www.cya.unam.mx/index.php/cya

Contaduría y Administración

Contaduría y Administración 68 (1), 2023, 289-323

La popularidad de las marcas y su valor económico en el marco de las finanzas corporativas: un análisis de aprendizaje máquina

The popularity of brands and their economic value in the framework of corporate finance: A machine learning analysis

Víctor Miguel Morales González, Griselda Dávila Aragón*, Francisco Ortiz Arango

Universidad Panamericana, México

Recibido el 17 de mayo de 2022; aceptado el 10 de noviembre de 2022
Disponible en Internet el: 14 de noviembre de 2022

Resumen

A lo largo del tiempo, la marca ha tomado un papel significativo en el ámbito empresarial, la percepción de la imagen comercial y el valor agregado. Este estudio está enfocado en explorar los componentes del concepto del valor de marca a partir de un diagnóstico y técnicas de aprendizaje máquina, para desarrollar una serie de modelos asociados a las dimensiones del valor de marca percibido desde un concepto más actual de la popularidad. La metodología de aprendizaje máquina, prioriza la predicción frente a la inferencia. No impone una especificación ni una teoría, a diferencia de la estadística clásica, donde se requiere especificar un modelo; esto representa una forma dinámica alternativa para entender cómo uno de los recursos más importantes de las empresas en el mercado está presente, lo que sin duda repercute en la gestión financiera y de riesgos de la empresa. Los resultados obtenidos mediante tres técnicas diferentes de aprendizaje máquina, muestran que las once variables propuestas en el estudio influyen positivamente con diferente intensidad en la popularidad de la marca.

* Autor para correspondencia
Correo electrónico: gdavila@up.edu.mx (G. Dávila Aragón).
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

http://dx.doi.org/10.22201/fca.24488410e.2023.4665

Código JEL: C19, C69, G40, G41
Palabras clave: popularidad; marcas; aprendizaje máquina; redes sociales

Abstract

Over time, the brand has played a significant role in the business sphere, the perception of commercial image, and added value. This study is focused on exploring the components of brand value from a diagnosis and machine learning techniques to develop a series of models associated with the dimensions of perceived brand value from a more current concept of popularity. The machine learning methodology prioritizes prediction over inference. Unlike classical statistics, it does not impose a specification or a theory, where a model is required to be specified; this represents an alternative dynamic way to understand how one of the most critical resources of companies is present in the market, which undoubtedly has repercussions on the financial and risk management of the company. The results obtained through three different machine learning techniques show that the eleven variables proposed in the study positively influence brand popularity with different intensities.

JEL Code: C19, C69, G40, G41
Keywords: popularity; brands; machine learning; social networks

Introducción

En un mundo como el actual las redes sociales pueden ser consideradas como una importante fuente de indicadores de opinión, comportamiento de consumo y prestigio mediático que tienen algunas marcas. Esta situación refleja una forma de percepción de la imagen de las marcas y de su valor mismo. El modelo de comunicación y marketing se enfoca en desarrollar la marca para conseguir máxima visibilidad y beneficio en los portales de venta y así alcanzar mayores volúmenes y con ello generar importantes beneficios económicos en corto plazo (Pérez Curiel y Sanz-Marcos, 2019).

En la antigüedad la apreciación de una marca era percibida de boca en boca por los clientes y la popularidad era evidenciada por la cantidad de ventas, estos conceptos han cambiado totalmente (Cuellar, 2019). Actualmente en el medio de las redes sociales, los consumidores emiten los juicios acerca de los productos y servicios por medio de “likes” o comentarios que se difunden, mientras que las empresas toman estos elementos en tiempo real (Cuellar, 2019).

A lo largo del tiempo, las marcas han tomado un papel protagónico en el ámbito empresarial, constituyen un elemento importante en el proceder comercial y en la generación de valor. Por ello, ya no es adecuado referir a la marca tan solo como un símbolo o imagen del producto o servicio en el mercado, sino como el valor que implica elementos distintivos que reflejan de alguna forma el reconocimiento de la marca, la lealtad de los clientes, la calidad percibida y la popularidad (Horna & Prado, 2015). De aquí la importancia de conocer el verdadero significado del valor de marca, como un indicador que mide la

percepción del consumidor frente a la competencia, y de permitir direccionar cada una de las estrategias y toma de decisiones, al cumplimiento de las necesidades y satisfacción de los clientes y con ello ver el nivel de popularidad que genera una marca por tales efectos (Horna & Prado, 2015).

El presente estudio está enfocado en explorar las componentes del concepto del valor de marca a partir de un diagnóstico y técnicas de aprendizaje máquina (Sneider Castillo & Ortegón Cortazar, 2016), para desarrollar una serie de modelos asociados a las dimensiones del valor de marca percibida desde algunos factores identificados en las redes sociales (Facebook, Instagram y Twitter), basados en lo que pudiera ser considerado como popularidad.

Hablar de la popularidad de las marcas en el medio de las redes sociales se ha considerado en la mayor parte de la literatura, como una forma de relación de las marcas con el mercado, su pertinencia tiene que ver con aspectos mercadológicos. Para Aggrawal, Ahluwalia, Khurana & Arora (2017), el marketing online es una de las mejores prácticas que se utiliza para establecer una marca y aumentar su popularidad. Los anuncios son una de las mejores maneras para mostrar el producto/servicio de la empresa y dar lugar a una valiosa línea estratégica de mercadeo. Publicar anuncios en las llamadas páginas web utilitarias ayuda a maximizar el alcance de la marca y obtener una mejor retroalimentación. Estos autores han propuesto un marco de referencia que permite analizar la popularidad de marca en términos de la relación de su presencia en páginas web y en las redes sociales. Para ello utilizan algoritmos de análisis de textos, de sentimientos en los mensajes y la construcción de redes. La popularidad de la marca se expone en términos de frecuencia de aparición en las páginas web y del análisis de sentimientos en los textos de las redes sociales.

En un sentido similar, Kim, Moon & Iacobucci (2019), reportan un estudio realizado para proponer una herramienta de gestión mercadológica de las marcas globales basada en la su popularidad por país y las actividades de los consumidores en las redes sociales. En su análisis usan como variables de estudio, las ventas, las utilidades y el valor de las marcas, adicionales a la medición de la popularidad. Aquí se entiende la popularidad como una forma de receptividad y preferencia de los consumidores por las marcas a través de las redes sociales, aspecto que se contrasta lo que en el pasado se ha asociado al grado en que la población en general busca y compra un producto o servicio por su marca.

También se han realizado estudios empíricos en relación con el concepto de lealtad de las marcas en términos de la recompra de los productos. Por ejemplo, en el trabajo descrito por Sriram, Prabhub & Bhat (2019), se analiza la lealtad a la marca en relación con el deseo de recompra de teléfonos celulares, teniendo como marco de análisis la norma ISO 9241(1992/2001), enfocada a la calidad en usabilidad y ergonomía tanto de hardware como de software. En este trabajo se aplican técnicas estadísticas tradicionales a los resultados de encuestas entre usuarios de teléfonos celulares respecto de la eficiencia, eficacia y satisfacción de los productos. Este trabajo no solo se enfoca a términos mercadológicos sino en

la llamada usabilidad de los productos y consecuentemente a una forma de lealtad hacia la marca. Estos aspectos son importantes desde la perspectiva comercial y consecuentemente del valor de la marca.

Con el objetivo de analizar la popularidad de las marcas publicadas en redes sociales, Robson, Banerjee & Kaur (2022) muestran una revisión de literatura de años recientes en la cual identifican hasta 22 conceptos relacionados con popularidad asociadas a las publicaciones de marcas, llegando a una de sus conclusiones de que pocos trabajos han examinado las interacciones entre estos conceptos. Aun cuando consideran que el estudio de la interacción de estos conceptos de popularidad de las marcas es de importancia para propósitos mercadológicos, no consideran estas interacciones con el valor de las marcas como un activo de las empresas. Se reconoce que la dimensión mercadológica solo representa una parte de la gerencia estratégica de la empresa.

En sentido similar a lo discutido por Kim, Moon & Iacobucci (2019) y Robson, Banerjee & Kaur (2022), en el presente estudio se propone partir de 11 variables asociadas a un concepto de popularidad, no solo enfocadas a aspectos mercadológicos, sino a aquellos aspectos que se relacionan con la gerencia y el valor económico de las marcas. Estas variables han sido planteadas considerando dos criterios importantes: los considerados a la fecha por empresas consultoras internacionales en la valoración de marcas globales, como Interbrand (2020); y teniendo como marco de análisis los elementos sugeridos por las normas ISO 10668 (2010) y la ISO 20671 (2019), en materia de valuación y evaluación de marcas, respectivamente. Para realizar este análisis se propone un procedimiento de aprendizaje máquina, que resulta ser pertinente si se consideran los alcances de los análisis de datos y conclusiones a las que han llegado estudios previos, como los realizados y reportados por Kim, Moon & Iacobucci (2019) y Robson, Banerjee &Kaur (2022), respecto a resultados contradictorios en los análisis realizados mediante técnicas estadísticas tradicionales. Por ello se hace pertinente incursionar en el uso de herramientas como big data analitics, herramientas de ciencia de datos y aprendizaje máquina.

De acuerdo con Crespo (2013), el aprendizaje máquina o (machine learning) es un conjunto de técnicas que tienden a mejorar el comportamiento o desempeño de un sistema a través de la experiencia adquirida. Es una disciplina del campo de la Inteligencia Artificial (Rich, Knight, Calero & Bodega, 1994), que, mediante algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo) (Espino Timón, 2017).

La metodología de aprendizaje máquina, prioriza la predicción frente a la inferencia. No impone una especificación ni una teoría, a diferencia de la estadística clásica donde se especifica un modelo (Viera, 2017). El modelo estadístico busca, con bases teóricas, encontrar la relación entre las variables para identificar variables explicativas, dependencia o independencia entre ellas y el sentido de su relación, así como probar hipótesis y realizar inferencias. En tanto que la metodología de machine learning permite que los datos “hablen” o se expresen por sí mismos, prioriza la importancia de la predicción frente a la

inferencia, mediante un algoritmo que encuentre la relación input-output con lo que busca que el modelo replique los datos utilizando la herramienta de validación cruzada extramuestral (Mergel, 1998). Con base en lo anteriormente descrito, se demostrará que las variables propuestas en el estudio influyen en la popularidad de la marca.

