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

Contaduría y Administración
Accounting & Management

Contaduría y Administración 68 (1), 2023, 1-33

Brand popularity and its economic value in corporate finance framework; A machine learning analysis

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

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

Universidad Panamericana, México

Received May 17, 2022; accepted November 10, 2022
Available online September 9, 2024

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 its diagnosis and machine learning techniques developing 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

*Corresponding author.
E-mail address: 0120549@up.edu.mx (V.M. Morales González).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

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

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.

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

Introduction

In today's world, social networks can be considered an important source of indicators of opinion, consumption behavior, and media prestige of some brands. This situation reflects a way of perceiving brands' image and value. The communication and marketing model focuses on developing the brand to achieve maximum visibility and profit in sales portals to reach higher volumes and generate significant economic benefits in the short term (Pérez Curiel & Sanz-Marcos, 2019).

Before social networks, clients perceived a brand's appreciation through word of mouth, and the number of sales evidenced popularity. These concepts have changed (Cuellar, 2019). Currently, in social networks, consumers issue judgments about products and services through "likes" or spread comments, while companies take these elements in real-time (Cuellar, 2019).

Throughout time, brands have taken a leading role in the business environment; they constitute an important element in commercial procedures and in the generation of value. Therefore, it is no longer appropriate to refer to the brand only as a symbol or image of the product or service in the market but as the value that involves distinctive elements that somehow reflect brand recognition, client loyalty, perceived quality, and popularity (Horna & Prado, 2015). Hence, knowing the true meaning of brand equity is important as an indicator that measures the consumer's perception of the competition. Furthermore, it enables the direction of each strategy and decision to satisfy the client's needs and thus to see the level of popularity that a brand generates for such effects (Horna & Prado, 2015).

The present study is focused on exploring the components of the concept of brand equity using a diagnosis and machine learning techniques (Sneider Castillo & Ortegón Cortazar, 2016) to develop a

series of models associated with the dimensions of perceived brand equity from some factors identified in social networks (Facebook, Instagram, and Twitter), based on what could be considered as popularity.

Commenting on the popularity of brands in the medium of social networks has been considered in most of the literature as a form of brand relation with the market, and its relevance has to do with marketing aspects. For Aggrawal, Ahluwalia, Khurana, and Arora (2017), online marketing is one of the best measures to establish a brand and increase its popularity. Advertisements are among the best ways to showcase the company's products/services, resulting in a valuable strategic marketing line. Running ads on so-called utility web pages helps to maximize brand reach and get better feedback. These authors have proposed a framework that enables analyzing brand popularity regarding its presence on websites and social networks. They use algorithms for text analysis, sentiment analysis in messages, and network construction. Brand popularity is demonstrated regarding the frequency of appearance on web pages and sentiment analysis in social network texts.

Similarly, Kim, Moon, and Iacobucci (2019) report a study conducted to propose a marketing management tool for global brands based on their popularity by country and consumer activities in social networks. In their analysis, they use the study variables sales, profits, and brand value, in addition to the measurement of popularity. In this case, popularity is understood as a form of consumer receptivity and preference for brands through social networks, an aspect that contrasts with what in the past has been associated with the degree to which the general population searches for and buys a product or service because of its brand.

Empirical studies have also been conducted regarding brand loyalty in terms of product repurchase. For example, in the work described by Sriram, Prabhub, and Bhat (2019), brand loyalty is analyzed with the desire to repurchase cell phones, having as a framework of analysis the ISO 9241(1992/2001) regulation, focused on quality in usability and ergonomics of both hardware and software. In this work, traditional statistical techniques are applied to the results of surveys among cell phone users regarding the efficiency, effectiveness, and satisfaction of the products. This work focuses not only on marketing terms but also on the so-called usability of the products and, consequently, on brand loyalty. These aspects are important from a commercial perspective and, consequently, from a brand value perspective.

To analyze the popularity of brands published on social networks, Robson, Banerjee, and Kaur (2022) show a review of literature from recent years in which they identify up to 22 concepts related to popularity associated with brand publications, reaching as one of their conclusions that few studies have examined the interactions between these concepts. Although they consider the study of the interaction of these brand popularity concepts important for marketing purposes, they do not consider these interactions

with the value of brands as an asset of the companies. It is recognized that the marketing dimension only represents a part of the company's strategic management.

In a similar sense to that discussed by Kim, Moon, and Iacobucci (2019) and Robson, Banerjee, and Kaur (2022), this study proposes to start from 11 variables associated with a concept of popularity, not only focused on marketing aspects but also on those aspects that concern the management and economic value of brands. These variables have been proposed considering two important criteria: those considered to date by international consulting companies in the valuation of global brands, such as Interbrand (2020), and those that have as a framework of analysis the elements suggested by the regulations ISO 10668 (2010) and ISO 20671 (2019), regarding brand valuation and evaluation, respectively. In order to conduct this analysis, a machine learning procedure is proposed, which turns out to be relevant considering the scope of data analysis and conclusions reached by previous studies, such as those conducted and reported by Kim, Moon, and Iacobucci (2019) and Robson, Banerjee, and Kaur (2022), regarding contradictory results in analyses conducted using traditional statistical techniques. Therefore, using tools such as big data analytics, data science tools, and machine learning is pertinent.

According to Crespo (2013), machine learning is a set of techniques that tend to improve the behavior or performance of a system through acquired experience. It is a discipline in the field of Artificial Intelligence (Rich, Knight, Calero, & Bodega, 1994), which, through algorithms, endows computers with the ability to identify patterns in massive data and devise predictions (predictive analysis) (Espino Timón, 2017).

The machine learning methodology prioritizes prediction over inference. Unlike classical statistics, where a model is specified, it does not impose a specification or a theory (Viera, 2017). The statistical model seeks, with theoretical bases, to find the relation between variables to identify explanatory variables, dependence or independence between them, and the sense of their relation, as well as to test hypotheses and conduct inferences. Whereas the machine learning methodology enables the data to "speak" or express themselves, it prioritizes the importance of prediction over inference through an algorithm that finds the input-output relation and seeks for the model to replicate the data using the extra-sample cross-validation tool (Mergel, 1998). Based on the above, it will be shown that the variables proposed in the study influence brand popularity.

The present work is developed as follows: The following section describes the three machine learning algorithms used in this research, and the codes and instructions in R language employed to conduct the calculations of each algorithm; a brief justification of the use of R language is also presented. Subsequently, in section three, the 11 variables that comprised the database used are described, where three of them are qualitative and 8 quantitative. This section concludes with an analysis of the results obtained with the three algorithms. Subsequently, the conclusions are presented, which can be synthesized

by establishing that the results obtained employing the three machine learning techniques show that the eleven variables proposed in the study positively influence the brand's popularity, although with different intensities. Finally, the bibliographical references are listed.

