

Emotion Analysis in Gasoline Consumption in Mexico using Machine Learning

Dafne Rosso Pelayo, Joel Armando Colín Pacheco, Luis Miralles-Pechuán

Universidad Panamericana, Facultad de Ingeniería,
Mexico

drosso@up.edu.mx, jacolinp@yahoo.com, lmiralles@up.edu.mx

Abstract. Emotions and Sentiment Analysis has had an important role in increasing business benefits on commerce sector. Emotions Analysis as well as Sentiment Analysis is a common machine learning technique used to analyze opinions of people about certain company aspects such as products image, product consumption, marketing campaigns, client's preferences and social or political movements. The relevance of Emotions Analysis research lies in the enormous economic impact that it provides to enterprises. In this work, we present an emotion analysis to obtain the principal feature set related to the emotions that make consumers prefer a gas station over others. Our approach, to understand gasoline consumption behavior in Mexico, is based on machine learning and statistical analysis. We use a conventional statistical approach to analyze the characteristics of gas stations preferred by customers based on their emotions. Finally, supervised Machine Learning classification methods are applied in order to predict the probability that a gas station is selected on the basis of customers' emotions.

Keywords: Gasoline consumption, machine learning, emotions.

1 Introduction

Emotions are sensations generated by feelings and perceptions. Emotions are accompanied with thinking and actions. Emotions never exist alone; they are always accompanied by thoughts [1]. Emotions are the result of brain stimulation; they can be originated by a remembrance, an action, an observation, a comment or any other stimulus that impacts feelings or perceptions of people.

Emotions reflect an important impact in client's decisions [2, 3, 4, 5, 6]. When emotions are positive, the likelihood of finding positive action from clients is higher; otherwise when the emotion is negative the likelihood of finding positive action is lower.

Antonio Damasio, neurologist and director of the Brain and Creativity Institute (BCI), formulated a theory about how emotions and their biological underpinnings are involved in decision-making. His main field of study is the neurological system based

on memory, language, emotions and decision-making process. Damasio in [7] shows that "basic emotions like happiness, sadness, shame and empathy, are a set of complex chemical responses as neurological both forming a distinctive pattern" [8, 9].

In this work we present an emotion analysis to obtain the principal features and characteristics of a gas station that produce an emotion and make impact over the selection process of the clients.

To this end, we analyze customer's preferences on gas stations from two different approaches. First, a classical statistical analysis is developed in order to explore the collected information and to observe how the gas station selection process can be affected by perceptions. Perceptions are originated from customer experiences when they consume gasoline.

Secondly, we perform a supervised classification ML model to identify patterns that make people select a particular gas station based in their emotions. Emotions are generated when people consume the gas station's services. Several ML models were applied; their results and comparative accuracies are presented in results section.

The paper is organized in 6 sections. Section 1 is for this introduction. In section 2 we present related work about emotion analysis. Section 3 describes the conventions used in the research. In section 4 we describe acquisition and preprocessing of data; we include the experiments and results: we achieve a classical statistical analysis to model client's behavior based on clients' perception of gas stations, and we present a supervised ML approach to discover the emotions and the factors that most influence on the gas station selection process. In section 5 conclusions and future work are presented. Finally section 6 lists the references.

2 Related Works

There is a lot of research on how emotions affect customer preferences [10, 11]. However, few investigations focus on gasoline consumption.

What drive consumers to prefer a gas station over another? At first sight one may think that it is the price of gasoline at that station. In [4] Turrentine, Kurani and Heffer conducted interviews to people who had recently bought a vehicle, giving more attention to the context and validity of the information than to its statistical reliability.

They found that consumers do not budget, manage or track fuel costs. They also discovered that fuel economy decisions are based more by emotions than by critical analysis, and that are more influenced by social awareness than by its monetary value. Buyers of hybrid vehicles did not think of a payback when they made the purchase but they paid a lot of attention to fuel economy [4].

