Predicting Profitability of Neighbourhood Stores in Colombia

Review of Integrative Business & Economics Research

David Goyeneche
School of Business, Bond University, Gold Coast, Queensland, Australia

ABSTRACT

This paper examines the neighbourhood stores as part of the Colombian micro businesses landscape. Decision models were constructed to identify the profitability of neighbourhood stores, with the use of statistical exploration methods of analysis such as t-Test, Logistic Regression, Discriminatory Analysis, Decision Tree, Random Forest and Artificial Neural Networks. The database used for this study is the result of a survey carried out to the shopkeepers in 336 different neighbourhood stores in the city of Bucaramanga, Colombia, in 2014. The main finding from this research is that statistical analysis methods can offer significant accuracy when analysing micro business such as neighbourhood stores and should be considered in future research and in practice. The results indicate that prediction models can be a very useful decision-making tool for neighbourhood store owners, managers, suppliers, and investors. Therefore, practitioners and policy makers will also be able to make comparisons of past and present profitability, as well as future predictions. This will allow them to have a consistent perspective in the development of projects related to neighbourhood stores as part of the small business sector.

Keywords: Neighbourhood store; t-Test; Logistic regression; Discriminatory analysis; Decision tree; Random forest; Artificial neural networks.

Received 18 February 2021 | Revised 3 May 2021 | Accepted 12 June 2021.

1. INTRODUCTION

In Colombia, the first national and international supermarket chains started to stand out in the 1990s (Silva, 2011). Like many other countries, before the 1990s, Colombian consumers had mostly relied on mini-markets and neighbourhood stores to make their daily purchases (Elias, 2008). However, the entry and growth of supermarket chains were perceived as a great threat to neighbourhood stores. It was assumed local stores could not compete with the economies of scale, the variety of products and brands and the low prices they offer (Arboleda, 2001; Giraldo, 2009). Faced with this change, it was predicted that neighbourhood stores would disappear at the beginning of the year 2000 (Revista Dinero, 2003). Despite the predictions and disadvantages of neighbourhood stores compared to large retail companies, they continue to exist today.

However, neighbourhood stores today are still be faced with many threats that can affect their future if they are not prepared (Kraiwanit, 2017; Widaningrum, 2015). On a national scale, new retail investors and large supermarkets have chosen to develop their brands with their own small supermarkets or express supermarkets, located in strategic places in the middle of neighbourhoods, with areas not bigger than 150 m^2 (Cardona & Garcia,

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

2017; Gaitan, 2010; Guerra, 2012; Ramirez, 2013). On an international scale, well-known retail brands such as OXXO, which have taken a great portion of the convenience stores market in Mexico, are now growing in Colombia too (Vargas, 2016), and business models such as Amazon Go are beginning to take part not only in the convenience stores market, but also in grocery stores, and fast food outlets throughout the world (Bullard, 2016; Ives et al., 2019; Polacco, 2018). Additionally, the internet has become a competitor, and technology plays a high impact role in the fight for sales (Jin & Ji, 2018; Poon & Swatman, 1999).

Therefore, through statistical analysis, this paper seeks to construct models to predict to what extent a neighbourhood store is profitable or not based on available data. The results of this paper are highly relevant to clearly understand the seldomly explored profile of Colombian neighbourhood stores.

For neighbourhood store owners, practitioners and policy makers, these statistical models can help them to recognise the current situation of this business sector (Bullard, 2016). From the results, it is possible to identify a profitability position of the neighbourhood stores and thus be able to define strategies for their future plans (Craig & Moores, 2010; Kaplan & Norton, 2004). Also, the results could be used to support the creation of new stores, as well as for the search of funding from investors (Gepp et al., 2010; Mcgurr & Devaney, 1998). Additionally, having an instrument to predict business failure has the potential to benefit not only small business such as neighbourhood stores, but the whole economy of any country (Khaled, 2019).

2. RESEARCH QUESTIONS AND OBJECTIVES

The focus of this research is the creation of academic material that can facilitate the understanding of neighbourhood stores. Although the data provided comes from the city of Bucaramanga, its results can be considered similar to those that would be obtained from neighbourhood stores in other cities of Colombia (Acosta, 2014; Cordoba & Cano, 2009; Uribe, 2015).