El presente trabajo, se desarrolla de la siguiente manera: En la siguiente sección se describen los tres algoritmos de aprendizaje máquina que se utilizan en esta investigación, así como los códigos e instrucciones en lenguaje R empleadas para realizar los cálculos de cada algoritmo; también se presenta una breve justificación del uso del lenguaje R. Posteriormente en la sección tres, se describen las 11 variables que conformaron la base de datos utilizada, donde tres de ellas son cualitativas y 8 cuantitativas. Se concluye dicha sección con el análisis de los resultados obtenidos con los tres algoritmos. Posteriormente se presentan las conclusiones, que pueden sintetizarse estableciendo que los resultados obtenidos mediante las tres técnicas de aprendizaje máquina empleadas, demuestran que las once variables propuestas en el estudio influyen positivamente, aunque con diferente intensidad en la popularidad de la marca. Finalmente se listan las referencias bibliográficas.

Metodología

De acuerdo con Carta, Podda, Recupero, Saia & Usai (2020) existen varios tipos de algoritmos de aprendizaje máquina o machine learning, cuya elección o apropiabilidad depende de la estrategia objetivo, el tipo de datos de entrada/salida involucrados y el tipo de problema a ser analizado. De este modo, en la literatura se proponen diversos tipos de algoritmos de aprendizaje máquina, tales como: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi-supervisado, aprendizaje por refuerzo, autoaprendizaje, aprendizaje de características, etc. (Nieto Jeux, 2021), siendo los algoritmos de aprendizaje supervisados, los no supervisados y los semi-supervisados, los más utilizados.

De acuerdo con Rojas (2020), en el aprendizaje supervisado se aprenden funciones, relaciones que asocian entradas con salidas, se ajustan a un conjunto de ejemplos de los que se conocen la relación entre la entrada y la salida deseada. En lo referente a los modelos de aprendizaje no supervisado, son aquellos en los que no hay interés en ajustar las entradas y salidas, sino que se trata de aumentar el conocimiento estructural de los datos disponibles. Para Ni (2022), el tercer algoritmo más importante es el Semi-supervisado, el cual combina algunas propiedades los otros dos tipos de algoritmos anteriormente descritos. Consiste en manejar conjuntos de datos donde se incorporan atributos adicionales en la variable objetivo y en otras variables que se consideren convenientes.

Modelos empleados para el aprendizaje máquina

En el presente trabajo se utilizó el programa RStudio, para generar modelos de aprendizaje máquina, tales como K-nn, Arboles de clasificación y Naive Bayes. La razón fundamental para utilizar el lenguaje R a través de la suite RStudio, es que ésta constituye un entorno y lenguaje de programación empleado primordialmente para efectuar análisis estadístico de datos y construcción de gráficos. Dada su calidad, versatilidad y la existencia de un sinnúmero de librerías de código libre, donde se han programado suficientes algoritmos útiles para el desarrollo de machine learning, por esta razón su uso se ha convertido ya en un estándar para este tipo de análisis, y en general para aplicaciones estadísticas y econométricas. Actualmente R es ampliamente usado en áreas como la bioestadística, el data mining, la econometría, la visualización de datos, etc. (Fernández Lizana, 2020).

Conforme con lo planteado por Sasikala, Biju & Prashanth (2017), uno de los objetivos de este trabajo es decidir cuál de los modelos de aprendizaje máquina utilizados es más eficiente para medir la popularidad de las marcas con base en sus variables financieras, económicas y tecnológicas. Se utilizarán los tres algoritmos mencionados anteriormente para generar modelos predictivos.

La forma de aplicar los algoritmos mencionados es mediante el uso de modelos propios de la ciencia de datos, entre los cuales destacan: el Método K-nn (K nearest neighbors); árboles de clasificación y el modelo Naïve Bayes, entre otros. Debido a que en este trabajo se utilizó una variable de respuesta binaria, se ha considerado que el empleo de estos modelos puede ser apropiado para el análisis de datos (analytics), normalmente empleados en aplicaciones económico-empresariales (Brunton & Kutz, 2022).

El método K-nn (K nearest neighbors), (Fix y Hodges, 1989) y (Dudani, 1976) es un método de clasificación supervisada (aprendizaje, estimación basada en un conjunto de entrenamiento y prototipos), que permite estimar la función de densidad de probabilidad. Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j[1] a partir de la información proporcionada por el conjunto de prototipos o ejemplos (Dudani, 1976). En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

Este es un método que se ha utilizado en el reconocimiento de patrones durante los últimos 40 años. Una de sus ventajas es que se ha aplicado a la categorización en estrategias de investigación, donde: en primera instancia se calcula la distancia entre la nueva muestra y la muestra de entrenamiento; posteriormente y de acuerdo con la categoría a la que pertenece el vecino, se determina la nueva muestra

[1] Es la asignación que se le da a una etiqueta de clase sobre la base que se representa con más frecuencia alrededor de un punto de datos determinado (Raschka, 2018) Fuente: https://sebastianraschka.com/pdf/lecture-notes/stat479fs18/02_knn_notes.pdf

y se verifica si todos pertenecen a la misma categoría (Wang, 2019) y (Zaki & Meira, 2020). La ecuación fundamental de este método es la siguiente:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(ar(x_i) - ar(x_j)\right)^2}$$

Donde:

$d(x_i,x_j)$: es la distancia euclidiana del elemento x_i con relación a x_j

n: es el número de atributos

$ar(x_i)$: representa el i-ésimo atributo de cada elemento

De acuerdo con Zaki & Meira (2020), el método consiste en promediar el valor de Y en las K observaciones más cercanas al punto x_0 para obtener una estimación de la variable respuesta asociada $\hat{Y}_0$.

$$\hat{Y}_0 = \sum_{x_i \in Nk(x_0)} y_i$$

Dónde: $Nk(x_0)$ es la vecindad de x_0 definida por los k puntos más cercanos en los datos de entrenamiento.

En el caso de clasificación con una variable respuesta categórica binaria (popularidad), es decir Y = 0, 1, el promedio de las k observaciones más cercanas representará la estimación de la probabilidad de que el punto x tome el valor Y = 1. De forma general:

$$Pr(Y_0 = j|X = x_0) = \frac{1}{k} \sum_{x_i \in Nk(x0)} (y_i = j)$$

Se trata de un método de estimación local, que no considera supuestos rigurosos sobre los datos, y está basado en el mejor estimador empírico para una función de pérdida cuadrática.

$$\widehat{Y} = E(Y|X)$$

Este clasificador siempre elegirá la categoría para la cual la probabilidad de clasificación se maximiza.

$$\max P(Y = j|X = x_0)$$

El algoritmo construido utilizado en el programa RStudio es el siguiente:

Se genera una semilla de datos aleatorios, de modo que los resultados posteriores no se vean modificados en el proceso. De acuerdo con Camaño & Goyeneche (2011) y Casajús (2022), se entiende

por semilla al valor inicial que se introduce en el programa de computación para que el algoritmo utilizado genere la serie de números aleatorios. Como los números aleatorios que se utilizarán para hacer el ordenamiento inicial de los candidatos son generados por un programa de computación, su aleatoriedad depende justamente que el número que se la da para el arranque (la semilla) sea un número aleatorio

De acuerdo con lo anterior, se parte de la siguiente instrucción en R:

```
set.seed(128)
```

Para ser aplicado este método, la base de datos descrita en la Sección 3. Selección de datos y resultados fue fraccionada para las fases de entrenamiento y prueba con el fin de calibrar la funcionalidad del modelo. Para lo cual, la prueba se realiza con un conjunto de datos que se divide en dos partes: datos de entrenamiento y datos de prueba o test. Los datos de entrenamiento o «training data» se usan para dar instrucciones de entrenamiento al modelo, mientras que los datos de prueba generan las predicciones con base en el modelo y permite comparar los valores generados con los valores reales de la muestra. Normalmente el conjunto de datos se suele repartir en un 70% de datos de entrenamiento y un 30% de datos de prueba (Vabalas, Gowen, Poliakoff, & Casson, 2019).

Conforme a lo anterior, en la primera fase se selecciona el 70% de los datos para entrenar al algoritmo y darle información para que encuentre los patrones necesarios y pueda hacer predicciones. En la fase de prueba el resto de los datos (30%) se usan para evaluar si la respuesta del modelo es confiable como modelo predictor. Posteriormente se incorporó el siguiente código para la elaboración del modelo K-nn. Donde su relevancia radica en el hecho de poder describir el comportamiento de la variable respuesta “Popularidad”

```
SP_knnEntrenado <- train(Popularidad ~ .,
                         data = SP_entrena,
                         method = "knn",
                         tuneLength = 20)
class (SP_knnEntrenado)
SP_knnEntrenado
plot(SP_knnEntrenado)
```

En el caso del modelo de árbol de clasificación, la variable dependiente es categórica y el valor en el nodo terminal es igual a la moda de las observaciones del conjunto de entrenamiento que han “caído” en esa región (Merino & Chacón, 2017). Los árboles de decisión o de clasificación surgieron en el ámbito del aprendizaje máquina y de la Inteligencia Artificial (Román & Lévy, 2003).