This social awareness is also related to social responsibility on environmental issues. In [5] it is concluded that consumer behavior regarding gasoline products is affected substantially within environmentally sensitive target groups. In their research they used a survey method to collect data regarding three gasoline brands (one green brand, one non-green brand and one partially green brand) and analyzed them with multiple regression analysis. Their results show a pattern that favors gasoline brands perceived as environmental friendly and that even people are willing to drive far in

order to locate their favorite station. They did not find any gender consumer behavior difference.

Emotional experience with product's brands is also an important element that affects customer choice. Experience, if fully processed, is stored in different parts of our memory; including the emotion associated with it. In [6] Hansen, Christensen and Lundsteen state that when a memory is recalled, all of its components get together and the emotional association with the brand comes up too; these emotional responses are the frame of conscious cognitive process.

According to them, if emotional associations with a brand recalled in memory are understood then marketing communication that triggers emotional responses consistent with the brand should help to reinforce positive brand attitudes as well as purchase intentions. Toward that end a study was undertaken where they looked at 16 categories (gasoline included) covering 64 brands and then they recruited a random sample of consumers who answered a questionnaire. To measure the emotions associated with brands they developed a set of scales and then they used factor analysis to reduce them and make them reliable and easily used in advertising and brand research. This resulted in 2-factor solutions that reflect a strong positive – negative distinction. Scores were then calculated for each subject based upon the intensity of felt emotion (on a 6-point scale) and the factor loadings to produce a positive and negative score for each brand and category. Their results showed important insights into how people perceive brands and the emotions associated with that perception.

ML techniques are being widely applied to discover emotions people reflect in their text work, their speech and music, their facial expression or their body movement and so on. Those efforts are mainly focused on discovering what the subject felt or intended to transmit when creating his work [12].

Here we focus on finding and predicting the emotions that elicit the actual features of gas stations in Mexico that make a difference in gas station selection.

3 Conventions

The following conventions are used in our research. We can express the customer behavior in the following way: the independent object's features x_i derives in a particular emotion e_i . Even more, the combination of dependent or independent features of objects x_i can lead to other particular emotions, represented by $e_j \cup \dots \cup e_m, \forall e_i \in E$ where E is the set of possible emotions.

In this way client's feelings and perception over any object x_i becomes an input to produce an emotion e_i . The emotions e_i influence clients' actions and decisions. Emotions have been effectively used to analyze customers' behavior and to help business units to develop marketing strategies [5, 6].

Each gas station G_i possesses a feature set X of size m , these features are independent, and each feature x_i has been rated for each client in an online questionnaire. We have defined a set of personal client's features Y , where each $y_i \in Y$ represents a client feature i.e. $y_i = \{age, gender, average gasoline consumption, etc.\}$.

Each feature $x_i \in X$ represents a feature that clients can get from a gas station G_i during the service time, i.e. $x_i = \{\text{clean bathrooms; store; card of rewards, windshield wiper, etc.}\}$.

The surveyed people rated the gas station feature x_i according to their preferences. These features x_i were associated to a particular emotion e_i in an unconsciously way to the responder.

4 Experiments and Results

4.1 Dataset Description

The dataset used in this research was obtained by conducting an online survey. The survey was answered only by Mexican residents that consume gasoline.

The survey was distributed over the internet and many people were invited to answer it through several social networks.

According to the “National Statistics and Geographic Institute” (INEGI), in 2014 Mexico had 35,353,077 vehicles. We represent this population by MV .

Through our survey we obtained from MV a representative sample of the population S of size n , where $S \in MV$, of enough size to obtain statistically significant results, we select a constant value for standard deviation σ , a 95% confidence level ($Z=1.96$), and an error e of 0.056%. The analysis of this article is based on the set S . The sample size responding to a minimal selected 95% of confidence level was estimate according statistical principles with the following formula:

$$n = \frac{Z^2 \sigma^2 N}{e^2 (N - 1) + Z^2 \sigma^2} \quad (1)$$

For every client t_j that answered our survey, we obtain a set of personal features Y and a vector of ranked features R , where r is a ranked feature, $r \in R$. And $r(i,j)$ is the rating given from a client t_j to a gas station feature x_i with $x \in X$, where X represents the set of gas station’s features evaluated. The ratings $r(i,j)$ varies in an interval $[1,10]$ where “1” is the minimum rating to evaluate a gas station’s feature and “10” represents the maximum rating.