2.1 Research questions

The questions that are intended to be answered are:
RQ1: What are the most relevant variables causing the profitability of neighbourhood stores in Colombia?
RQ2: How is it possible to identify and predict when a neighbourhood store in Colombia is profitable or not for their owners?

To answer the above questions, the following objectives are proposed:
2.1.1 To identify the most relevant variables generating profitability of a neighbourhood stores in Colombia.
2.1.2 To construct a series of models with different statistical analysis techniques to identify and predict the profitability of neighbourhood stores in Colombia.

3. LITERATURE REVIEW

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

3.1 Description of Neighbourhood Stores in Colombia

In Colombia, neighbourhood stores are usually created by adapting part of a house, such as the garage or the living room, where shelves, fridges and showcases are installed (Ramirez, 2013). The goods are acquired by direct purchase in marketplaces, supermarkets or by visits from suppliers. Typically, more than one store can be found per block in residential areas (Arboleda, 2001). Their main competitors are other stores, mini-markets and supermarkets (Cardona & Garcia, 2017). Among the products with the highest turnover are bread, milk, eggs, rice, produce, snacks and soft drinks. Their main customers are neighbours, people who pass by and people who work or study nearby (Giraldo, 2009). They have commercial areas below 50 m^2 (Gaitan, 2010; Goyeneche, 2014). Unlike supermarkets, neighbourhood stores in Colombia are preferred for their proximity as well as for being a place of socialization for neighbours (Cordoba & Cano, 2009). Additionally, shopkeepers establish friendship ties with their customers and also provide other services such as small-scale loans and credit payments for products (Coen et al., 2008).

3.2 Data description

The data selected for the current study is the result of an investigation undertaken at 336 neighbourhood stores in the city of Bucaramanga, in 2014 (Goyeneche, 2014). A total of 24 variables are considered for the present study (See Appendix A). To put the data of this research into a cultural context, it is necessary to consider that Colombia has a socioeconomic distribution called stratums (Martinsson et al., 2015). The stratum in Colombia represents a division of areas in the cities and towns, which run from level 1 to 6, as can be seen in Table 1 (Arrieta, 2018).

Table 1: Socioeconomic distribution of the Colombian population by stratum

Stratum	Socioeconomical Description
1	Low-low class
2	Low class
3	Low-middle class
4	Middle class
5	Middle-high class
6	High class

The majority of Colombians live in stratum 1, 2 or 3 homes, which represent about 80% of the housing in Colombia. The wealthy stratum 6 represents only about 3.5% of the housing in Colombia. In 2014, Bucaramanga had approximately 527,451 inhabitants. Its population by stratum (see Table 2; Acosta, 2014). A map of Bucaramanga with the distribution by stratum can be seen in Appendix B.

Table 2: Socioeconomic distribution of the Bucaramanga population by stratum in 2014

Stratum	Percentage	Population
1	11.07%	58,389
2	16.57%	87,399
3	28.89%	152,381
4	32.57%	171,791

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

5	3.81%	20,096
6	7.09%	37,396
	Total	**527,451**

Based on the current data, stratum areas 5 and 6 do not have a high impact on the consumption of neighbourhood stores (Gaitan, 2010; Giraldo, 2009; Goyeneche, 2014). Generally, the inhabitants of areas stratum 5 and 6 make their purchases in supermarkets (Cordoba & Cano, 2009; Silva, 2011). Therefore, to reduce the bias of the information in the present study, the 10 subjects from stratum 5 and 6 are discarded from the database as can be seen in Table 3.

Table 3: Selected sample of neighbourhood stores in Bucaramanga by stratum

Stratum	Count of Stratum	% of Stratum	Sum % of Stratum
1	48	14.29%	14.29%
2	44	13.10%	27.38%
3	139	41.37%	68.75%
4	95	28.27%	97.02%
5	9	2.68%	99.70%
6	1	0.30%	100.00%
	336		100.00%

3.3 Data analysis methods

Big data can be described in two ways: structured data and unstructured data. Structured data is highly specific and is stored in a predefined format, while unstructured data is a conglomeration of many varied types of data that are stored in their native formats. Different types of data can support different initiatives within retail (Bullard, 2016). There are two main data mining methods: unsupervised and supervised. The unsupervised method has no target variable and the algorithm therefore looks for patterns among all variables. Supervised models, conversely, have a pre-specified target and the algorithm finds the association between this target and the predictor variables. Furthermore, when statistics are used in the real world, it is not to simply apply tests to the data. The data and purpose of the tests must be thought through, because without context the numbers are just numbers and signify nothing (Takahashi & Inoue, 2016).