Methodology

According to Carta, Podda, Recupero, Saia, and Usai (2020), there are several machine learning algorithms whose choice or appropriateness depends on the target strategy, the input/output data type involved, and the type of problem to be analyzed. Thus, several types of machine learning algorithms are proposed in the literature, such as supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, self-learning, feature learning, etcetera (Nieto Jeux, 2021), with supervised, unsupervised, and semi-supervised learning algorithms being the most widely used.

According to Rojas (2020), in supervised learning, functions are learned, and relations that associate inputs with outputs are adjusted to a set of examples of which the relation between the input and the desired output is known. Unsupervised learning models are those in which there is no interest in adjusting inputs and outputs but rather in increasing the structural knowledge of the available data. For Ni (2022), the third most important algorithm is the Semi-supervised, which combines some properties of the other two types of algorithms described above. It consists of handling data sets where additional attributes are incorporated in the target variable and other variables considered convenient.

Models used for machine learning

This work used the RStudio program to generate machine learning models like K-nn, Classification Trees, and Naïve Bayes. The essential reason for using the R language through the RStudio suite is that it is an environment and programming language used primarily for statistical data analysis and graph construction. Given its quality, versatility, and countless open-source libraries, where enough useful algorithms have been programmed for machine learning development, its use has already become a standard for this type of analysis and, in general, for statistical and econometric applications. R is widely used in biostatistics, data mining, econometrics, data visualization, etcetera (Fernández Lizana, 2020).

In accordance with the proposal of Sasikala, Biju, and Prashanth (2017), one of the objectives of this work is to decide which of the machine learning models used is more efficient to measure the popularity of brands based on their financial, economic, and technological variables. The three algorithms mentioned above will be used to generate predictive models.

Applying the algorithms above requires using data science models, including the K-nn (K nearest neighbors) method, classification trees, and the Naïve Bayes model. Since a binary response variable was used in this work, these models may be appropriate for data analysis (analytics), normally used in economic-business applications (Brunton & Kutz, 2022).

The K-nn (K nearest neighbors) method (Fix & Hodges, 1989) and (Dudani, 1976) is a supervised classification method (learning, estimation based on a training set and prototypes), which enables estimation of the probability density function. This method is a non-parametric classification, which estimates the value of the probability density function or directly the a posteriori probability that an element x belongs to class C_j[1] from the information provided by the set of prototypes or examples (Dudani, 1976). In the learning process, no assumption is made about the distribution of the predictor variables.

This method has been used in pattern recognition for the last 40 years. One of its advantages is that it has been applied to categorization in research strategies, where, in the first instance, the distance between the new sample and the training sample is calculated. Subsequently, and according to the category to which the neighbor belongs, the new sample is determined, and it is verified whether they all belong to the same category (Wang, 2019) and (Zaki & Meira, 2020). The essential equation of this method is as follows:

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

Where:

$d(x_i,x_j)$: is the Euclidean distance of the element x_i in relation to x_j

n: is the number of attributes

$ar(x_i)$: represents the i-th attribute of each item

According to Zaki and Meira (2020), the method consists of averaging the value of Y over the K observations closest to the point x_0 to estimate the associated response variable $\hat{Y}_0$

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

Where: $Nk(x_0)$ is the vicinity of x_0 defined by the k nearest points in the training data.

[1]It is the assignment given to a class label on the basis that it is most frequently represented around a given data point (Raschka, 2018) Source: https://sebastianraschka.com/pdf/lecture-notes/stat479fs18/02_knn_notes.pdf

In the case of classification with a binary categorical response variable (popularity), i.e., Y = 0, 1, the average of the k closest observations will represent the estimate of the probability that the point x takes the value Y = 1:

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

This local estimation method does not consider rigorous assumptions about the data and is based on the best empirical estimator for a quadratic loss function.

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

This classifier will always choose the category for maximization of classification probability.

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

The constructed algorithm used in the RStudio program is as follows:

A random data seed is generated so the subsequent results are not modified. According to Camaño and Goyeneche (2011) and Casajús (2022), a seed is the initial value introduced into the computer program so that the algorithm generates the series of random numbers. Since the random numbers that will be used to perform the initial ordering of the candidates are generated by a computer program, their randomness depends precisely on the fact that the number given for the start (the seed) is random.

According to the above, it should start from the following instruction in R:

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

In order to apply this method, the database is described in Section 3. Data selection and results were split for the training and testing phases to calibrate the model's functionality. For this purpose, the test is performed with a data set divided into two parts: training and test data. The training data are used to give training instructions to the model, while the test data generate the predictions based on the model and enable a comparison of the generated values with the real values of the sample. The data set is typically split into 70% training and 30% test data (Vabalas, Gowen, Poliakoff, & Casson, 2019).

Accordingly, in the first phase, 70% of the data is selected to train the algorithm and provide information to find the necessary patterns and make predictions. In the test phase, the rest of the data (30%) is used to evaluate whether the model's response is reliable as a predictive model. Subsequently, the following code was incorporated to elaborate the K-nn model. Its relevance lies in being able to describe the behavior of the response variable "Popularity."

```
SP_knnEntrenado <- train(Popularidad ~ .,
data = SP_entrena,
```

method = "knn",

tuneLength = 20)

class (SP_knnEntrenado)

SP_knnEntrenado

plot(SP_knnEntrenado)

In the classification tree model, the dependent variable is categorical, and the value at the terminal node is equal to the mode of the observations of the training set that have "fallen" in that region (Merino & Chacón, 2017). Decision or classification trees emerged in machine learning and Artificial Intelligence (Román & Lévy, 2003).

According to Beltrán and Barbona (2021), it is a non-parametric binary segmentation method constructed by repeatedly dividing the data. The data are classified into mutually exclusive groups. The algorithm starts with an initial node, divided into two sub-groups or sub-nodes; finally, a variable is chosen, and the cut-off point is determined so that the units belonging to each newly defined group are as homogeneous as possible. The essential equation of this method is the following (Zaki & Meira, 2020):

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

Where:

i(t): known as the Gini impurity regarding how a randomly chosen item is mislabeled.

p(j / t): is the probability of an error in the categorization of an element

According to Zaki and Meira (2020), since the response variable (popularity) is qualitative, there are several alternatives to finding homogeneous nodes. The most commonly used are:

Classification error rate:

It is defined as the proportion of observations that do not belong to the most common class in the node.

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

Where $\widehat{p_{mk}}$ represents the proportion of observations of node m that belong to class k. Despite its simplicity, this measure is not sensitive enough to create models.

Gini Coefficient

It measures the total variance in the set of K classes of node m and is considered a measure of the node's homogeneity.

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

When $\widehat{p_{mk}}$ is close to 0 or 1, the node contains mostly observations of one class—consequently, the higher the node's homogeneity, the lower the value of the Gini G coefficient.