Personal features Y of set S were correlated to clients. This correlation analysis helps to understand gasoline consumption behavior and clients’ preferences.

Segmentation and statistical analyses bring us as a result a set of segments and statistician’s patterns about clients’ consumption. In the set S , each gasoline feature x_i is related in a supervised way to a particular emotion e_i . In this way, our data set S can be used to analyze features x_i , which had been rated by clients, and to discover the most relevant emotions that drive clients to make consumption in a specific gas station.

4.2 Statistical Analysis

Our statistical analysis involves an independent variable analysis of clients’ personal characteristics over their consumption using ranked gasoline features x_i . Table 1

represents an example of set Y . This set includes clients' personal features, where each $y_i \in Y$ represents a client characteristic.

Table 1. Personal client features.

Client t_i	Y_i
y_1	Age
y_2	Gender
\vdots	\vdots
y_n	Average gasoline consumption

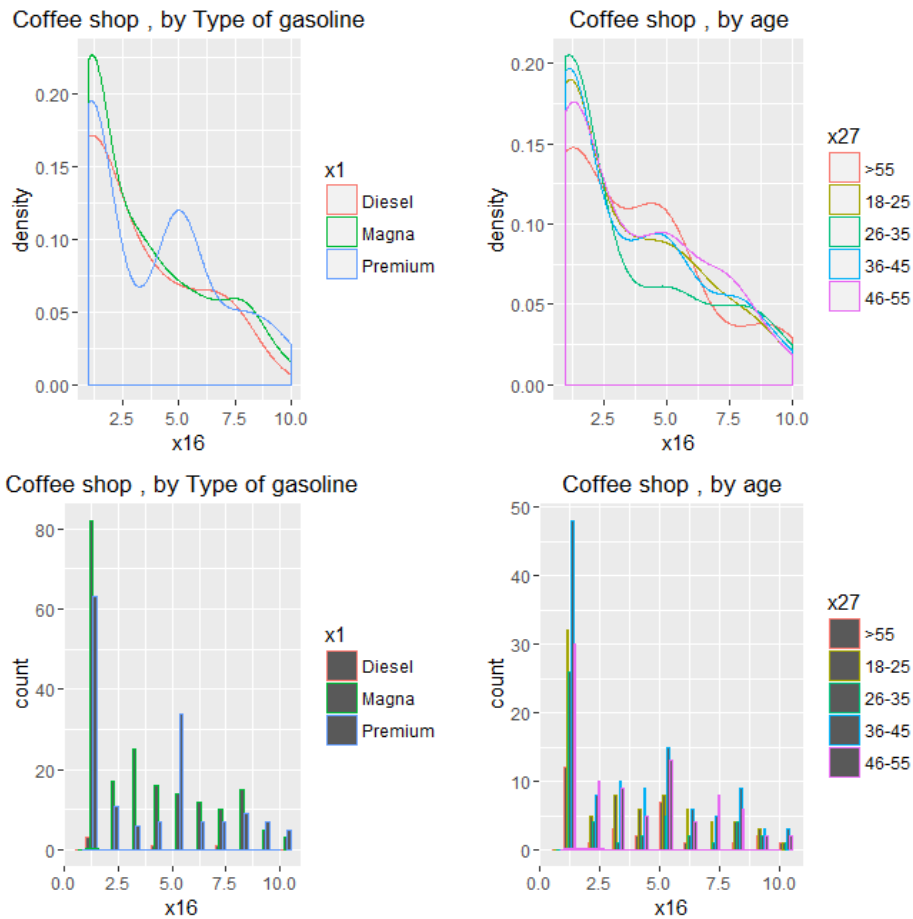


Fig 1. Distinct preferences distribution related to Coffee shops.

