operacion_Sistema","Navegador","Region","Grafico","Visitor
","Fin_Semana","Compra"),
                       remove_first_dummy =
TRUE,remove_selected_columns = TRUE)


# METODO FORWARD
coordenadas2<-
data.frame(coordenadas1,datosdummy$Operacion_Sistema_3,

datosdummy$Navegador_3,datosdummy$Region_8)
```

```
################## ANALISIS CLUSTER ######################

wss <- (nrow(coordenadas2)-
1)*sum(apply(coordenadas2,2,var))

for (i in 2:15) wss[i] <-
sum(kmeans(coordenadas2,i)$withinss)

plot(1:15, wss, type="b", xlab="Número de Clusters",
ylab="Suma de

cuadrados entre grupos")

cluster<-kmeans(coordenadas2, 4)

library('cluster')

clusplot(coordenadas2, cluster$cluster , main='Solución
cluster',color=TRUE, shade=TRUE, labels=4, lines=0)


km12 <- kmeans(coordenadas2[,c(1,2)], 4)

fviz_cluster(km12, data =
coordenadas2[,c(1,2)],xlab='Acceso Página Administrativa y
producto',ylab='Valoración Web',  ellipse.type =
"convex",repel = FALSE,geom = "point",    show.clust.cent =
FALSE,shape=19,  palette = "Set2",ggtheme
=theme_minimal(),main='Cluster')

km13 <- kmeans(coordenadas2[,c(1,3)], 4)

fviz_cluster(km13, data =
coordenadas2[,c(1,3)],xlab='Acceso Página Administrativa y
producto',ylab='Acceso Página Informativa', ellipse.type =
"convex",repel = FALSE,geom = "point",show.clust.cent =
FALSE,shape=19,   palette = "Set2",ggtheme =
theme_minimal(),main='Cluster')

km14 <- kmeans(coordenadas2[,c(1,4)],4)
```

```
fviz_cluster(km14, data =
coordenadas2[,c(1,4)],xlab='Acceso Página Administrativa y
producto',ylab='Cercanía dia especial', ellipse.type =
"convex",repel = FALSE,geom = "point",show.clust.cent =
FALSE,shape=19, palette = "Set2",ggtheme =
theme_minimal(),main='Cluster')

km23 <- kmeans(coordenadas2[,c(2,3)], 4)

fviz_cluster(km23, data =
coordenadas2[,c(2,3)],xlab='Valoración Web',ylab='Acceso
Página Informativa', ellipse.type = "convex",repel =
FALSE,geom = "point",show.clust.cent = FALSE,shape=19,
palette = "Set2",ggtheme = theme_minimal(),main='Cluster')

km24 <- kmeans(coordenadas2[,c(2,4)], 4)

fviz_cluster(km24, data =
coordenadas2[,c(2,4)],xlab='Valoración Web',ylab='Cercanía
dia especial', ellipse.type = "convex",repel = FALSE,geom =
"point", show.clust.cent = FALSE,shape=19, palette =
"Set2",ggtheme = theme_minimal(),main='Cluster')

km34 <- kmeans(coordenadas2[,c(3,4)], 4)

fviz_cluster(km34, data =
coordenadas2[,c(3,4)],xlab='Acceso página
Informativa',ylab='Cercania dia especial', ellipse.type =
"convex",repel = FALSE,geom = "point",show.clust.cent =
FALSE,shape=19, palette = "Set2",ggtheme =
theme_minimal(),main='Cluster')


############## KNN COMPONENTES ####################

#no compra = 0 y compra = 1. VARIABLE Y

datosinteresados$Compra1 <-
factor(datosinteresados$Compra,labels = c("0", "1"))
```

```
#Dataframe de entrenamiento

train=rbind(coordenadas2[1:561,1:7])

#Dataframe para probar su clasificación

test = rbind(coordenadas2[562:1123,1:7])

#Clasificación de entrenamiento

cl = as.factor(datosinteresados$Compra1[1:561])

#clasificaciones con diversos tipos de k.

library(class)

k1=knn(train, test, cl, k = 1, prob=TRUE)

k3=knn(train, test, cl, k = 3, prob=TRUE)

k5=knn(train, test, cl, k = 5, prob=TRUE)

k7=knn(train, test, cl, k = 7, prob=TRUE)

k9=knn(train, test, cl, k = 9, prob=TRUE)

k11=knn(train, test, cl, k = 11, prob=TRUE)