Chi-square

This approach consists of identifying whether there is a significant difference between the particular nodes and the general node, i.e., whether there is evidence that the division achieves an improvement. For this purpose, the chi-square goodness of fit statistical test is applied using the frequency of each class in the general node as the expected H_0 distribution. The higher the X^2 statistic, the greater the statistical evidence of a difference.

$$X^2 = \sum k \frac{(observed\ k - expected\ k)^2}{expected\ k}$$

The value of the measurement at each of the two resulting nodes is calculated for each possible division. The two values are summed by weighting each one by the fraction of observations in each node.

$$\left(\frac{n\ node\ observations\ A}{n\ total\ observations}\right) * purity\ A + \left(\frac{n\ node\ observations\ B}{n\ total\ observations}\right) * purity\ B$$

The division with the lowest or highest value (depending on the measure used) is selected as the optimal division. Purity is understood as the maximum probability of each node, according to the entropy or Gini coefficient. Consequently, impurity is usually measured as each node's minimum probability of occurrence (Zaki & Meira, 2020).

The constructed algorithm used in the RStudio program is as follows:

Starting from the fact that a seed is generated because it is a random process and it is desired that the subsequent results are not modified in this process, using the following instructions:

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

summary(arbol_clasificacion)

Finally, the Naïve Bayes model is one of the most widely used classifiers due to its simplicity and speed. It consists of a supervised classification and prediction technique that allows models to be built that predict the probability of possible outcomes based on Bayes' Theorem, also known as the conditional probability theorem (Webb, Keogh, & Miikkulainen, 2010). One limitation is that the assumption of attribute independence sometimes does not correspond to reality. Nevertheless, it has been suggested that in the face of such limitations, its impact may be less because binary classification is considered a function of probability estimation (Yang & Webb, 2002; Rrmoku, Selimi, & Ahmedi, 2022). The essential equation of this method is as follows (Zaki & Meira, 2020):

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

Where:

p(C=c|X=x):	refers to the conditional probability a posteriori
C:	refers to the dependent variable;
X:	is the conditional variable or attribute

According to Zaki and Meira (2020), the a posteriori probability of any hypothesis consistent with the training data set can be estimated to choose the most likely hypothesis.

Given an example x represented by k values, the Naïve Bayes classifier is based on finding the most probable hypothesis describing that example.

If the description of that example is given by the values $<a_1, a_2, ..., a_n>$, where a_i are the possible attributes of the target variable, the most probable hypothesis will be the one that fulfills:

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

That is, the probability that the known values describing that example belong to the class v_j , where v_j is the value of the classification function f(x) in the finite set V.

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

$P(v_j)$ can be estimated by counting the times the example v_j appears in the training set and dividing by the total number of examples in that set.

The constructed algorithm used in the RStudio program is as follows:

Partición <- createDataPartition(y = datos_train_prep$Popularidad, p = 0.7, list = FALSE)

```
SP_entrena <- datos_train_prep[Particion,]
SP_test <- datos_train_prep[-Particion,]
```

As already mentioned, the objective of this work is to identify the most efficient predictor model to avoid the two statistical errors (type 1 error and type 2 error) when analyzing the popularity of brands as a response variable, considering their financial, economic, and technological variables (presence in social networks). In particular, the aim is to find the model that offers the highest proportion of true positives in the form of so-called "accuracy," i.e., the best predictor model (Fleuren et al., 2020).

Data selection and results

The database comprises data from the 50 most valuable brands in the world, given by the consulting firm Interbrand in 2020 (Interbrand, 2020). The variables used in the database are shown in Table 1:

Table 1
Proposed variables for the construction of the sample database

	Variable	Description
1.	Brands	The most valuable in the world, according to the consulting firm Interbrand in 2020
2.	Value	The brand value according to Interbrand's 2020 valuation methodology based on financial factors and brand strength
3.	Sector	The sectors to which the brands belong, such as Technology, Beverages, Automotive, Restaurant, Media, Business Services, Sporting Goods, Luxury, Financial Services, Logistics, Retail, Diversified, Alcohol, Fast Moving Consumer Goods, Apparel
4.	Country	Place of origin of trademarks such as USA, South Korea, Japan, Germany, France, Sweden, Italy, Spain, Switzerland
5.	Sales	Total sales provided in 2020 by brands and reported by the Economatica database
6.	Profits	The total 2020 profits reported by the brands and reported by the Economatica database
7.	Share price	The share price as of the close of 2020, as reported by the brands and by the Economatica database
8.	Facebook	The number of likes on each page of the corresponding brand given by Facebook by the Meta platform
9.	Instagram	The number of likes on each page of the corresponding brand given by Instagram by Meta platform
10.	Twitter	The number of likes on each page of the corresponding brand given by the Twitter platform
11.	Popularity	Popularity, being the response variable, was generated from the average number of sales and the number of likes on the platforms; thus, each brand's popularity level was assigned.

Source: created by the authors.

As seen in Table 1, the database is composed of 11 variables, which, for this paper, refer to the concept of popularity, as discussed in the literature. It is important to note that unlike the concept of popularity established in the traditional literature, which refers to marketing aspects (Aggrawal, Ahluwalia, Khurana, & Arora, 2017), this work refers to a set of variables considering criteria referred to by Interbrand (2020) and having as a framework of analysis the elements suggested by the ISO 10668 (2010) and ISO 20671 (2019) regulations on brand valuation and evaluation, respectively.

Among the eleven variables proposed, three are qualitative: sector, country, and popularity, and the remaining eight are quantitative. The temporal factor of the data used is annual and refers to 2020.

Regarding the qualitative variables, they were coded as follows:

Table 2
Sector and coding variable

Sector	Coding
Technology	1
Beverages	2
Automotive	3
Restaurant	4
Mean	5
Business Services	6
Sporting goods	7
Luxury	8
Financial Services	9
Logistics	10
Retail	11
Diversified	12
Alcohol	13
Immediate consumption	14
Apparel	15

Source: created by the authors with data from Interbrand 2020.

Table 3
Country and coding variable

Country	Coding
United States of America (USA)	1
South Korea	2
Japan	3
Germany	4
France	5
Sweden	6
Italy	7
Spain	8
Switzerland	9

Source: created by the authors with data from Interbrand 2020.

The popularity variable was coded according to the abovementioned popularity level, giving binary values where 0 corresponds to low popularity, and 1 corresponds to high popularity.

The steps followed in the development of the models using RStudio software were as follows:

a. Exploratory analysis of the database
b. KNN Model
c. Classification tree model
d. Naïve Bayes Model

Results

According to the methodology used, the following results were obtained:

Exploratory analysis

The exploratory analysis consisted of examining the database's distribution, cleaning, and preprocessing to ensure the proper use of the proposed models. The data are based on the 50 most valuable brands in the world according to Interbrand (2020) and the 11 variables described in Table 1. The first step was ensuring that no null values were in the database.

Having quantitative and qualitative variables, the variables sector, country, and popularity were transformed with the code "as.factor" and named according to the number they belong to. Subsequently, the existence of null values was identified, where in this base and with the previous cleaning, it was verified that there were no null data. The distribution of the response variable is visualized as follows. (See Figure 1).