The statistical analysis was performed to every gasoline characteristic and personal feature as well as their combinations. Along this article we mention some of them. One

of the first relevant discovered patterns is that local stores or coffee shops are not an important factor for gas station selection. Figure 1 represents the distribution of the

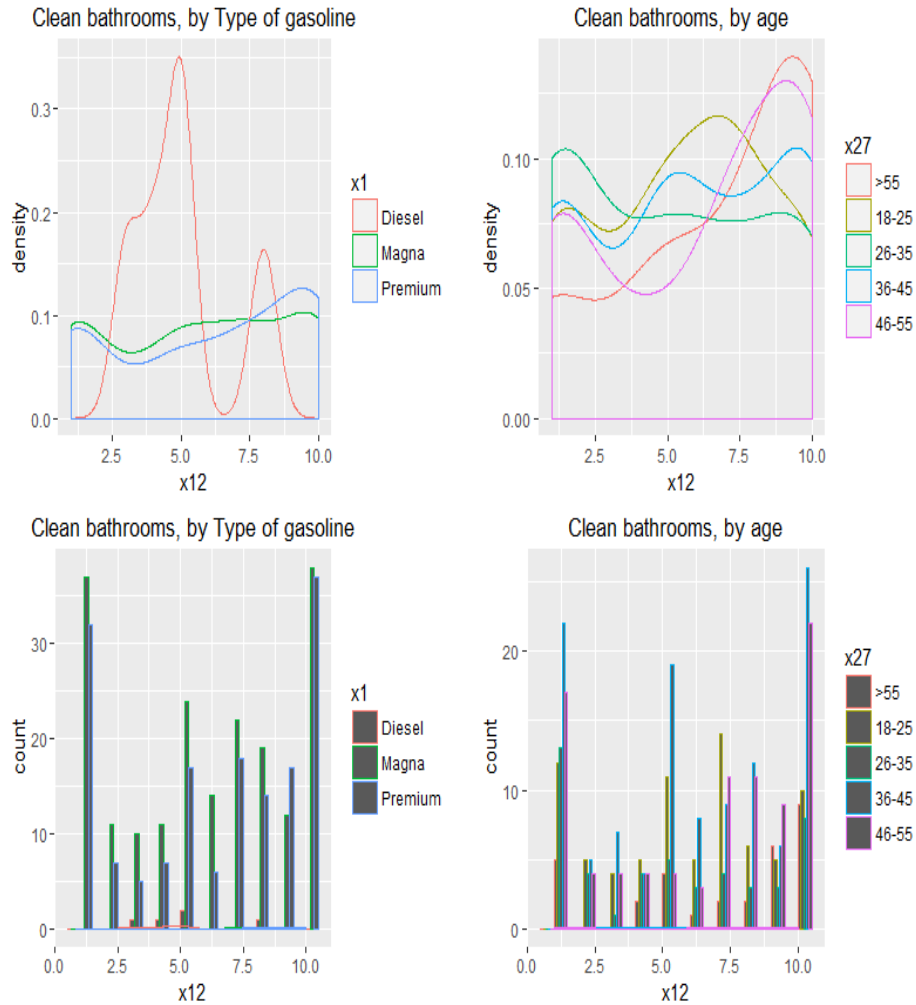


Fig 2. Distinct preferences distribution related to clean bathrooms.

different consumption preferences related to coffee shops.

This kind of analysis shows how clients' preference varies according to their profile. And more important, it identifies those gas station characteristics that are not relevant in any client segment or that are relevant only in a few. Figure 1 shows density distribution of 'coffee shop' feature. This distribution changes for different customer ages and the type of products consumed.

However, we can conclude that regardless clients' age, they do not select a particular gas station based on the 'coffee shop' feature. Statistical analysis of the gas station

characteristics shows that clients prefer other kind of intrinsic characteristics related to gas station instead of additional services as it is shown in Figure 2.