De acuerdo con Beltrán & Barbona (2021), se trata de un método no-paramétrico de segmentación binaria y se construye dividiendo repetidamente los datos. Los datos son clasificados en grupos mutuamente excluyentes. El algoritmo comienza con un nodo inicial, el cual se divide en dos sub-

grupos o sub-nodos, finalmente se elige una variable y se determina el punto de corte de modo que las unidades pertenecientes a cada nuevo grupo definido sean lo más homogéneas posible. La ecuación fundamental de este método es la siguiente (Zaki & Meira, 2020):

$$i(t) = \sum_{j=1}^{k} p(j/t)\, ln\, p(j/t)$$

Donde:

i(t): conocido como la impureza de Gini, en términos de cómo un elemento elegido aleatoriamente es etiquetado incorrectamente;

p(j / t): es la probabilidad de un error en la categorización de un elemento

De acuerdo con Zaki & Meira (2020), al ser la variable respuesta (popularidad) cualitativa, existen varias alternativas con el objetivo de encontrar nodos homogéneos. Las más empleadas son:

Classification Error Rate:

Se define como la proporción de observaciones que no pertenecen a la clase más común en el nodo.

$$Em = 1 - max\, k(\widehat{p_{mk}})$$

Donde $\widehat{p_{mk}}$) representa la proporción de observaciones del nodo m que pertenecen a la clase k. A pesar de la sencillez de esta medida, no es suficientemente sensible para crear modelos.

Gini Index

Es una medida de la varianza total en el conjunto de las K clases del nodo m. Se considera una medida de homogeneidad del nodo.

$$G_m = \sum \widehat{p_{mk}}\,(1 - \widehat{p_{mk}})$$

Cuando $\widehat{p_{mk}}$ es cercano a 0 o a 1 el nodo contiene mayoritariamente observaciones de una clase. Como consecuencia, cuanto mayor sea la homogeneidad del nodo, menor es el valor del índice Gini G.

Chi-Square

Esta aproximación consiste en identificar si existe una diferencia significativa entre los nodos particulares y el nodo general, es decir, si hay evidencias de que la división consigue una mejora. Para ello, se aplica una prueba estadística "chi-square goodness of fit" empleando como distribución esperada H_0 la

frecuencia de cada clase en el nodo general. Cuanto mayor el estadístico X^2, mayor la evidencia estadística de que existe una diferencia.

$$X^2 = \sum k \frac{(observado\ k - esperado\ k)^2}{esperado\ k}$$

Para cada posible división se calcula el valor de la medida en cada uno de los dos nodos resultantes. Se suman los dos valores ponderando cada uno por la fracción de observaciones que contiene cada nodo.

$$\left(\frac{n\ observaciones\ nodo\ A}{n\ observaciones\ totales}\right) * pureza\ A + \left(\frac{n\ observaciones\ nodo\ B}{n\ observaciones\ totales}\right) * pureza\ B$$

La división con menor o mayor valor (dependiendo de la medida empleada) se selecciona como división óptima. Entendiendo pureza como la máxima probabilidad de cada nodo, de acuerdo con la entropía o índice de Gini. Consecuentemente, la impureza suele medirse como la mínima probabilidad de ocurrencia de cada nodo (Zaki & Meira, 2020).

El algoritmo construido utilizado en el programa RStudio es el siguiente:

Partiendo de que se genera una semilla dado a que es un proceso aleatorio y se quiere que los resultados posteriores no se vean modificados en dicho proceso, utilizando la siguiente instrucción:

```
set.seed(123)
arbol_clasificacion <- tree(
  formula = Popularidad ~ .,
  data    = SP_entrena,
  minsize = 10)
summary(arbol_clasificacion)
```

Finalmente, el modelo Naïve Bayes, el cual es uno de los clasificadores más utilizados por su simplicidad y rapidez. Consiste en una técnica de clasificación y predicción supervisada que permite construir modelos que predicen la probabilidad de posibles resultados, con base en el Teorema de Bayes, también conocido como teorema de la probabilidad condicionada (Webb, Keogh & Miikkulainen, 2010). Una limitación es que la suposición de independencia de atributos en ocasiones no corresponde a la realidad. Sin embargo, se ha sugerido que ante tales limitantes su impacto puede ser menor debido a que la clasificación binaria es considerada como una función de la estimación de la probabilidad (Yang & Webb, 2002) y (Rrmoku, Selimi & Ahmedi, 2022). La ecuación fundamental de este método es la siguiente (Zaki & Meira, 2020):

$$p(C = c|\, X = x) = \frac{p(C = c)p(X = x|\, C = c)}{p(X = x)}$$

Donde:

p(C=c|X=x): se refiere a la probabilidad condicional a posteriori

C: se refiere a la variable dependiente;

X: es la variable condicional o atributo

De acuerdo con Zaki & Meira (2020), se pueden estimar las probabilidades a posteriori de cualquier hipótesis consistente con el conjunto de datos de entrenamiento para así escoger la hipótesis más probable.

Dado un ejemplo x representado por k valores, el clasificador Naïve Bayes se basa en encontrar la hipótesis más probable que describa a ese ejemplo.

Si la descripción de ese ejemplo viene dada por los valores $<a_1, a_2, ..., a_n>$, siendo a_i los posibles atributos de la variable objetivo, la hipótesis más probable será aquella que cumpla:

$$v_{MAP} = \text{argmax}_{vj \epsilon V}\, P(v_j | a_1, \dots, a_n)$$

Es decir, la probabilidad de los valores conocidos que describen a ese ejemplo, pertenezcan a la clase v_j , donde v_j es el valor de la funcion de clasificacion f(x) en el conjunto finito V.

$$v_{MAP} = \text{argmax}_{vj \epsilon V}\, \frac{P(a_1, \dots, a_n | v_j) p(v_j)}{P(a_1, \dots, a_n)}$$

Se puede estimar $P(v_j)$ contando las veces que aparece el ejemplo v_j en el conjunto de entrenamiento y dividendo por el número total de ejemplos que forman ese conjunto.

El algoritmo construido utilizado en el programa RStudio es el siguiente:

```
Partición <- createDataPartition(y = datos_train_prep$Popularidad, p = 0.7, list = FALSE)
SP_entrena <- datos_train_prep[Particion,]
SP_test <- datos_train_prep[-Particion,]
```

Como ya se mencionó el objetivo de este trabajo es identificar cual es el modelo predictor que resulte ser más eficiente para evitar los dos errores estadísticos (error tipo 1 y error tipo 2), al analizar como variable respuesta la popularidad de las marcas, considerando sus variables financieras, económicas y tecnológicas (presencia en redes sociales). Particularmente se trata de encontrar el modelo que ofrezca obtener la mayor proporción de verdaderos positivos, en forma del llamado “accuracy” es decir el mejor modelo predictor (Fleuren, et al; 2020).

Selección de datos y resultados

La base de datos está compuesta por datos de las 50 marcas más valiosas del mundo, dadas por la consultora Interbrand en 2020 (Interbrand, 2020). Las variables utilizadas en la base de datos se muestran en la Tabla 1:

Tabla 1
Variables propuestas para la construcción de la base de datos muestra

	Variable	Descripción
1.	Marcas	Las más valiosas del mundo dadas por la consultora Interbrand en 2020
2.	Valor	El valor de la marca dada por la consultora Interbrand en 2020 dada por su metodología de valuación ante factores financieros y fuerza de la marca
3.	Sector	El sector que pertenecen las marcas dadas por Tecnología, Bebidas, Automotriz, Restaurante, Media, Servicios de negocios, Artículos deportivos, Lujo, Servicios financieros, Logística, Retail, Diversificado, Alcohol, Consumo inmediato, Vestido
4.	País	El lugar de origen de las marcas tales como EUA, Corea del Sur, Japón, Alemania, Francia, Suecia, Italia, España, Suiza
5.	Ventas	El total de ventas dadas en 2020 por parte de las marcas y reportadas por la base de datos Economatica
6.	Utilidades	El total de las Utilidades dadas en 2020 por parte de las marcas y reportadas por la base de datos Economatica
7.	Precio de Acción	El precio de cotización de la acción a fecha de cierre de 2020 por parte de las marcas y reportadas por la base de datos Economatica
8.	Facebook	El número de likes dados en cada página de la respectiva marca dados por la plataforma Facebook by Meta
9.	Instagram	El número de likes dados en cada página de la respectiva marca dados por la plataforma Instagram by Meta
10.	Twitter	El número de likes dados en cada página de la respectiva marca dados por la plataforma Twitter
11.	Popularidad	La popularidad siendo la variable respuesta, se generó a partir del número de ventas en promedio como también el número de likes de las plataformas y con ello se asignó el nivel de popularidad de cada marca.

Fuente: Elaboración propia

Como se puede apreciar en la Tabla 1, la base de datos está compuesta por 11 variables, las cuales para efectos de este trabajo se refieren a un concepto de popularidad, como ha sido discutido en la literatura. Es importante señalar que a diferencia del concepto de popularidad establecido en la literatura tradicional, la cual se refiere a aspectos mercadológicos (Aggrawal, Ahluwalia, Khurana & Arora, 2017), para efectos del presente trabajo nos referimos a un conjunto de variables considerando criterios referidos por Interbrand (2020); y teniendo como marco de análisis los elementos sugeridos por las normas ISO 10668 (2010) y la ISO 20671 (2019) en materia de valuación y evaluación de marcas, respectivamente.

De las once variables propuestas tres son de tipo cualitativo: Sector, país y popularidad; y las ocho restantes son cuantitativas. La temporalidad de los datos manejados es anual, y referidos al año 2020.

En lo que respecta a las variables cualitativas se codificaron de la siguiente manera. Ver Tabla 2 y Tabla 3:

Tabla 2
Variable sector y codificación

Sector	Codificación
Tecnología	1
Bebidas	2
Automotriz	3
Restaurante	4
Media	5
Servicios de negocios	6
Artículos deportivos	7
Lujo	8
Servicios financieros	9
Logística	10
Retail	11
Diversificado	12
Alcohol	13
Consumo inmediato	14
Vestir	15

Fuente: Elaboración propia con datos de Interbrand 2020

Tabla 3
Variable país y codificación

País	Codificación
Estados Unidos de América (EUA)	1
Corea del Sur	2
Japón	3
Alemania	4
Francia	5
Suecia	6
Italia	7
España	8
Suiza	9

Fuente: Elaboración propia con datos de Interbrand 2020

En cuanto a la variable popularidad, se codificó con respecto a su nivel de popularidad ya mencionado anteriormente, dando valores binarios, donde el 0 corresponde a baja popularidad y el 1 corresponde a alta popularidad.