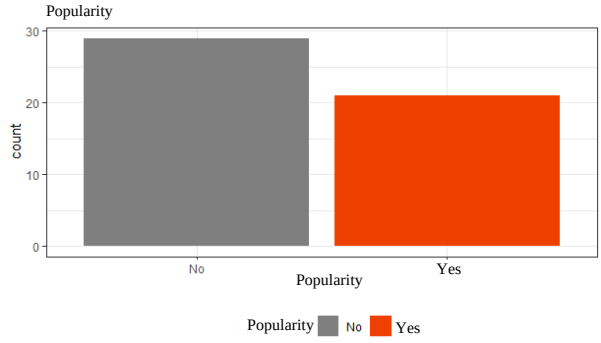


Figure 1. Distribution of the response variable
Source: created by the authors, using RStudio software.

Figure 1 shows that out of the 50 brands given by the standards for calculating popularity, most brands are not very popular, and the rest are. This magnitude can be visualized in Table 4 for a more precise view.

Table 4
Popularity distribution

No popularity	Popularity
29	21

Source: created by the authors.

This statistic indicates that 58% of the total sample are not considered popular brands, and 42% are considered popular.

Regarding the frequency distribution of the country variable, 29 brands belong to the United States of America, the highest number, followed by Germany with 7 brands and France with 5 brands. On the other hand, in the frequency distribution by sector, the brands with the highest presence are those of the automotive sector, with 9 brands, followed by technology and financial services brands with 6, and business and media services brands with 5.

The following figures, showing the distribution of the continuous variables for each variable, were generated to show the interaction of the popularity response variable with the other continuous variables.

Figures 2 to 7 show two ways of visualizing the relation between the proposed variables and the target variable, which is popularity: the first is in terms of smoothed empirical probability density (on the left side), where the distribution of quantitative data in a continuous time interval or period is visualized; the second, on the right side, is in the form of a boxplot or box and whiskers and shows the distribution of data in quartiles, highlighting the average and the outliers. Both show the analyzed data distribution, and trends and dispersions are observed. An integrated interpretation of what these figures represent is given below.

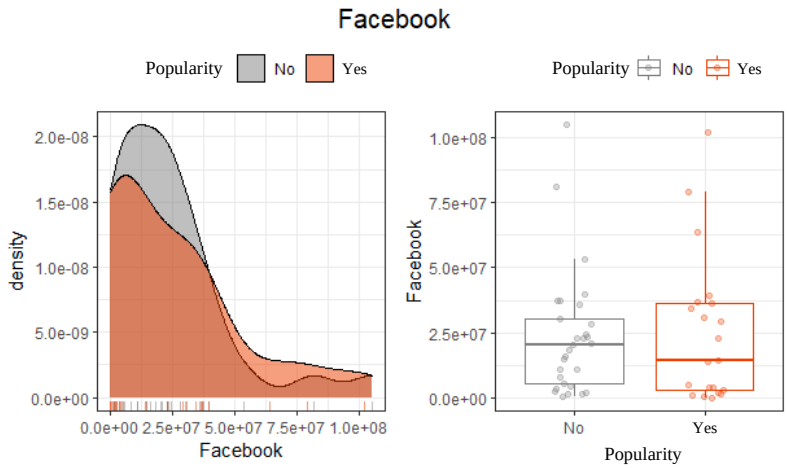


Figure 2. Distribution of continuous variables (popularity and Facebook)
Source: created by the authors, using RStudio software.

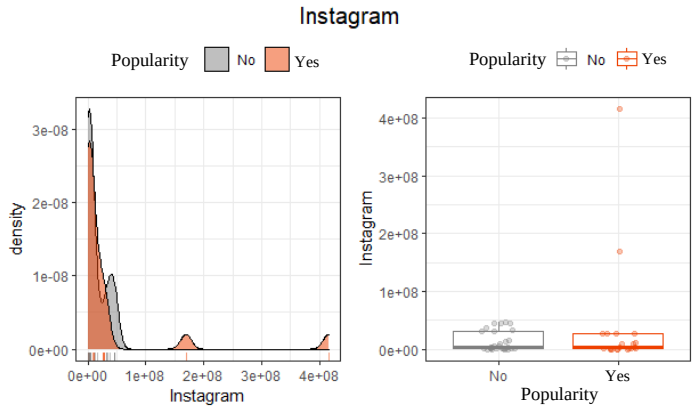


Figure 3. Distribution of continuous variables (popularity and Instagram)
Source: created by the authors, using RStudio software.

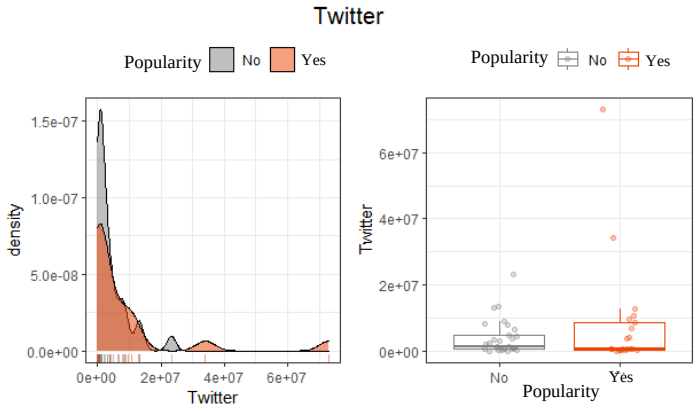


Figure 4. Distribution of continuous variables (popularity and Twitter)
Source: created by the authors, using RStudio software.

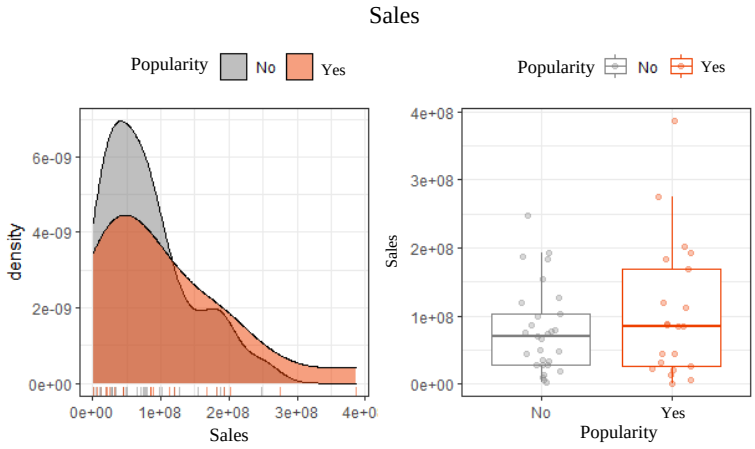


Figure 5. Distribution of continuous variables (popularity and sales)
Source: created by the authors, using RStudio software.