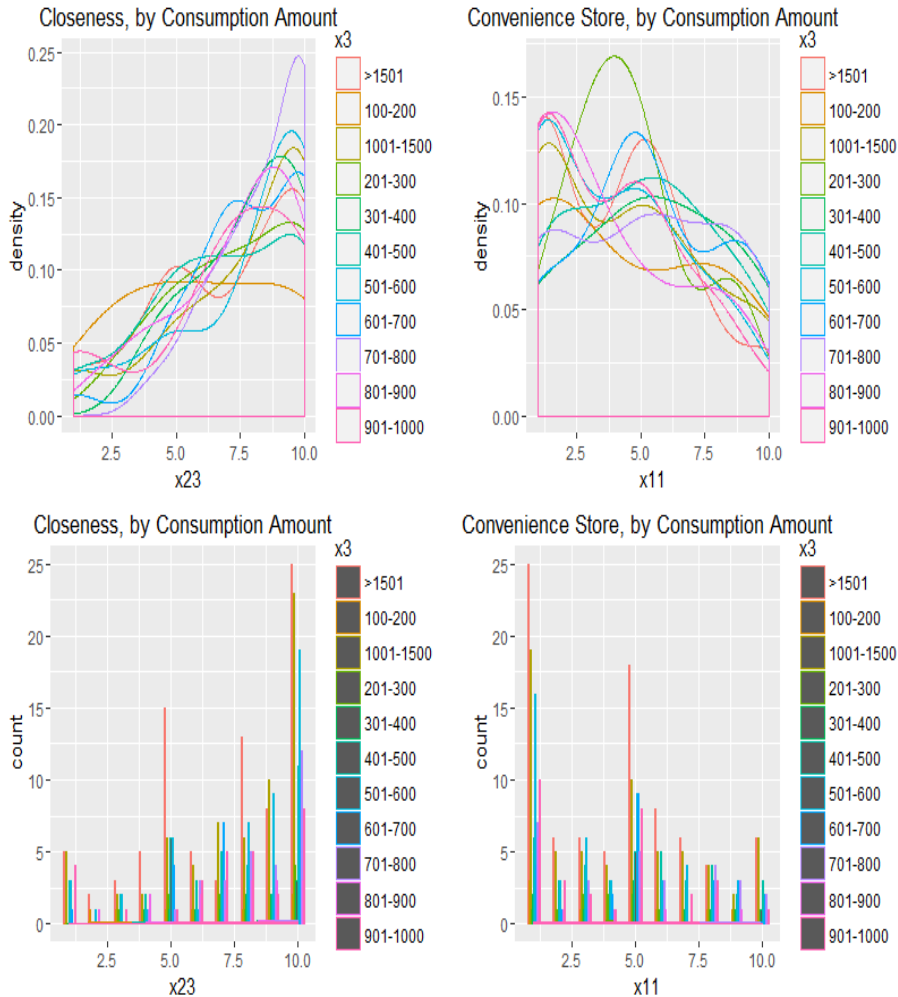


Fig 3. Distinct preferences distribution over gas stations characteristics.

In Figure 2, the analysis of the distribution of the 'Clean bathrooms' feature is very different from the distribution analysis of 'Coffee shop' feature as shown in Figure 1. Although this is not a defining feature to select a gas station, it is a very influential one. As we can see many customers prefer a station with clean bathrooms so the tendency is to select a gas station with this feature.

However, there are some variables that are critical in the selection of the gas station. Figure 3 shows the behavior of these detected variables related to the amount of consumption.

This statistical analysis shows that there are much more important features than others in the gas station selection process. In the second part of our analysis, the emotion variables added in association with the characteristics of gas stations will show not only the features that customers prefer when they select a gas station, but also a set of features that trigger one or more emotions on the client which influences the selection of a particular gas station.

4.3 Data Transformation

Data transformation is a fundamental stage in ML analysis. In this phase, we perform a transformation process over the data in order to be able to mine our data set S . We alter data dimensionality in the following way, for every gasoline characteristic x_i we add an associated emotion e_i , in such a way that personal characteristics are preserved.

The emotion e_i , was mapped to characteristics x_i based in studies that shows that emotions are produced by external stimulus and they produce an instantaneous impact causing an immediate response over people that they can't realize in a complete conscientious way about what is happening at that moment [8]. To distinct between positive and negative emotions we take into account that people exposed to constant external stimulus, incorporate in their neurological patterns those that are beneficial to their life and discard those that are not [10,13]. In our research, best ranking characteristics were taken as accepted for people and were mapped with positive emotions, in other way, bad ranking characteristic were associated with negative emotions, nevertheless all characteristics about gas station services focused to bring a commodity to clients like convenience stores or automatic cashiers were mapped to positive emotions.