Los pasos seguidos en el desarrollo de los modelos utilizando el software RStudio fueron de la siguiente manera.:

a. Análisis exploratorio de la base de datos.

b. Modelo KNN.

c. Modelo Árbol de clasificación

d. Modelo Naive Bayes.

Resultados

De acuerdo con la metodología empleada se llegaron a los siguientes resultados:

Análisis exploratorio

El análisis exploratorio consistió en ver la distribución de la base de datos, así como su limpieza y pre-procesamiento para el buen uso de los modelos propuestos. La distribución de los datos está dada a partir de las 50 marcas más valiosas del mundo según Interbrand (2020) y las 11 variables descritas en la Tabla 1, de las cuales como primer paso se vio que no existan valores nulos en dicha base.

Al tener variables cuantitativas y cualitativas, las variables sector, país y popularidad se transformaron con el código "as.factor" y nombrándolas de acuerdo al número que pertenece, con ello se procedió a identificar la existencia de valores nulos, que en dicha base y con la limpieza previa, se verificó que no existieron datos nulos. La distribución de la variable respuesta se visualiza de la siguiente manera. Ver (Figura 1).

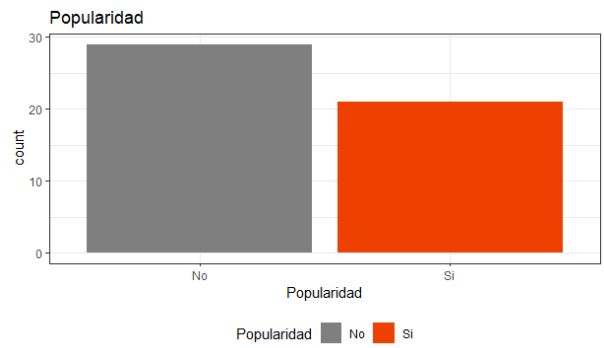


Figura 1. Distribución de la variable respuesta
Fuente: Elaboración propia, utilizando el programa RStudio

En la figura 1 se puede observar que de las 50 marcas dadas por los estándares para el cálculo de la popularidad, hay más marcas que no tienen tanta popularidad y el resto que si presentan mayor popularidad. Para ser más precisos se puede visualizar tal magnitud en la Tabla 4.

Tabla 4
Distribución de la popularidad

No popularidad	Si popularidad
29	21

Fuente: Elaboración propia

Esto indica que el 58% del total de nuestra muestra son marcas que no son consideradas populares y el 42% son consideradas populares.

En cuanto a la distribución de frecuencias de la variable país, 29 marcas pertenecen a los Estados Unidos de América, siendo la cantidad más alta, seguida por Alemania con 7 marcas y Francia con 5 marcas. Por otro lado, en la distribución de frecuencias por sector, las marcas que tienen mayor presencia son las del sector automotriz con 9 marcas, seguida por las marcas tecnológicas y servicios financieros con 6, y finalmente las de servicios de negocio y media con 5.

Para mostrar la interacción de la variable respuesta popularidad con las demás variables continuas se generaron las siguientes figuras, donde se presenta la distribución de las variables continuas de cada variable.

En principio las figuras 2 a 7 muestran dos formas de visualizar la relación que hay entre las variables propuestas y la variable objetivo que es la popularidad: la primera es en términos de densidad de probabilidad empírica suavizada (al lado izquierdo), donde se visualiza la distribución de datos cuantitativos en un intervalo o período de tiempo continuo; la segunda en el lado derecho, está en forma de un gráfico de caja o bigotes (boxplots o box and whiskers) y muestran la distribución de datos en cuartiles, resaltando el promedio y los valores atípicos. Ambas muestran la distribución de los datos analizados, y se observan tendencias y dispersiones. Más adelante se da una interpretación integrada de lo que representan estas figuras en su conjunto.

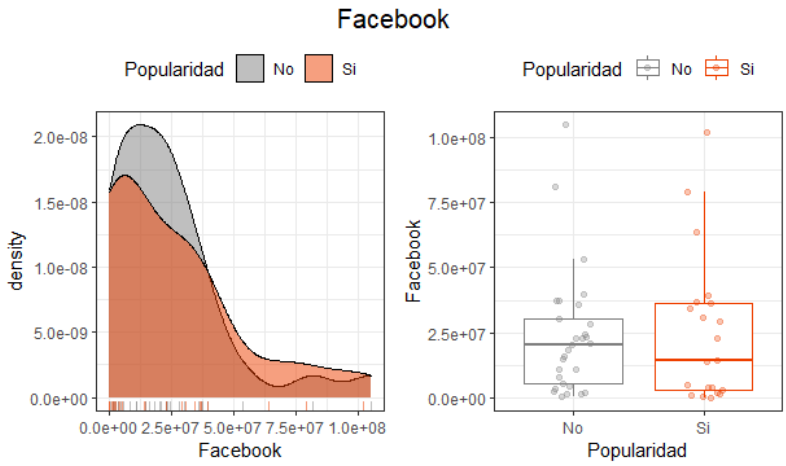


Figura 2. Distribución variables continuas (popularidad y Facebook)
Fuente: Elaboración propia, utilizando el programa RStudio

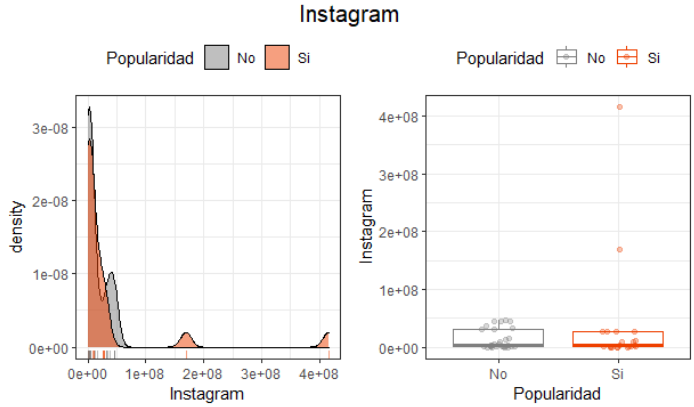


Figura 3. Distribución variables continuas (popularidad e Instagram)
Fuente: Elaboración propia, utilizando el programa RStudio

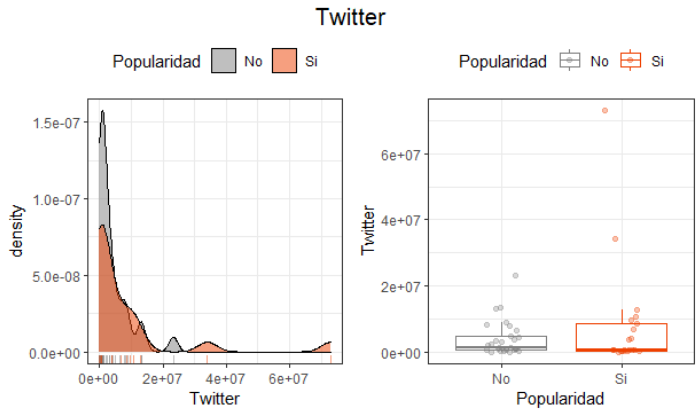


Figura 4. Distribución variables continuas (popularidad y Twitter)
Fuente: Elaboración propia, utilizando el programa RStudio

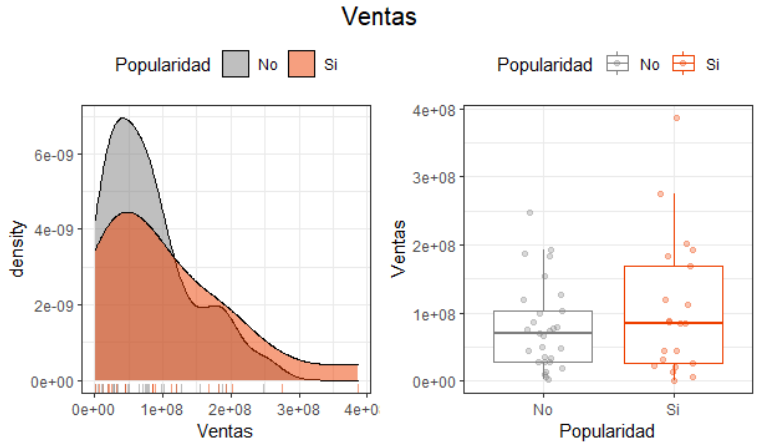


Figura 5. Distribución variables continuas (popularidad y Ventas)
Fuente: Elaboración propia, utilizando el programa RStudio

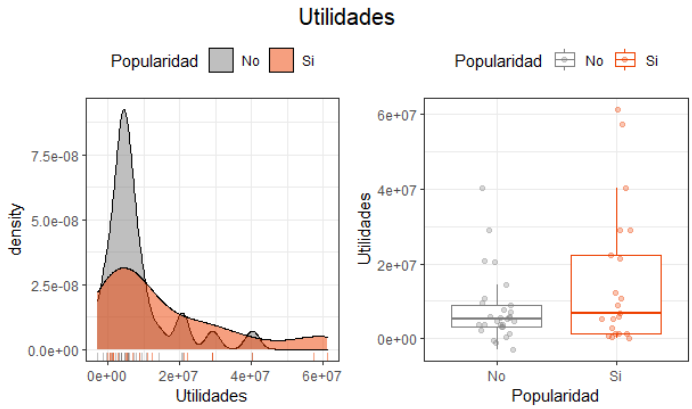


Figura 6. Distribución variables continuas (popularidad y Utilidades)
Fuente: Elaboración propia, utilizando el programa RStudio

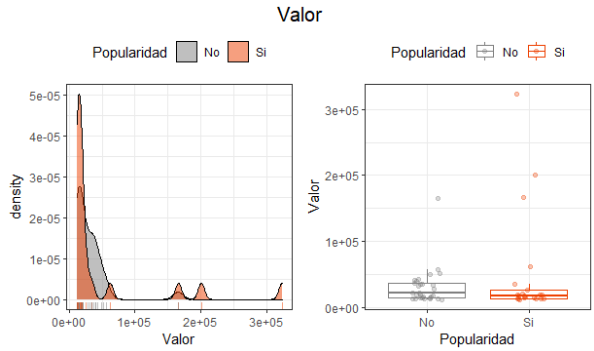


Figura 7. Distribución variables continuas (popularidad y Valor)
Fuente: Elaboración propia, utilizando el programa RStudio

Considerando las figuras 2, 3 y 4 podemos observar que el nivel de popularidad dado por las redes sociales de las plataformas de Facebook, Instagram y Twitter no presenta una distribución normal y la curva se encuentra más cargada hacia la izquierda. Por otro lado, de acuerdo con el grafico de caja, se observa la existencia de valores atípicos. Cabe mencionar que la plataforma Twitter, al observar el grafico de caja, concentra más valores dentro de la caja dado a que es considerada una de las redes sociales con mayor difusión entre los usuarios y las empresas.