# Elegimos los datos que queremos predecir

p<-datosinteresados$Compra1[562:1123]

tabla1<-table(p,k1)

acierto<-
((tabla1[1,1]+tabla1[2,2])/(tabla1[1,1]+tabla1[1,2]+tabla1[
2,1]+tabla1[2,2]))*100

acierto

tabla3<-table(p,k3)

acierto<-
((tabla3[1,1]+tabla3[2,2])/(tabla3[1,1]+tabla3[1,2]+tabla3[
2,1]+tabla3[2,2]))*100

acierto
```

```
tabla5<-table(p,k5)

acierto<-
((tabla5[1,1]+tabla5[2,2])/(tabla5[1,1]+tabla5[1,2]+tabla5[
2,1]+tabla5[2,2]))*100

acierto


tabla7<-table(p,k7)

acierto<-
((tabla7[1,1]+tabla7[2,2])/(tabla7[1,1]+tabla7[1,2]+tabla7[
2,1]+tabla7[2,2]))*100

acierto

tabla9<-table(p,k9)

acierto<-
((tabla9[1,1]+tabla9[2,2])/(tabla9[1,1]+tabla9[1,2]+tabla9[
2,1]+tabla9[2,2]))*100

acierto

tabla11<-table(p,k11)

acierto<-
((tabla11[1,1]+tabla11[2,2])/(tabla11[1,1]+tabla11[1,2]+tab
la11[2,1]+tabla11[2,2]))*100

acierto


##########SUPER VECTOR MACHINE FUNCION####################

#modelo svm con datos train

library(e1071)

library(caret)
```

```
#Modelo SVM con todos los datos

attach(datosinteresados)

ddata<-data.frame(coordenadas2,Compra1) #Tengo 9 var

x<-subset(ddata,select=-c(Compra1)) #Tengo 8 var

y<-ddata$Compra1

x1=x[1:561,]

y1=y[1:561]

svm_model<-svm(as.factor(y1)~
PC1+PC2+PC3+PC4+datosdummy.Operacion_Sistema_3+
datosdummy.Navegador_3+datosdummy.Navegador_3,kernel='linea
r',data=x1)

summary(svm_model)

#Ahora predecimos el resto

x2=x[562:1123,]

clasificacion<-predict(svm_model,x2)

clasificacion

#Tabla de confusiones

cont<-ddata[562:1123,8]

tabla1<-table(cont,clasificacion)

error<-((tabla1[1,2]+tabla1[2,1])/1123)*100

tabla1

100-error


################## FUNCION SVM MANUAL #################


library(readr)
```

```
library(lpSolve)
datos<-transform(ddata, yi=ifelse(Compra1==0,-1,1))
datos<-subset(datos,select=-c(Compra1))
datosmat=datos[1:7]


p=ncol(datosmat)
n=nrow(datosmat)
fo=rep(0,(2*p)+n+1)
fo[1:((2*p)+n)]=1
eror=diag(n)
A=matrix(0,nrow = n,ncol = 2*p)
columna=1
for(i in 1:(2*p)){
  if(i%%2!=0){
    A[,i]=datosmat[,columna]
    columna=columna+1
  }
}
columna=1
for(i in 1:(2*p)){
  if(i%%2==0){
    A[,i]=-datosmat[,columna]
    columna=columna+1
  }
}
```

```
A=datos$yi*A

b=rep(1,n)

A=cbind(A,eror,datos$yi)

signo=rep('>=',n)

sol=lp('min',fo,A,signo,b)

solucion=sol$solution

b=tail(solucion,1)

Ws=solucion[1:(2*p)]

parametros=rep(0,3)

vectorp=seq(from=2,to=(2*p),by=2)

vectori=seq(from=1,to=(2*p),by=2)

parametros=solucion[vectori]-solucion[vectorp]

datosmat$svm=1

for(i in 1:n){

  datosmat[i,p+1]=rowSums(datosmat[i,1:p]*parametros)

}

datosmat<-transform(datosmat, clasifica=ifelse(svm<0,0,1))

comprobacion<-data.frame(datosmat,datos$yi)

tabla2<-table(datos$yi,datosmat$clasifica,
dnn=c("real","predicho"))

tabla2

error2<-((tabla2[1,2]+tabla2[2,1])/1123)*100

error2
```

```
MÉTODOS APLICADOS A LA BASE DE DATOS DE 5CP + VAR.
CATEGÓRICAS

# METODO FORWARD

coordenadas3<-
data.frame(coordenadas,datosdummy$Grafico_6,datosdummy$Regi
on_3)


################## PREDICCION KNN ########################

#no compra = 0 y compra = 1. VARIABLE Y

datosinteresados$Compra1 <-
factor(datosinteresados$Compra,labels = c("0", "1"))

#Dataframe de entrenamiento

train=rbind(coordenadas3[1:561,1:7])

#Dataframe para probar su clasificación

test = rbind(coordenadas3[562:1123,1:7])

#Clasificación de entrenamiento

cl = as.factor(datosinteresados$Compra1[1:561])