Our supervised analysis also requires a class C to classify the items set into selected and not selected. Thus, we add a binary class called "Selection" with two possible values {selected, not selected}. We classified every characteristic in a supervised way based in its ranking, if it is higher than 5.0 then it is a characteristic that will most probably be accepted by clients; in other way it is rejected (not selected).

After the transformation process we call our new data set as S' .

Table 2. Example of client Emotions in S' .

Client t_i	x_i	e_i
e1	Automatic charge	Security
e2	Rapid service	Happiness
⋮	⋮	⋮
en	Illuminated station	Security

It is important to mention that the mapping between emotion e_i and characteristic x_i is constant and has a unique value thorough the analysis. The emotion e_i was assigned according to the type of characteristic, nevertheless the set of emotions of a single client has a particular combination and proportion of emotions and characteristics that make

him to select or not select a specific gas station; thus an ML analysis can tell us with accuracy the probability of a gas station of being selected or not selected.

4.4 Machine Learning Model

The objective of supervised ML model is to predict the selection probability of a gasoline station based on the emotions triggered in clients through the gas station features. The contribution of our model is the novelty of associating the selection probability with the analysis of emotions rather than basing the analysis only on the gas stations feature set.

We accentuate the requirement of an ML model because it emphasizes prediction and can deal with the uncertainty founded in the missing data; it also provides the advantage that its predictions are based on historical data and will remain valid if the population increases, while statistical analysis emphasizes inference and the confidence intervals will not be valid under other parameters of the distribution.

The greater the probability of selection, the greater the acceptance of the gas station according to customer preferences and the greater the realization of future consumption.

From an economic point of view, this type of analysis can bring huge profits on future investments in the construction of new gas stations in Mexico. Building gas stations that trigger emotions in customers that ensure greater acceptance will get preference over others and thus higher economic gains.

Figure 4 shows the correlation between emotions and the class C where blue color represents “Selected” and red color represents “Not selected”. It can be observed how emotions are related to gas station features in the selection or in the rejection of a gas station.

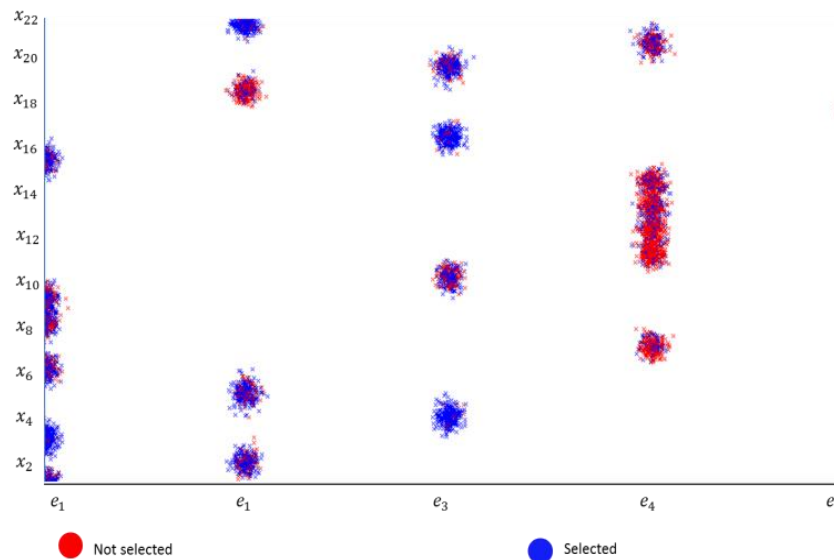


Fig. 4. Correlation between features and emotions.