Por otra parte, analizando las figuras 5, 6 y 7 se encuentra que la variable popularidad y las ventas muestran una distribución casi parecida a la normal, dado al gran volumen que ciertas empresas manejan de inventario y el gran impacto que tienen en la sociedad. En cuanto a la utilidad se muestra el mismo efecto, sin embargo, muestra una distribución normal cuando las empresas están generando

utilidades y tienen popularidad, considerando las diferentes estructuras de costos y gastos que presentan las empresas analizadas. Y en cuanto al valor de la marca con base a su popularidad, no se observa una distribución normal dado a la existencia de datos atípicos que sobrepasan a la media muestral del valor de la marca, esto se aprecia con más detalle en el gráfico de caja.

La distribución de las variables cualitativas se expresa de la siguiente manera. Ver (Figura 8 y Figura 9).

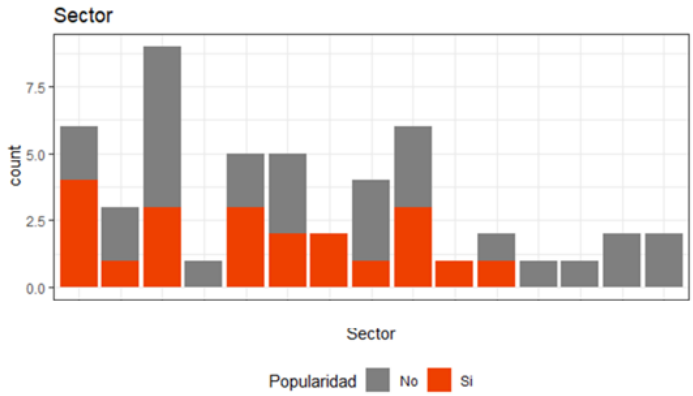


Figura 8. Distribución variables cualitativas (popularidad y Sector)
Fuente: Elaboración propia, utilizando el programa RStudio. El sector va en orden de acuerdo con el listado de la Tabla 2

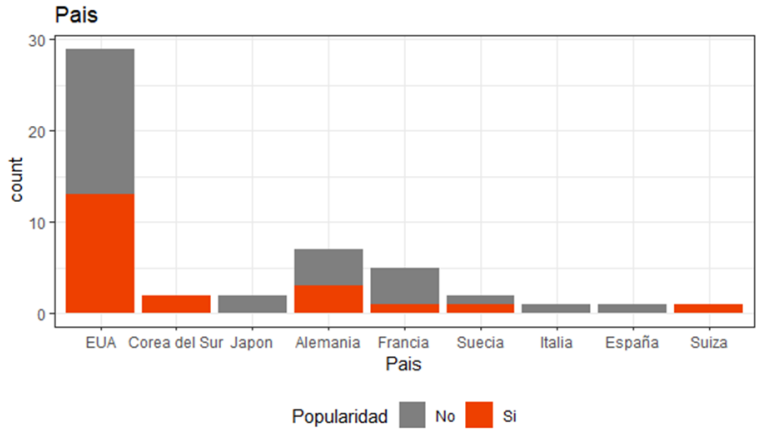


Figura 9. Distribución variables cualitativas (popularidad y País)
Fuente: Elaboración propia, utilizando el programa RStudio

De acuerdo con la figura 8 se presenta una mayor popularidad en el sector de tecnología, servicios de negocios y servicios financieros. Sin embargo, no presentan tanta popularidad el sector de bebidas, automotriz, restaurantes, diversificado y consumo inmediato.

En cuanto al país, la figura 9 muestra que se presenta una mayor popularidad de marcas en los países de Corea del Sur, Suiza, Estados Unidos de América y Alemania. Mientras que no se presentan niveles importantes de popularidad las marcas en países como Japón, Francia, Italia y España.

Importancia de las variables cuantitativas con respeto a nuestra base de datos, se encontraron los siguientes resultados. Ver (Figuras 10 a 13).

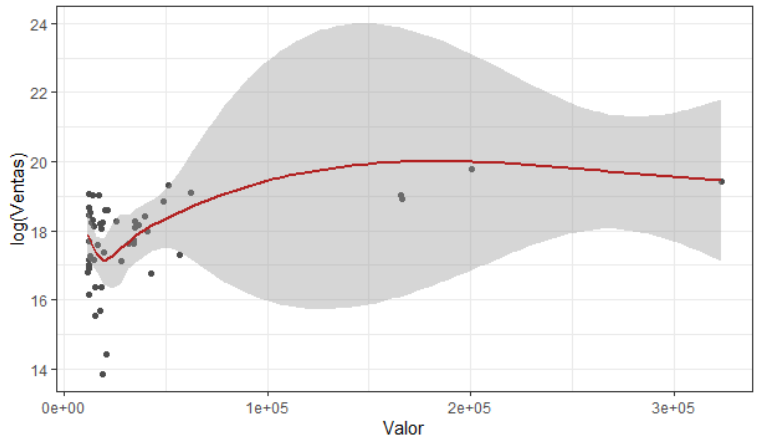


Figura 10. Importancia de las variables cuantitativas (Valor y Ventas)
Fuente: Elaboración propia, utilizando el programa RStudio

La variable valor con respecto a las ventas tienen una correlación de 0.6296, siendo significativa $t = 5.61, p < .05$ (4).

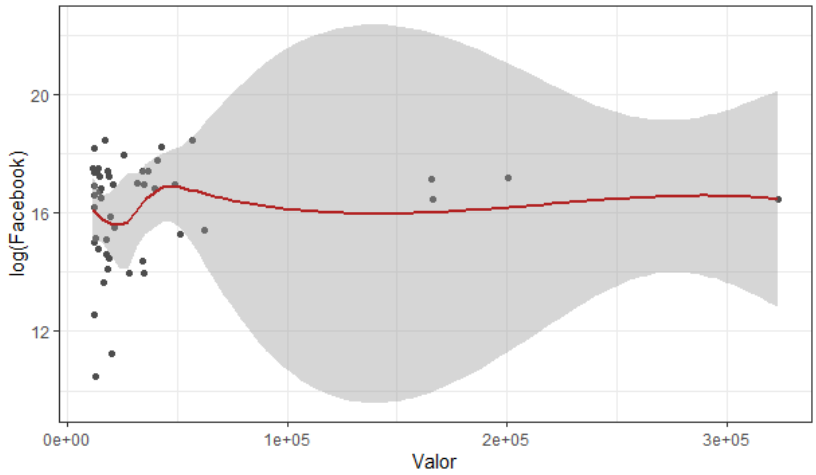


Figura 11. Importancia de las variables cuantitativas (Valor y Facebook)
Fuente: Elaboración propia, utilizando el programa RStudio

En cuanto al valor con respecto a la variable de Facebook no presentan resultados significativos y también se tiene una correlación positiva baja de 0.0058.

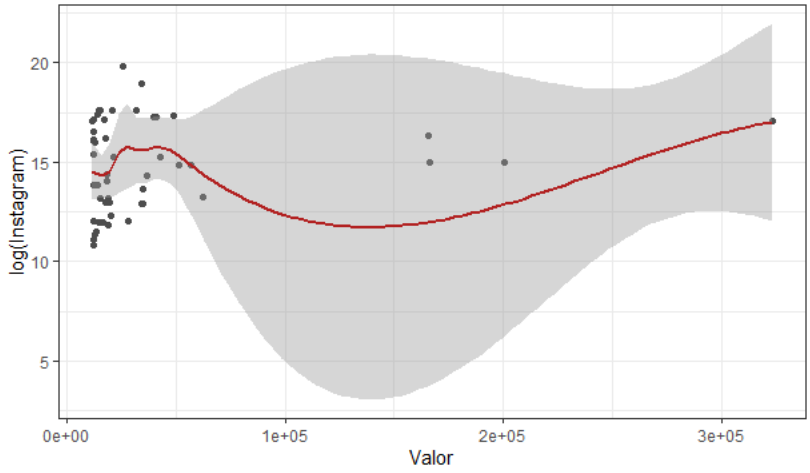


Figura 12. Importancia de las variables cuantitativas (Valor e Instagram)
Fuente: Elaboración propia, utilizando el programa RStudio

Para Instagram no se presentan resultados significativos y se obtiene una correlación negativa de -0.0202.

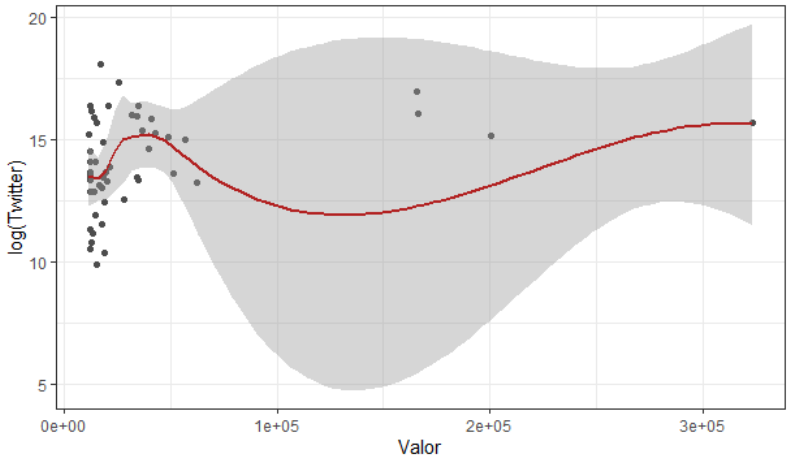


Figura 13. Importancia de las variables cuantitativas (Valor y Twitter)
Fuente: Elaboración propia, utilizando el programa RStudio

Y finalmente para la variable de Twitter, tampoco presenta una significancia estadística, pero se tiene una correlación de 0.0926.