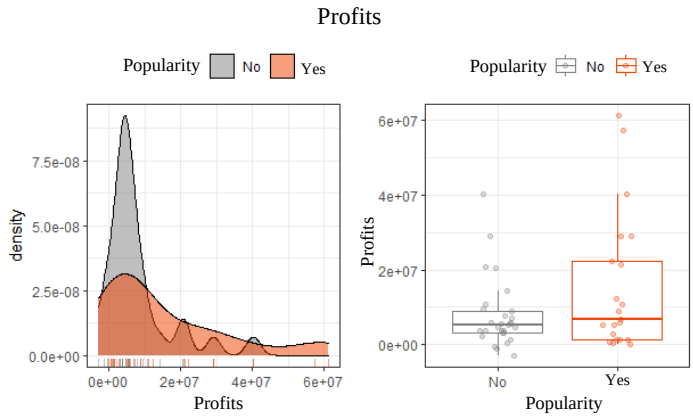


Figure 6. Distribution of continuous variables (popularity and Profits)
Source: created by the authors, using RStudio software.

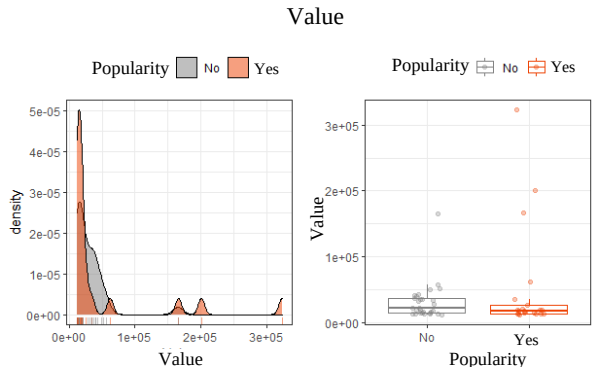


Figure 7. Distribution of continuous variables (popularity and value)
Source: created by the authors, using RStudio software.

Considering Figures 2, 3, and 4, it can be observed that the level of popularity given by the social networks of the Facebook, Instagram, and Twitter platforms does not present a normal distribution, and the curve is steeper toward the left. On the other hand, outliers are observed according to the box plot. It is worth mentioning that the Twitter platform, when observing the box plot, concentrates more values within the box since it is considered one of the most widespread social networks among users and companies.

On the other hand, analyzing Figures 5, 6, and 7, it is found that the popularity and sales variable shows a distribution almost similar to the normal distribution, given the large volume of inventory that certain companies handle and the great impact they have on society. The same effect is shown for profit;

nevertheless, it shows a normal distribution when the companies are generating profits and have popularity, considering the different cost and expense structures of the companies analyzed. Regarding the brand value based on its popularity, a normal distribution is not observed due to outliers that exceed the sample mean of the brand value, which can be seen in more detail in the box plot.

The distribution of the qualitative variables is expressed as follows. (See Figure 8 and Figure 9).

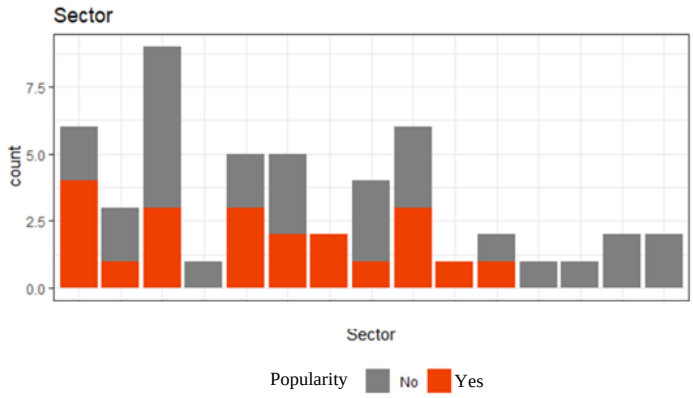


Figure 8. Distribution of qualitative variables (popularity and sector)
Source: created by the authors, using the RStudio program. The sector goes in order according to the list in Table 2.

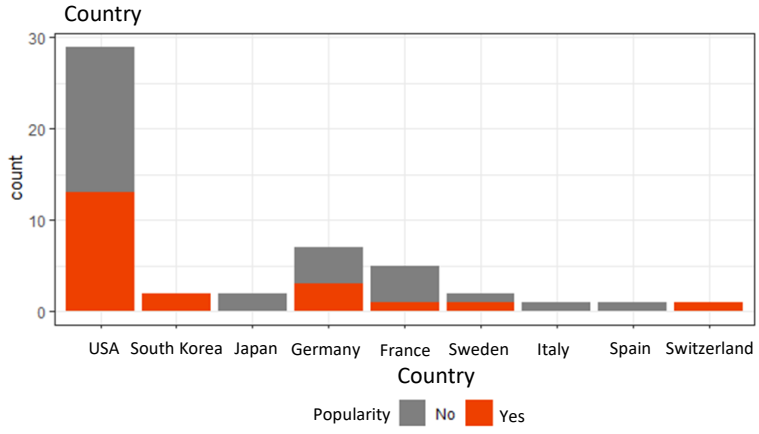


Figure 9. Distribution of qualitative variables (popularity and country)
Source: created by the authors, using the Rstudio program.

According to Figure 8, the technology, business, and financial services sectors are more popular, while the beverages, automotive, restaurants, diversified, and immediate consumption sectors are less popular.

Concerning the country, Figure 9 shows a higher popularity of brands in South Korea, Switzerland, the United States of America, and Germany. In comparison, there are no significant levels of brand popularity in countries such as Japan, France, Italy, and Spain.

For the importance of quantitative variables concerning the authors' database, the following results were found. (See Figures 10 to 13).

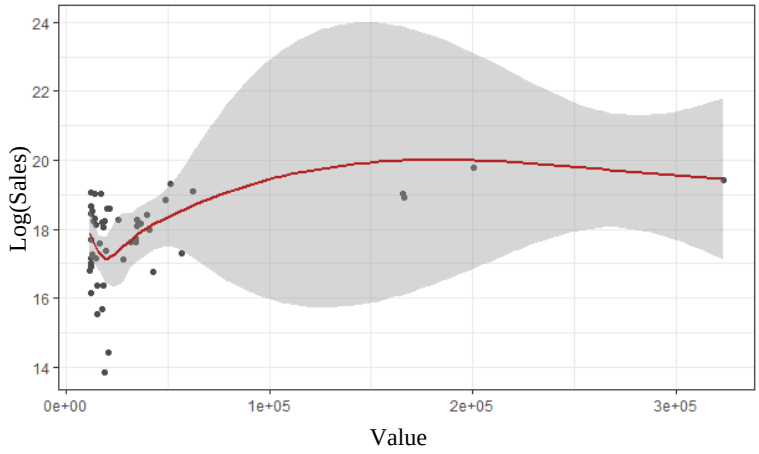


Figure 10. Importance of quantitative variables (Value and Sales)
Source: created by the authors, using RStudio software.

The variable value concerning sales correlates to 0.6296, meaning $t = 5.61, p < .05$ [(4)].