The classification model can be represented in the following way. Probability P that has a gas station g_i to be selected is given by $P(x_i|\text{selected})$. This is expressed in equation 2:

$$P(g_i|\text{selected}) = \alpha \text{ and } P(g_i|\text{not selected}) = 1 - P(g_i|\text{selected}) = 1 - \alpha. \quad (2)$$

Our supervised model estimates the likelihood of a gas station to be selected or not, which observes a Bernoulli distribution with class C , where $C = \{\text{selected, not selected}\}$ and where the set of features X is discrete. For this module we built a discriminant model also known as a binary classifier r_i^t where $X = \{x^t, r^t\}_{t=1}^N$ and:

$$r_i^t = \begin{cases} 1 & \text{si } x^t \in C_i \\ 0 & \text{si } x^t \in C_j \text{ for } j \neq i. \end{cases}$$

The selection probability is given by:

$$P(x^t|\text{selected}) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}. \quad (3)$$

Thus, the probability of being not selected is:

$$P(x^t|\text{not selected}) = 1 - P(x^t|\text{selected}). \quad (4)$$

Linearizing (1) by transforming $Y = \ln \frac{y}{1-y}$ we have:

$$\frac{\log P(x^t)}{1-P(x^t)} = \alpha + \beta x \quad (5)$$

and then

$$y = \alpha + \beta x. \quad (6)$$

In the next section we present the results obtained for this classification model.

4.5 Results Comparisons

In this section are presented the results of the analysis performed by varying the training set at 80% and taking the rest of the set for testing. The following table shows the best results obtained with the classification ML methods. For the experiments, we used version 3.1.2 (2014-10-31) of R Studio Software. This software has been executed on an Intel R Core TM i5-2400 CPU @ 3.10 GHz with 16 Gb RAM computer, over Windows 7 Pro operating system, Service Pack 1 64 bit.

For the experiments we used the following classification methods: k-Nearest Neighbors (KNN), Random Forest (RF), C4.5-like Trees (C4.5-Trees), Rule-Based

Classifier (RBC), AdaBoost.M1 (AB), Stochastic Gradient Boosting (SGB), Neural Network (NN), Model Averaged Neural Network (ANN), C5.0, Single Rule Classification (SRC), Neural Network (NN), SVM with Radial Basis Function Kernel (SVM), Multivariate Adaptive Regression Splines (MARS), Mixture Discriminant Analysis (MDA), Naive Bayes (NB), Shrinkage Discriminant Analysis (SDA), Penalized Discriminant Analysis (PDA).

Table 4. Results using 80% of the dataset for training and 20% for testing.

N°	Method name	Acc	Sens	Spec	Prec	F1-score	Bal. Acc.	Train time*	Test time*
1	KNN	87.54	81.15	91.91	87.27	89.53	86.53	2.05	0.29
2	RF	84.63	76.79	89.99	83.99	86.88	83.39	5.32	0.12
3	C4.5-Trees	80.77	74.77	84.88	77.17	80.84	79.82	0.45	0.07
4	RBC	78.12	81	76.14	69.89	72.88	78.57	1.06	0.04
5	AB	76.85	69	82.22	72.62	77.12	75.61	9.51	0.55
6	SGB	76.28	66.2	83.17	72.9	77.7	74.69	0.19	0.02
7	NN	75.84	65.89	82.64	72.18	77.06	74.26	0.72	0.01
8	ANN	75.71	63.4	84.13	73.2	78.29	73.76	3.72	0.01
9	C5.0	75.4	61.37	84.98	73.64	78.91	73.18	0.18	0.05
10	SRC	75.08	60.28	85.2	73.57	78.96	72.74	0.23	0.04
11	NN	74.95	62.77	83.28	71.96	77.21	73.03	0.71	0.01
12	SVM	74.45	53.43	88.82	76.56	82.24	71.12	6.36	0.77
13	MARS	73.88	55.45	86.47	73.71	79.58	70.96	0.33	0.02
14	MDA	73.75	56.07	85.84	73.02	78.91	70.96	0.29	0.01
15	NB	73.06	56.85	84.13	71.01	77.02	70.49	0.89	2.49
16	SDA	71.92	52.02	85.52	71.06	77.62	68.77	0.04	0.01
17	PDA	71.41	57.79	80.72	67.21	73.35	69.26	0.05	0.02
Average		75.36	62.97	83.83	72.68	77.75	73.4	1.7	0.25

* The training time is expressed in seconds and the test time in milliseconds.