Como conclusión preliminar de los resultados generados hasta este punto del análisis exploratorio, se percibe que existe relación entre variables cuantitativas y cualitativas. Cabe mencionar además que, de acuerdo con los gráficos generados, la correlación que existe entre las variables relacionadas con aspectos monetarios y de preferencia que hay en las redes sociales se presentan resultados atípicos, es decir que desde la inferencia estadística se han generado valores que no tienen relevancia entre sí. Sin embargo, al utilizar algoritmos de aprendizaje máquina se permite procesar datos que no serían significativos para la estadística tradicional (Lara, Mora & Londoño, 2022).

Segmentación de datos de entrenamiento y prueba

De acuerdo con el preprocesado de las variables se hicieron las siguientes transformaciones: para las variables continuas se hizo un proceso de normalización, que consiste en trabajar sobre la misma base numérica. En cuanto a la variable cualitativa, como es la popularidad, se realizó el proceso de binarización (0,1) con el objeto de que no existan outliers (valores atípicos) en la base de datos.

Modelo K-nn

La secuencia seguida para aplicar este modelo fue la siguiente:

- Aplicación del modelo en la base de datos no pre-procesada, es decir sin que las variables hayan sido normalizadas.;
- Aplicación del modelo en la base de datos no pre-procesados, pero tomando como arranque los resultados de la aplicación del modelo anterior;
- Finalmente, con base en los resultados de la aplicación de los dos modelos antes mencionados, se utilizan los datos normalizados para obtener el modelo optimo que servirá como el mejor modelo predictor con base en su "accurancy".

En la figura 14, se ilustra el funcionamiento de este método de clasificación, donde se encuentran representadas 50 muestras pertenecientes a dos clases distintas: la Clase 1 formada por los que si tienen popularidad y la Clase 2 formada por los que no tienen popularidad. En este caso, el modelo arrojó un resultado de veintitrés vecinos, es decir, k=23 de las 50 muestras, que es suficiente para hacer el análisis predictivo.

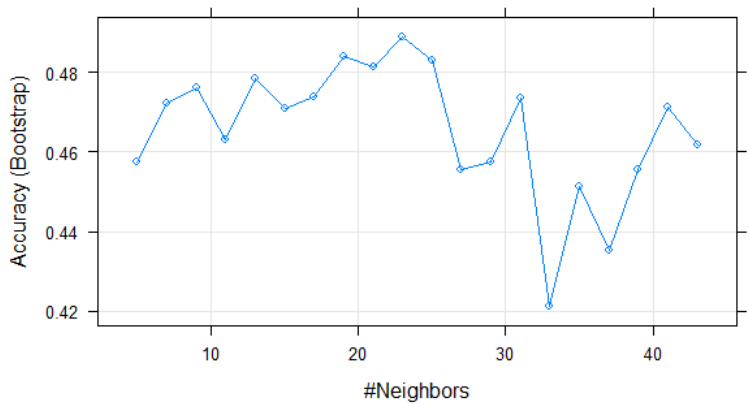


Figura 14. Modelo K-nn
Fuente: Elaboración propia, utilizando el programa RStudio

Con el propósito de observar el comportamiento de los datos, con relación a la variable objetivo, primero se aplicó el modelo en la base de datos sin las transformaciones de variables, es decir con los datos no pre-procesados. En la Tabla 4 se muestra el resultado obtenido en la aplicación del modelo sobre los datos no pre-procesados.

Tabla 4
Resultados K-nn (datos no pre-procesados)

No pre-processing
Resampling: Booststrapped (25 reps)
Summary of sample sizes: 26, 26, 26, 26, 26 ,26, 26,
Resampling results across tuning parameters:

k	Accuracy	Kappa
5	0.4576929	-0.027876321
7	0.4722657	0.030622951
9	0.4761565	0.016015804
11	0.4631363	-0.029195171
13	0.4784870	0.003882498

Fuente: Elaboración propia, utilizando el programa RStudio

Partiendo del resultado anterior de datos no pre-procesados, que arrojó un valor de k= 23, se ejecuta el siguiente modelo con datos no pre-procesados, dando ahora como resultado el vecino más cercano con k = 43. Esto significa que al aumentar el nivel de k se capta mayor cantidad de variables que corresponden al atributo objetivo que es la popularidad. Ver (Figura 15 y Tabla 5).

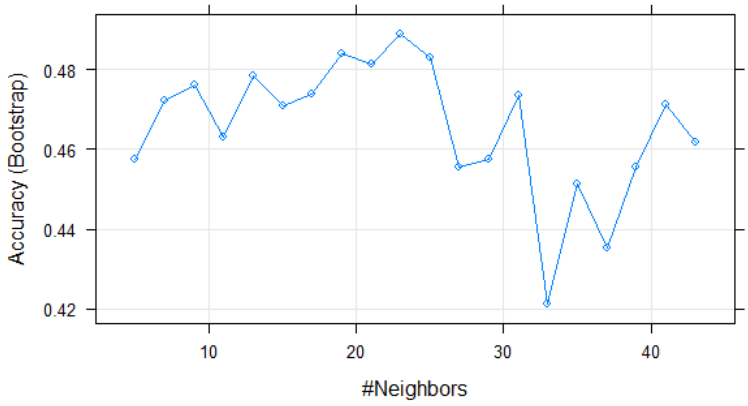


Figura 15. Modelo K-nn k=43
Fuente: Elaboración propia, utilizando el programa RStudio

Tabla 5
Resultados K-nn k=43

No pre-processing
Resampling: Cross-Validated (23 fold)
Summary of sample sizes: 25, 24, 25, 25, 24 ,25,
Resampling results across tuning parameters:

k	Accuracy	Kappa
5	0.5000000	0.0000000
7	0.5000000	0.0000000
9	0.5833333	-0.09090909
11	0.4444444	-0.15384615
13	0.5555556	0.0000000

Fuente: Elaboración propia, utilizando el programa RStudio

Cabe mencionar que la precisión (accuracy) es mejor con un mayor número de k que a con un valor menor. Como se mencionó previamente se utilizaron datos no pre-procesados.

Derivado de los resultados anteriores, se considera que arroja mejores resultados al considerar un valor de k=43, no obstante haber utilizado datos no pre-procesados. A continuación, se aplicó el modelo ahora con los datos ya normalizados o procesados, es decir considerando los datos que se relacionan o no con la variable objetivo que es la popularidad. Por lo tanto, se dividen en dos clases: tienen o no popularidad Ver (Figura 16).

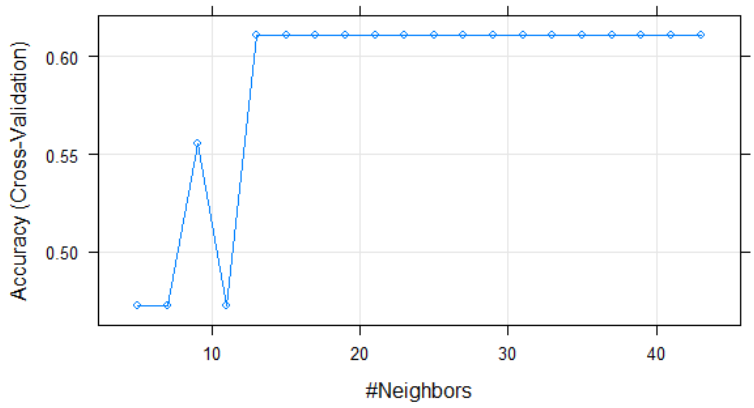


Figura 16. Modelo K-nn (datos procesados)
Fuente: Elaboración propia, utilizando el programa RStudio

El método K-nn supone que los considerados vecinos más cercanos generan una mejor clasificación utilizando los atributos, siendo en este caso el atributo objetivo la "popularidad" como la única variable utilizada, se procedió a realizar el análisis de predicción, como se muestra en la Tabla 6.

Tabla 6
Resultados predicción K-nn

	No	Si
1	0.5769230	0.4230760
2	0.5769231	0.4230761
3	0.5769232	0.4230762
4	0.5769233	0.4230763
5	0.5769234	0.4230764
6	0.5769235	0.4230765
7	0.5769236	0.4230766
8	0.5769237	0.4230767
9	0.5769238	0.4230768
10	0.5769239	0.4230769

Fuente: Elaboración propia, utilizando el programa RStudio

De acuerdo con estas diez muestras, es más probable que con base en los datos obtenidos un 57.69% que no se tenga tanta popularidad dado a las redes sociales y las ventas, y un 42.31% de probabilidad de que se tenga probabilidad dado a las redes sociales y las ventas. Ver (Tabla 7).

Tabla 7
Resultados Accuracy

Confusion Matrix and Statistics		
Reference		
Prediction	No	Si
No	6	4
Si	0	0
		Accuracy: 0.6
		95% CI: (0.2624, 0.8784)
		No Information Rate: 0.6
		P-Value [Acc > NIR]: 0.6331

Fuente: Elaboración propia, utilizando el programa RStudio

De acuerdo con la matriz de confusión, Accuracy de: $\frac{6+0}{(6+0+0+4)} = 0.6$ (5)

Para lo cual será comparado con los siguientes modelos para ver cual resulta óptimo.

Árbol de clasificación

Se procedió a generar un árbol de clasificación empleando como variable respuesta popularidad y como predictores todas las variables disponibles tales como las ventas. Ver (Tabla 8 e Figura 17).