#clasificaciones con diversos tipos de k.

library(class)

k1=knn(train, test, cl, k = 1, prob=TRUE)

k3=knn(train, test, cl, k = 3, prob=TRUE)

k5=knn(train, test, cl, k = 5, prob=TRUE)

k7=knn(train, test, cl, k = 7, prob=TRUE)

k9=knn(train, test, cl, k = 9, prob=TRUE)

k11=knn(train, test, cl, k = 11, prob=TRUE)

# Elegimos los datos que queremos predecir
```

```
p<-datosinteresados$Compra1[562:1123]

tabla1<-table(p,k1)

acierto<-
((tabla1[1,1]+tabla1[2,2])/(tabla1[1,1]+tabla1[1,2]+tabla1[
2,1]+tabla1[2,2]))*100

acierto


tabla3<-table(p,k3)

acierto<-
((tabla3[1,1]+tabla3[2,2])/(tabla3[1,1]+tabla3[1,2]+tabla3[
2,1]+tabla3[2,2]))*100

acierto

tabla5<-table(p,k5)

acierto<-
((tabla5[1,1]+tabla5[2,2])/(tabla5[1,1]+tabla5[1,2]+tabla5[
2,1]+tabla5[2,2]))*100

acierto

tabla7<-table(p,k7)

acierto<-
((tabla7[1,1]+tabla7[2,2])/(tabla7[1,1]+tabla7[1,2]+tabla7[
2,1]+tabla7[2,2]))*100

acierto

tabla9<-table(p,k9)

acierto<-
((tabla9[1,1]+tabla9[2,2])/(tabla9[1,1]+tabla9[1,2]+tabla9[
2,1]+tabla9[2,2]))*100

acierto

tabla11<-table(p,k11)
```

```
acierto<-
((tabla11[1,1]+tabla11[2,2])/(tabla11[1,1]+tabla11[1,2]+tab
la11[2,1]+tabla11[2,2]))*100

acierto



##########SUPER VECTOR MACHINE FUNCION#####################

#modelo svm con datos train

library(e1071)

library(caret)

#Modelo SVM con todos los datos

attach(datosinteresados)

ddata<-data.frame(coordenadas3,Compra1) #Tengo 9 var

x<-subset(ddata,select=-c(Compra1)) #Tengo 8 var

y<-ddata$Compra1

x1=x[1:561,]

y1=y[1:561]

svm_model<-svm(as.factor(y1)~
PC1+PC2+PC3+PC4+PC5+datosdummy.Grafico_6+datosdummy.Region_
3,kernel='linear',data=x1)

summary(svm_model)

#Ahora predecimos el resto

x2=x[562:1123,]

clasificacion<-predict(svm_model,x2)

clasificacion

#Tabla de confusiones
```

```
cont<-ddata[562:1123,8]

tabla1<-table(cont,clasificacion)

error<-((tabla1[1,2]+tabla1[2,1])/1123)*100

tabla1

100-error

################## FUNCION SVM MANUAL ####################

library(readr)

library(lpSolve)

datos<-transform(ddata, yi=ifelse(Compra1==0,-1,1))

datos<-subset(datos,select=-c(Compra1))

datosmat=datos[1:7]

p=ncol(datosmat)

n=nrow(datosmat)

fo=rep(0,(2*p)+n+1)

fo[1:((2*p)+n)]=1

eror=diag(n)

A=matrix(0,nrow = n,ncol = 2*p)

columna=1

for(i in 1:(2*p)){

  if(i%%2!=0){

    A[,i]=datosmat[,columna]

    columna=columna+1

  }

}

columna=1
```

```
for(i in 1:(2*p)){
  if(i%%2==0){
    A[,i]=-datosmat[,columna]
    columna=columna+1
  }
}
A=datos$yi*A
b=rep(1,n)
A=cbind(A,eror,datos$yi)
signo=rep('>=',n)
sol=lp('min',fo,A,signo,b)
solucion=sol$solution
b=tail(solucion,1)
Ws=solucion[1:(2*p)]
parametros=rep(0,3)
vectorp=seq(from=2,to=(2*p),by=2)
vectori=seq(from=1,to=(2*p),by=2)
parametrossolucion[vectori]-solucion[vectorp]
datosmat$svm=1
for(i in 1:n){
  datosmat[i,p+1]=rowSums(datosmat[i,1:p]*parametros)
}
datosmat<-transform(datosmat, clasifica=ifelse(svm<0,0,1))
comprobacion<-data.frame(datosmat,datos$yi)
```

```
tabla2<-table(datos$yi,datosmat$clasifica,
dnn=c("real","predicho"))

tabla2

error2<-((tabla2[1,2]+tabla2[2,1])/1123)*100

error2

100-error2
```