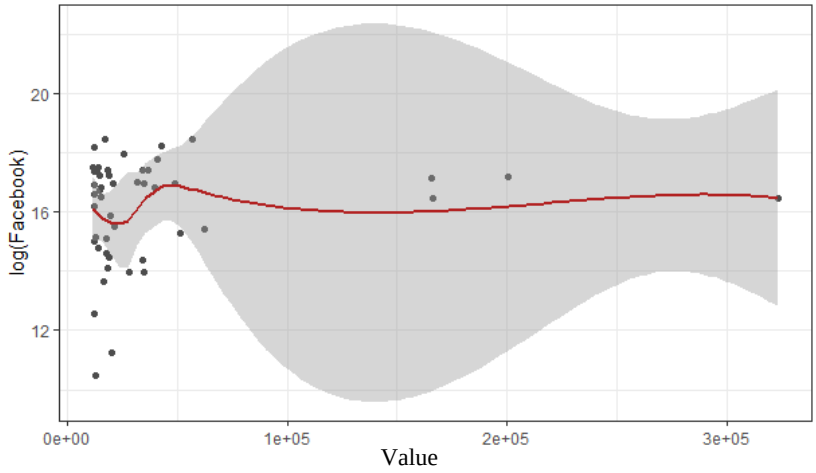


Figure 11. Importance of quantitative variables (Value and Facebook)
Source: created by the authors, using RStudio software.

There are no relevant results regarding the value of the Facebook variable, and there is also a low positive correlation of 0.0058.

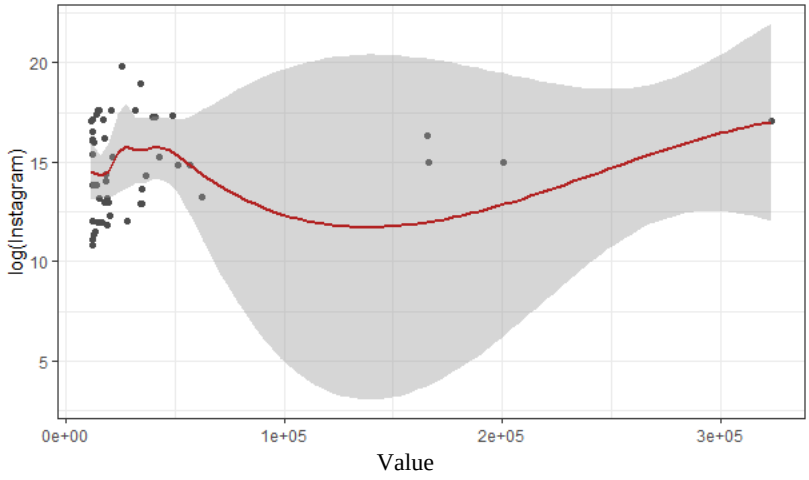


Figure 12. Importance of quantitative variables (Value and Instagram)
Source: created by the authors, using RStudio software.

No relevant results are presented for Instagram, and a negative correlation of -0.0202 is obtained.

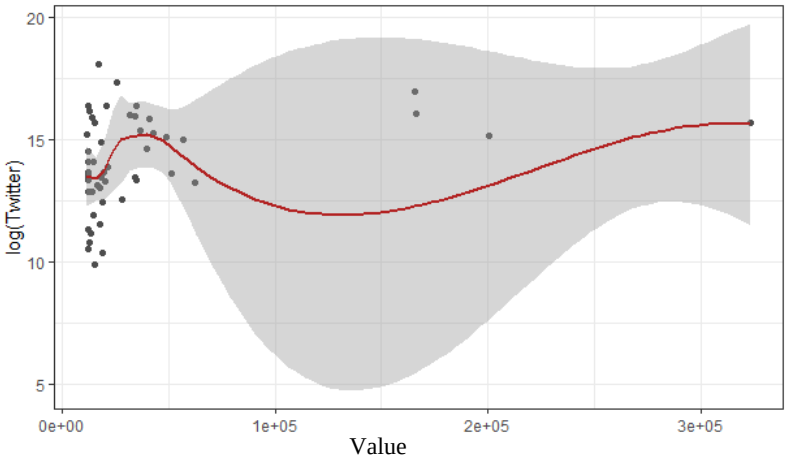


Figure 13. Importance of quantitative variables (Value and Twitter)
Source: created by the authors, using RStudio software.

Finally, there is no statistical significance for the Twitter variable, but there is a correlation of 0.0926.

As a preliminary conclusion of the results generated up to this point of the exploratory analysis, it is perceived that there is a relation between quantitative and qualitative variables. It is also worth mentioning that, according to the graphs generated, the correlation between the variables related to monetary aspects and preference in social networks presents atypical results, i.e., from the statistical inference, values irrelevant to each other have been generated. Nevertheless, machine learning algorithms can process data that would not be relevant to traditional statistics (Lara, Mora, & Londoño, 2022).

Segmentation of training and test data

The variables were preprocessed, and the following transformations were made: For the continuous variables, a normalization process was performed, which consists of working on the same numerical base. As for the qualitative variables, such as popularity, a binarization process (0,1) was carried out to avoid outliers in the database.

K-nn model

The sequence followed to apply this model was as follows:

- Application of the model on the non-preprocessed database, i.e., without the variables being normalized.
- Application of the model on the non-preprocessed database, but taking as a starting point the results of the application of the previous model;
- Finally, based on applying the two models mentioned above, the normalized data are used to obtain the optimal model that will serve as the best predictor model based on its accuracy.

Figure 14 illustrates the operation of this classification method, where 50 samples belonging to two different classes are represented: Class 1, formed by those who are popular, and Class 2, formed by those who are not popular. In this case, the model yielded a result of twenty-three neighbors, i.e., k=23 of the 50 samples, which is sufficient to perform the predictive analysis.

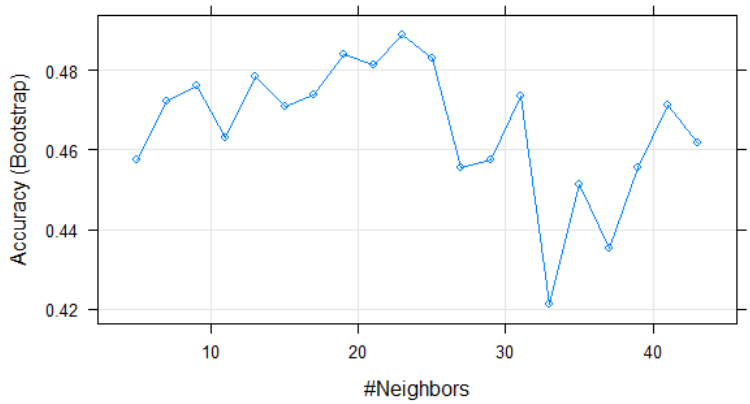


Figure 14. K-nn model
Source: created by the authors, using RStudio software.

In order to observe the data's behavior concerning the target variable, the model was first applied to the database without the variable transformations, i.e., with the non-preprocessed data. Table 5 shows the results of applying the model to the non-preprocessed data.

Table 5
K-nn results (non-preprocessed data)

No preprocessing
Resampling: Bootstrapped (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

Source: created by the authors, using RStudio software.