Table 5. Supervised classification methods parameter configuration.

N°	Method name	Parameters name	Parameter values	Config.
1	AB	(mfinal, maxdepth, coeflearn)	(150, 3, 3)	27
2	BLR	(nIter)	(21)	3
3	C4.5 Trees	(C)	(0.25)	1
4	C5.0	(trials, model, winnow)	(10, 2, FALSE)	12
5	KNN	(kmax, distance, kernel)	(9, 2, 1)	3
6	MDA	(subclasses)	(2)	3
7	ANN	(size, decay, bag)	(5, 0.1, FALSE)	9

Nº	Method name	Parameters name	Parameter values	Config.
8	MARS	(degree)	(1)	1
9	NB	(fL, usekernel)	(0, TRUE)	2
10	NSC	(threshold)	(0.6496)	3
11	NN	(size, decay)	(5, 0.1)	9
12	PDA	(lambda)	(2)	3
13	RF	(mtry)	(2)	3
14	RBC	(threshold, pruned)	(0.25, 1)	1
15	SDA	(diagonal, lambda)	(FALSE, 1)	3
16	SRC	(-)	(-)	1
17	SGB	(n.trees, interaction.depth, shrinkage)	(150, 3, 0.1)	9
18	SVM	(C)	(0.25)	3

5 Conclusions and Future Work

In this work we show how client decisions are influenced by the emotions produced from the gas station characteristics.

We show that an ML approach can obtain the probability of selecting a gas station over another based on the emotions it produces in customers and it will remain valid if the population grows.

We would like to stress the importance of analytical decision-making. In this study it is found that ML approaches can improve the economical profit of investments made in Mexico in generating marketing strategies that make the clients prefer a national gas station over competency, emphasizing the gas station characteristics in adequate proportions according to the preferences of clients and their emotions.

References

1. Ratner, C.: A cultural-psychological analysis of emotions. *Culture & Psychology*, 6(1), pp. 5–39 (2000)
2. Frijda, N. H., Manstead, A. S., Bem, S.: *Emotions and beliefs: How feelings influence thoughts*. Cambridge University Press (2000)
3. Cambria, E., Melfi, G.: Semantic Outlier Detection for Affective Common-Sense Reasoning and Concept-Level Sentiment Analysis. In: *FLAIRS Conference*, pp. 276–281 (2015)
4. Turrentine, T., Kurani, K. S., Heffner, R. R.: *Fuel economy: what drives consumer choice?* Institute of Transportation Studies (2008)
5. Singh, S., Vrontis, D., Thrassou, A.: Green marketing and consumer behavior: The case of gasoline products. *Journal of Transnational Management*, 16(2), pp. 84–106 (2011)
6. Hansen, F., Christensen, S. R., Lundsteen, S.: Measuring emotions in a marketing context. *Innovative Marketing*, 2 (2), pp. 68–75 (2006)
7. Damasio, A. R.: *En busca de Spinoza: neurobiología de la emoción y los sentimientos*. Ed. Crítica, S.L. (2005)
8. Goleman, D.: *Inteligencia emocional*. Editorial Kairós (2012)

9. Ahn, H., Picard, R. W.: Affective-cognitive learning and decision making: A motivational reward framework for affective agents. In: *International Conference on Affective Computing and Intelligent Interaction*, Springer Berlin Heidelberg, pp. 866–873 (2005)
10. Ferrer, A.: *Neuromarketing, la tangibilización de las emociones*. Universidad Abat Oliba CEU, Francia (2009)
11. Minsky, M.: *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster (2007)
12. Ponce, H., Martínez-Villaseñor, L., Miralles-Pechuán, L.: A Novel Wearable Sensor-Based Human Activity Recognition Approach Using Artificial Hydrocarbon Networks. *Sensors*, 16(7), pp. 1033 (2016)
13. Aurier, P., Guintcheva, G.: The Dynamics of Emotions in Movie Consumption: A Spectator-Centred Approach. *International Journal of Arts Management*, 17(2), pp. 5 (2015)