Tabla 8
Árbol de clasificación

Classification tree:
Tree(formula = Popularidad ~., data = SP_entrena, minsize = 10
Number of terminal nodes: 4
Residual mean deviance: 1.021 = 22.46 / 22
Misclassification error rate: 0.2308 = 6 / 26

Fuente: Elaboración propia, utilizando el programa RStudio

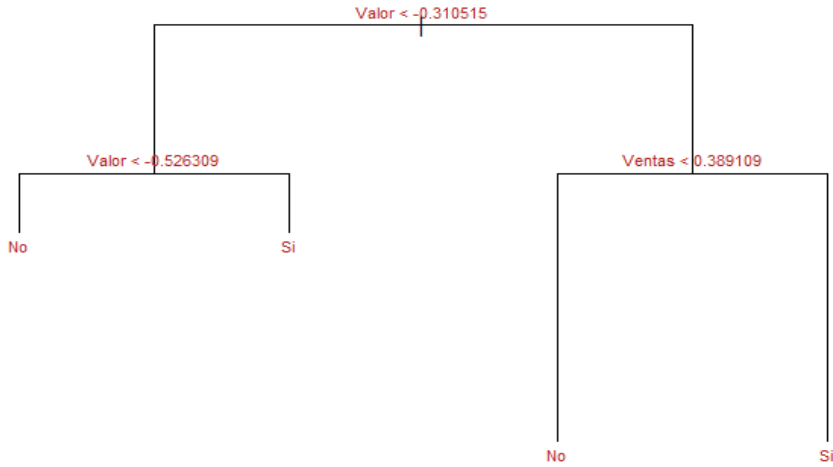


Figura 17. Modelo de árbol de clasificación
Fuente: Elaboración propia, utilizando el programa RStudio

Posteriormente, el proceso de "pruning[2]", es aplicado con el propósito de comparar los resultados frente al modelo inicial mostrado en la Tabla 9. La función summary () muestra que el árbol ajustado tiene un total de 4 nodos terminales y un classification error rate de entrenamiento del 23.08%. El término Residual mean deviance[3] mostrado en el summary, es simplemente la desviación residual

[2] El proceso de división binaria recursiva puede conseguir buenas predicciones con los datos de entrenamiento, ya que reduce el RSS (Residual Sum of Squares) de entrenamiento, lo que implica un sobreajuste a los datos (derivado de la facilidad de ramificación y posible complejidad del árbol resultante), reduciendo la capacidad predictiva para nuevos datos. Recuperado de https://rpubs.com/

[3] El residual mean deviance es aquella medida de error que queda en el árbol de clasificación, después de la construcción. Corresponde con variable entrenamiento RSS (Residual Sum of Squares), dividido entre el número de

dividida entre (número de observaciones – número de nodos terminales) dando como resultado 1.021. Cuanto menor es la deviance mejor es el ajuste del árbol a las observaciones de entrenamiento, en este caso la deviance es mayor y por lo tanto el ajuste del árbol no será tan bueno. Ver (Tabla 9).

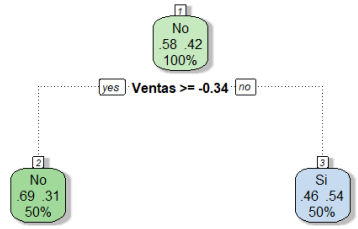
Tabla 9
Predicciones

predicciones	No	Si
No	0	0
Si	8	6

Fuente: Elaboración propia, utilizando el programa Rstudio

De acuerdo con la matriz de confusión, Accuracy de: $\frac{0+6}{0+0+8+6} = 0.4286$ (6)

Generando el proceso de "pruning", se llega a un óptimo de nodos de 2, por lo tanto, con base en las ventas se llega a tener sí o no popularidad, tal y como se muestra en la siguiente imagen. Ver (Figura 18).




```
Rule number: 3 [Popularidad=Si cover=13 (50%) prob=0.54]
  Ventas< -0.3407

Rule number: 2 [Popularidad=No cover=13 (50%) prob=0.31]
  Ventas>=-0.3407
```

Figura 18. Árbol de clasificación con 2 nodos
Fuente: Elaboración propia, utilizando el programa Rstudio

Naive-Bayes

Al ser un modelo de clasificación en Minería de Datos, donde se clasifica si de la muestra tiene popularidad o no, para este caso también llamados instancias, se caracterizará por una serie de atributos. Sin embargo, al contar en el modelo como un único atributo de tener o no popularidad respecto a tener el

observaciones menos el número de nodos. Cuanto menor es este valor, mejor se ajusta el modelo a los datos de entrenamiento. Recuperado de https://rpubs.com/

valor de marca, se llega a dicho resultado de ser el estatus con un accuracy más alto, dado que únicamente se cuenta con un atributo ya mencionado y no se consideran las otras variables cualitativas, siendo ventas la única variable causal. Es por ello que su grado de importancia es de 100, al igual que tiene popularidad, dado a que es el único atributo que se consideró para el análisis. Ver (Tabla 10).

Tabla 10
Resultados Naive-Bayes

26 samples
3 predictor
2 classes: `No´, `Si´
No pre-processing
Resampling: Cross-validated (10 fold)
Summary of sample sizes: 24, 24, 23, 23, 24, 24, …
Resampling results across tuning parameters:

usekernel	Accuracy	Kappa
TRUE	1.0000000	1.0
FALSE	0.9333333	0.8

Fuente: Elaboración propia, utilizando el programa RStudio

Con el objeto de identificar y evaluar los modelos que tengan mayor proporción de verdaderos positivos en sus resultados, se utiliza la llamada curva ROC (Receiver Operating Characteristic). Consiste en una representación gráfica del rendimiento del clasificador que muestra la distribución de las fracciones de verdaderos positivos y de falsos positivos. La bondad de una prueba diagnóstica que produce resultados continuos es lo que se le conoce como el área bajo la curva (AUC), que representa la probabilidad de que, en el caso de la investigación, la popularidad de las marcas tiene un efecto positivo o negativo en su valor económico. La figura 19, muestra el área debajo de la curva resultante a partir de las variables seleccionadas. El valor del área bajo la curva (AUC) es de 0.62, esto indica el nivel mínimo aceptado para los modelos de predicción anteriormente descritos. Con base en este criterio, el modelo que más se acerca a este criterio es el modelo de K-nn, debido a que tiene un uso más amplio de la base de datos, lo que permite apreciar mejor el comportamiento de las variables propuestas.

No obstante, a lo anteriormente dicho, se puede concluir de manera preliminar que para la aplicación de este modelo K-nn al igual que de los otros dos modelos, se requiere el uso de otras variables adicionales a las consideradas en este estudio, para que cumplan con los parámetros mínimos aceptables.

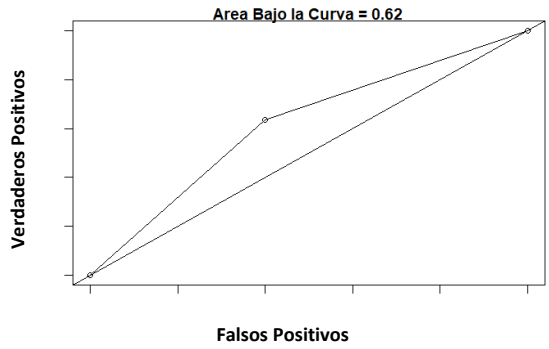


Figura 19. Curva ROC (Receiver Operating Characteristic)
Fuente: Elaboración propia, utilizando el programa RStudio

Conclusiones

En cuanto al objetivo establecido respecto a que, si las variables propuestas en el estudio influyen en la popularidad de la marca, se concluye que todas influyen de manera positiva en diferente medida. En lo referente a la popularidad de las marcas y las redes sociales, se percibe que el efecto es mayor en Twitter, debido a que tiene mayor presencia entre los líderes de opinión y representa un medio de comunicación formal y oficial entre las empresas y consumidores. En otros aspectos, variables como el valor de la marca y las ventas tienen un efecto en la popularidad importante considerando que en los medios se difunde cada vez con mayor importancia la existencia de empresas y marcas líderes globales y sus tendencias de desarrollo. Por otro lado, tienen menor efecto las variables que se relacionan con aspectos más técnicos y especializados de la gestión de las empresas en términos financieros como son las utilidades.

En cuanto a la influencia que hay entre el valor de la marca y las variables de las redes sociales dadas como "likes", se identificó que los elementos que se relacionan con el valor de la marca y son indicadores de algún grado de popularidad, contribuyen a la construcción de marcas poderosas.

La estrecha relación que existe entre el valor de la marca y de las ventas es relevante, ya que existe una amplia correlación entre ellas. Sin embargo, es recomendable investigar qué otros factores son considerados en el impacto adicional en lo que hace que una marca sea valiosa y popular (Acuña Moraga & Severino-González, 2018). Tal y como lo indica González, Orozco & Barrios (2011) que la relación se basa en un conjunto de dimensiones tales como el conocimiento, relación y actitud hacia la marca, así como preferencia de marca desde diferentes niveles de involucramiento en el proceso de compra. Para entender la relación existente entre la popularidad de marca y la evaluación de los atributos por parte del

consumidor, la preferencia de marca y la lealtad a la misma se observa que dicha popularidad no está asociada únicamente con las ventas dadas por ella (Boix, Boluda, & López, 2019).

Debido a tales descubrimientos, para estudios posteriores y que den mayor explicación, se propone el uso de otro tipo de variables como lo indica García Granda & Gastulo Chuzónen (2018), en términos de la recolección de datos que indiquen el grado de lealtad, reputación y gusto de una marca. Así mismo, de acuerdo con Tapia Cedeño (2017), utilizar una variable de satisfacción, genera un punto favorable como un factor de popularidad.

También se puede afirmar que, considerando un conjunto de marcas valiosas en el mercado global, existen efectos en el valor de dichas marcas respecto a no solo a sus ventas sino a la presencia de factores internos y externos relevantes en la gestión de las marcas como activos de las empresas. Estos efectos varían entre los tipos de sectores de mercado en los cuales se desempeñan las marcas, lo cual conduce a considerar la existencia de un grado de popularidad dada por las empresas con base en la gestión de sus marcas.