Starting from the previous result of non-preprocessed data, which showed a value of k= 23, the following model is run with non-preprocessed data, resulting now in the nearest neighbor with k = 43. This means that increasing the k level captures more variables corresponding to the target attribute: popularity. (See Figure 15 and Table 6).

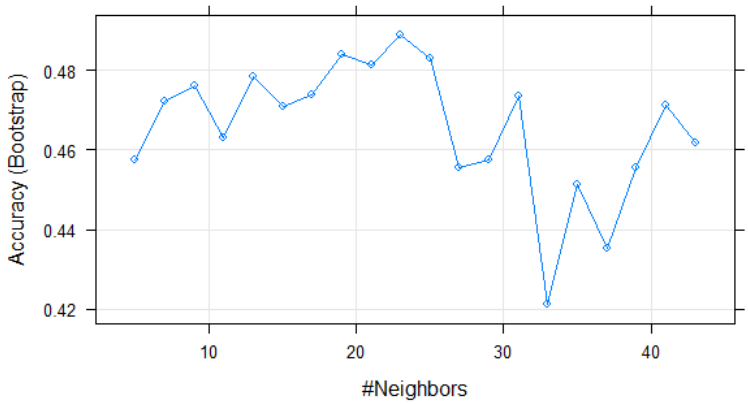


Figure 15. Model K-nn k=43
Source: created by the authors, using RStudio software.

Table 6
Results K-nn k=43

No preprocessing
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

Source: created by the authors, using RStudio software.

It is worth mentioning that the accuracy is better with a higher number of k than a with a lower value. As previously mentioned, non-preprocessed data were used.

Derived from the previous results, better results are obtained when considering a value of k=43 despite using non-preprocessed data. Next, the model was applied with the data already normalized or processed, i.e., considering the data that are or are not related to the target variable, which is popularity. Therefore, they are divided into two classes: have or do not have popularity (see Figure 16).

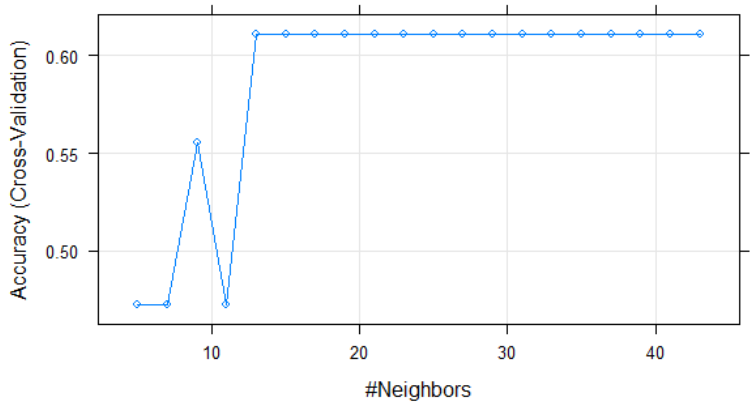


Figure 16. K-nn Model (processed data)
Source: created by the authors, using RStudio software.

The K-nn method assumes that those considered nearest neighbors generate a better ranking using the attributes. In this case, the target attribute “popularity” was the only variable used; next, the prediction analysis was performed, as shown in Table 7.

Table 7
K-nn prediction results

	No	Yes
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

Source: created by the authors, using RStudio software.

According to these ten samples, there is a greater probability, based on the data obtained, of 57.69%, that there is not so much popularity given the social networks and sales, and 42.31% probability that there is popularity given the social networks and sales. (See Table 8).

Table 8
Results Accuracy

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

Source: created by the authors, using RStudio software.

According to the confusion matrix, Accuracy of: $\frac{6+0}{(6+0+0+4)} = 0.6$ (5)

To this end, it will be compared with the following models to determine which is optimal.

Classification tree

A classification tree was generated using popularity as the response variable and all available variables, such as sales, as predictors. (See Table 9 and Figure 17).

Table 9
Classification tree

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

Source: created by the authors, using RStudio software.

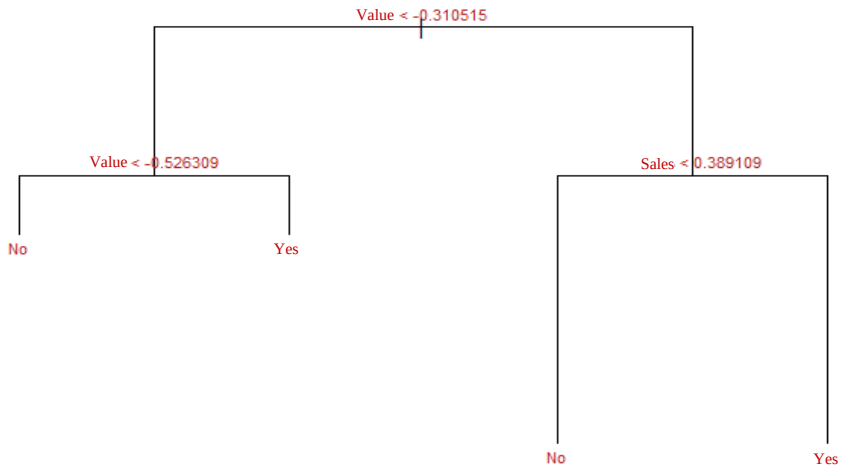


Figure 17. Classification tree model
Source: created by the authors, using RStudio software.

Subsequently, the pruning process[2] compares the results with the initial model shown in Table 9. The summary function () shows the adjusted tree has 4 terminal nodes and a classification training error rate of 23.08%. Residual mean deviance[3] shown in the summary is the residual deviance divided by (number of observations - number of terminal nodes), resulting in 1.021. The lower the deviance, the better the fit of the tree to the training observations; in this case, the deviance is higher, and therefore, the fit of the tree will not be as good. (See Table 10).

[2]The recursive binary splitting process can achieve good predictions with the training data, since it reduces the training RSS (Residual Sum of Squares), which implies an overfitting to the data (derived from the ease of branching and possible complexity of the resulting tree), reducing the predictive capacity for new data. Retrieved from https://rpubs.com/

[3]The residual mean deviance is that measure of error remaining in the classification tree after construction. It corresponds to the RSS (Residual Sum of Squares) training variable, divided by the number of observations minus the number of nodes. The lower this value, the better the model fits the training data. Retrieved from https://rpubs.com/

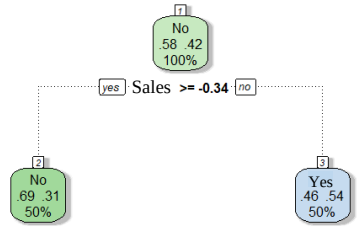
Table 10
Predictions

predictions	No	Yes
No	0	0
Yes	8	6

Source: created by the authors, using the Rstudio program.