Para el caso de México es importante analizar estos efectos que existen, tales como la popularidad en las marcas y su valor en los mercados globales, esto permitirá identificar parámetros de comparación entre las prácticas de empresas líderes mundiales y las que se desempeñan en mercados nacionales y con tendencias a su internacionalización. Las características de la gestión de las marcas y su valor tienen relación importante bajo la perspectiva del comportamiento y la conducta no solo individual, sino colectiva y a niveles corporativos de las empresas.

Referencias

Acuña Moraga, O., & Severino-González, P. E. (2018). Sustentabilidad y comportamiento del consumidor socialmente responsable. Opción: Revista de Ciencias Humanas y Sociales, (87), 299-324 Disponible en: http://repositorio.ucm.cl/handle/ucm/2450

Aggrawal, N., Ahluwalia, A., Khurana, P. & Arora, A (2017). Brand analysis framework for online marketing: ranking web pages and analyzing popularity of brands on social media. Soc. Netw. Anal. Min. 7, 21. https://doi.org/10.1007/s13278-017-0442-5

Beltrán, C., & Barbona, I. (2021). Comparación del desempeño de Árboles de clasificación y Redes Neuronales en la clasificación politómica mediante simulación. Revista de epistemología y ciencias humanas, Disponible en: http://hdl.handle.net/2133/21727

Boix, J. C., Boluda, I. K., & López, N. V. (2019). ¿Por qué las instituciones de educación superior deben apostar por la marca? Revista de investigación educativa, 37(1), 111-127. https://doi.org/10.6018/rie.37.1.291191

Brunton, S. L., & Kutz, J. N. (2022). Data-driven science and engineering: Machine learning, dynamical systems, and control. Cambridge University Press. https://doi.org/10.1017/9781108380690

Camaño, G y Goyeneche, J. (2011.). Selección de una muestra ordenada con semillas y algoritmos de números aleatorios. (Serie DT (11/00)). Udelar. FCEA-IESTA. https://hdl.handle.net/20.500.12008/10558

Casajús Setién, J. (2022). Autocodificador evolutivo de red Bayesiana para detección de anomalías aplicado a ciberseguridad. Tesis (Master), E.T.S. de Ingenieros Informáticos (UPM). https://oa.upm.es/71723/

Carta, S., Podda, A. S., Recupero, D. R., Saia, R., & Usai, G. (2020). Popularity Prediction of Instagram Posts. Information (2078-2489), 11(9). https://doi.org/10.3390/info11090453

Crespo, A. B. (2013). Aprendizaje máquina multitarea mediante edición de datos y algoritmos de aprendizaje extremo (Doctoral dissertation, Universidad Politécnica de Cartagena).

Cuellar, J. (2019). Popularidad de los contenidos de instagram en marcas de lujo. Repositorio Academico de la Universidad de Chile. Disponible en: http://repositorio.uchile.cl/handle/2250/179714

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, (4), 325-327. https://doi.org/10.1109/TSMC.1976.5408784

Espino Timón, C. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo-herramientas Open-Source que permiten su uso (Grado en Ingeniería Informática Business Inteligence, Universidad Oberta de Catalunya).

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247. https://doi.org/10.2307/1403796

Fleuren, L.M., Klausch, T.L.T., Zwager, C.L., Schoonmade, L. J., Guo, T., Roggeveen, L. F., Swart, E. L., Girbes, A. R. J., Thoral, P., Ercole, A., Hoogendoorn, M., & Elbers, P. W. G. (2020). Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 46, 383–400. https://doi.org/10.1007/s00134-019-05872-y

García Granda, T. L., & Gastulo Chuzón, D. N. (2018). Factores que influyen en la decisión de compra del consumidor para la marca Metro-Chiclayo. Tesis de pregrado, Universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú. Disponible en: http://hdl.handle.net/20.500.12423/1039

González, E., Orozco, M., & Barrios, A. (2011). El valor de la marca desde la perspectiva del consumidor. Revista Contaduría y Administración, 235, 217-239. https://www.redalyc.org/pdf/395/39519916011.pdf

Horna, K. S. A., & Prado, A. L. (2015). Valor de marca: un acercamiento conceptual mediante su origen y modelos. Revista de Investigación Valor Agregado, 2(1). https://doi.org/10.17162/riva.v2i1.837

Interbrand. (2020). Best Global Brands 2020: Methodology. Disponible en: https://interbrand.com/thinking/best-global-brands-2020-methodology/

ISO 10668:2010, Brand valuation — Requirements for monetary brand valuation

ISO 20671:2019, Brand evaluation — Principles and fundamentals

Kim, M. Y., Moon, S., & Iacobucci, D. (2019). The Influence of Global Brand Distribution on Brand Popularity on Social Media. Journal of International Marketing, 27(4), 22-38. https://doi.org/10.1177/1069031X19863307

Lara, P. H. V., Mora, F. A. G., & Londoño, C. M. G. (2022). Aprendizaje de máquina para mantenimiento predictivo: un problema de clasificación binaria. ConcienciaDigital, 5(2.1), 45-68. https://doi.org/10.33262/concienciadigital.v5i2.1.2150

Fernández Lizana, M. I. (2020). Ventajas de R como herramienta para el Análisis y Visualización de datos en Ciencias Sociales. Revista Científica De La UCSA, 7(2), 97–111. Recuperado a partir de https://revista.ucsa-ct.edu.py/ojs/index.php/ucsa/article/view/30

Merino, R. F. M., & Chacón, C. I. Ñ. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases, (10), 165-189. https://doi.org/10.26439/interfases2017.n10.1775

Mergel, B. (1998). Diseño instruccional y teoría del aprendizaje. Universidad de Saskatchewan, Canadá. Disponible en: www. usask. ca/education/coursework/802papers/mergel/espanol. pdf. [Consultado el 8 de mayo de 2006], 16.

Ni, Z. (2022). Sistema de extracción de datos (Doctoral dissertation, ETSI_Informatica). https://oa.upm.es/71408/

Nieto Jeux, A. (2021). Algoritmos de aprendizaje automático: un estudio de su difusión y utilización (Trabajo Fin de Grado, E.T.S. de Ingenieros Informáticos (UPM), Madrid, España). https://oa.upm.es/68484/

Pérez Curiel, C. y Sanz-Marcos, P. (2019). Estrategia de marca, influencers y nuevos públicos en la comunicación de moda y lujo. Tendencia Gucci en Instagram. Prisma Social: revista de investigación social, 24, 1-24. Disponible en: https://orcid.org/0000-0002-1888-0451 http://orcid.org/0000-0002-6103-6993

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808. https://doi.org/10.48550/arXiv.1811.12808

Rich, E., Knight, K., Calero, P. A. G., & Bodega, F. T. (1994). Inteligencia artificial (Vol. 1). McGraw-Hill. ISBN 8448118588, 9788448118587

Robson, S., Banerjee, S., & Kaur, A. (2022). Brand Post Popularity on Social Media: A Systematic Literature Review. In 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-6). IEEE. https://doi.org/10.1109/IMCOM53663.2022.9721784

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. Revista Ibérica de Sistemas e Tecnologias de Informação, (E28), 586-599. Disponible en https://www.proquest.com/docview/2388304894?pq-origsite=gscholar&fromopenview=true

Román, M.V. & Lévy, J.P. (2003). Clasificación y segmentación jerárquica. En J.-P. Lévy y J. Valera (Dietrs), Análisis Multivariable para las Ciencias Sociales (pp. 567-630). Madrid: Pearson Prentice Hall

Rrmoku, K., Selimi, B., & Ahmedi, L. (2022). Application of Trust in Recommender Systems—Utilizing Naive Bayes Classifier. Computation 10, 6. https://doi.org/10.3390/computation10010006

Sasikala, B. S., Biju, V. G., & Prashanth, C. M. (2017). Kappa and accuracy evaluations of machine learning classifiers. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 20-23). IEEE. https://doi.org/ 10.1109/RTEICT.2017.8256551

Sneider Castillo, J., & Ortegón Cortazar, L. (2016). Componentes del valor de marca en marketing industrial. Caso máquinas y herramientas. Revista Perspectivas, (37), 75-94. Disponible en: On-line ISSN 1994-3733

Sriram, K. V., Prabhu, H. M., & Bhat, A. A. (2019, November). Mobile Phone Usability and its Influence on Brand Loyalty and Re-Purchase Intention: An Empirical. In 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 1-4). IEEE. https://doi.org/10.1109/wiecon-ece48653.2019.9019911

Tapia Cedeño, G. A. (2017). Análisis de los factores que influyen al comportamiento del consumidor en los bares-restaurantes en la ciudad de Portoviejo. Trabajo de titulación. Carrera de Marketing. Portoviejo, USGP. Disponible en: http://repositorio.sangregorio.edu.ec/handle/123456789/365

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. PloS one, 14(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Viera, Á. F. G. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. Investigación bibliotecológica, 31(71), 103-126. https://doi.org/10.22201/iibi.0187358xp.2017.71.57812.

Wang, L. (2019). Research and implementation of machine learning classifier based on KNN. In IOP Conference Series: Materials Science and Engineering (Vol. 677, No. 5, p. 052038). IOP Publishing. https://doi.org/10.1088/1757-899X/677/5/052038

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15, 713-714. S. https://doi.org/10.1007/978-0-387-30164-8_576

Yang, Y., & Webb, G. I. (2002). A comparative study of discretization methods for naive-Bayes classifiers. In T. Yamaguchi, A. Hoffmann, H. Motoda, & P. Compton (Eds.), Proceedings of The 2002 Pacific Rim Knowledge Acquisition Workshop (pp. 159 - 173). Japanese Society for Artificial Intelligence. Disponible en: (https://users.monash.edu/~webb/Files/YangWebb02a.pdf)

Zaki, M. J., & Meira Jr, W. (2020). Data mining and machine learning: Fundamental concepts and algorithms. Cambridge University Press. https://doi.org/10.1017/9781108564175