According to the confusion matrix, accuracy of: $\frac{0+6}{0+0+8+6} = 0.4286$ (6)

By generating the "pruning" process, an optimum of 2 nodes is reached. Therefore, based on the sales, it is possible to have yes or no popularity, as shown in the following image. (See Figure 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
```

Figure 18. Classification tree with 2 nodes
Source: created by the authors, using the Rstudio program.

Naïve-bayes

Being a classification model in Data Mining, where it is classified if the sample has popularity or not, in this case also called instances, it will be characterized by a series of attributes. Nevertheless, by having in the model only one attribute of having or not having popularity concerning having brand value, the result is that it is the status with the highest accuracy, given that it only has the attribute mentioned above and the other qualitative variables are not considered, sales being the only causal variable. To this end, its degree of importance is 100, and it has popularity since it is the only attribute considered for the analysis. (See Table 11).

Table 11
Naïve-Bayes Results

26 samples
3 predictor
2 classes: `N", `Ye"
No preprocessing
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

Source: created by the authors, using RStudio software.

In order to identify and evaluate the models with the highest proportion of true positives in their results, the so-called Receiver Operating Characteristic (ROC) curve is used. It consists of a graphical representation of the classifier's performance, showing the distribution of the fractions of true and false positives. The goodness of a diagnostic test that produces continuous results is known as the area under the curve (AUC), which represents the probability that, in the case of the research, the popularity of brands has a positive or negative effect on their economic value. Figure 19 shows the area under the curve resulting from the selected variables. The value of the area under the curve (AUC) is 0.62, which indicates the minimum level accepted for the prediction models described above. Based on this criterion, the model that comes closest to this criterion is the K-nn model because it has a wider use of the database, which provides a better appreciation of the behavior of the proposed variables.

Nevertheless, it can be preliminarily concluded that applying this K-nn and the other two models requires using variables other than those considered in this study to comply with the minimum acceptable parameters.

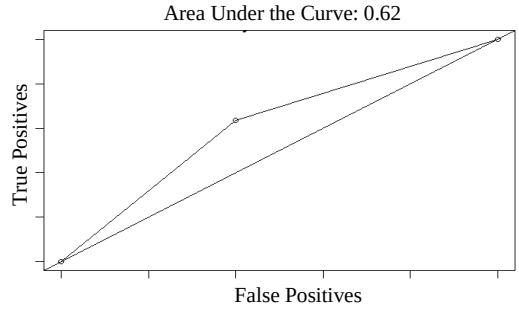


Figure 19. ROC (Receiver Operating Characteristic) curve
Source: created by the authors, using RStudio software.

Conclusions

Regarding the objective established as to whether the variables proposed in the study influence brand popularity, it is concluded that they all positively influence to different extents. Concerning the popularity of brands and social networks, it is perceived that the effect is greater on Twitter because it has a greater presence among opinion leaders and represents a formal and official means of communication between companies and consumers. In other aspects, variables such as brand value and sales have an important effect on popularity, considering that the existence of leading global companies and brands and their development trends are increasingly disseminated in the media. Moreover, variables related to more technical and specialized aspects of company management in financial terms, such as profits, have a lesser effect.

Regarding the influence between brand equity and social network variables given as "likes," it was identified that elements related to brand equity and indicators of some degree of popularity contribute to the construction of powerful brands.

The close relation between brand equity and sales is relevant, as they have a broad correlation. Nevertheless, it is advisable to investigate what other factors are considered in the additional impact on what makes a brand valuable and popular (Acuña Moraga & Severino-González, 2018). As indicated by González, Orozco, and Barrios (2011), the relation is based on a set of dimensions such as knowledge, relation, and attitude toward the brand, as well as brand preference from different levels of involvement in the purchase process. In order to understand the relation between brand popularity and consumer evaluation of attributes, brand preference, and brand loyalty, it is observed that brand popularity is not only associated with sales given by the brand (Boix, Boluda, & López, 2019).

Due to such findings, for further studies and to provide further explanation, the use of other types of variables is proposed, as indicated by García Granda and Gastulo Chuzónen (2018), in terms of data collection that indicate the degree of loyalty, reputation, and liking of a brand. Likewise, according to Tapia Cedeño (2017), using a satisfaction variable generates a favorable point as a popularity factor.

It can also be affirmed that considering a set of valuable brands in the global market, there are effects on their value with respect not only to their sales but also to the presence of relevant internal and external factors in the management of the brands as assets of the companies. These effects vary among the types of market sectors in which the brands operate, which leads to considering the existence of a degree of popularity given by the companies based on the management of their brands.

In Mexico, it is important to analyze these effects, such as brand popularity and brand value in global markets, to identify parameters for comparing the policies of leading global companies and those operating in national markets with internationalization tendencies. The characteristics of brand

management and its value have an important relationship in the perspective of behavior and conduct both individually and collectively and at corporate levels of companies.

References

Acuña Moraga, O., & Severino-González, P. E. (2018). Sustentabilidad y comportamiento del consumidor socialmente responsable. Opción: Revista de Ciencias Humanas y Sociales, Opción, Año 34, No. 87 (2018): 299-324 ISSN 1012-1587/ISSNe: 2477-9385. Available in http://repositorio.ucm.cl/handle/ucm/2450 Consulted 09/05/2022.

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. Available in: http://hdl.handle.net/2133/21727. Consulted 09/05/2022.

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. Available in: https://hdl.handle.net/20.500.12008/10558. Consulted 09/05/2022.

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). Available in: https://oa.upm.es/71723/. Consulted 09/05/2022.

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. Available in: http://repositorio.uchile.cl/handle/2250/179714. Consulted 09/05/2022.

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 mismoherramientas 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. Available in: http://hdl.handle.net/20.500.12423/1039. Consulted 09/05/2022.

González, E., Orozco, M., & Barrios, A. (2011). El valor de la marca desde la perspectiva del consumidor. Revista Contaduría y Administración, 235, 217-239. Available in: https://www.redalyc.org/pdf/395/39519916011.pdf. Consulted 09/05/2022.

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. Available in: https://interbrand.com/thinking/best-global-brands-2020-methodology/. Consulted 09/05/2022.

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. Available in: https://revista.ucsa-ct.edu.py/ojs/index.php/ucsa/article/view/30. Consulted 09/05/2022.

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. Available in: www.usask.ca/education/coursework/802papers/mergel/espanol.pdf. Consulted 09/05/2022.

Ni, Z. (2022). Sistema de extracción de datos (Doctoral dissertation, ETSI_Informatica). Available in: https://oa.upm.es/71408/ Consulted 09/05/2022.

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). Available in: https://oa.upm.es/68484/ Consulted 09/05/2022.

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. Available in: https://orcid.org/0000-0002-1888-0451 http://orcid.org/0000-0002-6103-6993. Consulted 09/05/2022.

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). McGrawHill. 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. Available in: https://www.proquest.com/docview/2388304894?pqorigsite=gscholar&fromopenview=true Consulted 09/05/2022.

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. Available in: http://www.scielo.org.bo/scielo.php?pid=S1994-37332016000100005&script=sci_arttext Consulted 09/05/2022.

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. Available in: http://repositorio.sangregorio.edu.ec/handle/123456789/365. Consulted 09/05/2022.

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. Available in: https://users.monash.edu/~webb/Files/YangWebb02a.pdf. Consulted 09/05/2022.

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