3.3.1 The independent t-Test

The independent t-Test is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups. The null hypothesis for the independent t-Test is that the population means from the two unrelated groups are equal: H_0: $u_1 = u_2$. In most cases, the purpose of the research is to see if it is possible to show that the null hypothesis can be rejected, and accept the alternative hypothesis, which is that the population means are not equal: H_A: $u_1 \neq u_2$. To do this, it is necessary to set a significance level (also called alpha) that can either reject or accept the alternative hypothesis. Most commonly, this value is set at 0.05 (Landau & Everitt, 2004).

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

To run an independent t-Test, the following conditions are required: One independent, categorical variable that has two levels; and at least one continuous dependent variable. Unrelated groups, also called unpaired groups or independent groups, are groups in which the cases (e.g., participants) in each group are different. Often, research is about differences in individuals, which means that when comparing two groups, an individual in one group cannot also be a member of the other group and vice versa. An example would be sex, where an individual would have to be classified as either male or female, but not both (Pallant, 2001).

3.3.2 Logistic Regression (LR)

Logistic regression analysis is a method for predicting probability, such as the probability of selling a particular cake based on a certain day of the week (Takahashi & Inoue, 2016). LR allows researchers to assess how well a group of variables predicts the categorical variables in a data set. It provides an indication of the relative importance of each predictor variable or the interaction among a predictor variable (Tabachnick, 2019). It is used to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

When selecting the model for the LR analysis, another important consideration is the model fit. Adding independent variables to a logistic regression model will always increase the amount of variance explained in the log odds (typically expressed as R^2). However, adding more and more variables to the model can result in overfitting, which reduces the generalizability of the model beyond the data on which the model is fit (Bender et al., 2007).

3.3.3 Decision Tree (DT)

A decision tree is a specific type of flow chart used to visualize the decision-making process by mapping out different courses of action, as well as their potential outcomes. The main characteristic of DTs is a recursive subsetting of a target field of data according to the values of associated input fields or predictors to create partitions, and associated descendent data subsets (called leaves or nodes), that contain progressively similar intra-leaf (or intra-node) target values and progressively dissimilar inter-leaf (or inter-node values) at any given level of the tree (de Ville, 2013). DTs are tools for choosing between several courses of action. They provide a highly effective structure within which it is possible to lay out options and investigate the possible outcomes of choosing different routes. They also help to form a balanced picture of the risks and rewards associated with each possible course of action. They can be simplified in terms of the clarity and accuracy of its final product. Ideally, the pruned DT should be more comprehensible than the original DT but should not be significantly less accurate when classifying unseen cases (Quinlan, 1987).

A significant advantage of a decision tree is that it forces the consideration of all possible outcomes of a decision and traces each path to a conclusion. It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis. Another advantage of DTs is that they are non-parametric. The only assumptions that DTs possess are that the successful and failing groups are discrete, non-overlapping and distinctly identifiable (Gepp et al., 2010). Evidence suggests that DT techniques demonstrate strengths as classifiers and predictors of business failure in

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

various circumstances (de Ville, 2013; Gepp et al., 2010). However, when using DTs, it must be considered that a small change in the data can cause a large change in the structure of the tree causing instability.

3.3.4 Random Forest (RF)

The Random Forest is a classification algorithm consisting of many DTs, which can be used for both classifications and regression tasks. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. In classification, the output is the mode of the classifications of the individual trees. In regression, the output is the mean from every generated tree (Khaled, 2019). RF will handle the missing values and maintain the accuracy of a large proportion of data. If there are more trees, RF would not allow overfitting trees in the model.

RF are among the most popular machine learning methods because of their relatively good accuracy, robustness and ease of use. They also provide two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy. RF is effective at classification but not for regression problems as it does not give precise continuous nature prediction. For regression, it does not predict beyond the range in the training data, and RF may over fit data sets that are particularly noisy.

3.3.5 Discriminant Analysis (DA)

Discriminant Analysis is a statistical technique used to classify observations into non-overlapping groups, based on scores on one or more quantitative predictor variables. It is an appropriate technique when the independent variable is not continuous (Coakes et al., 2008).

In many ways, discriminant analysis parallels multiple regression analysis. The main difference between these two techniques is that regression analysis deals with a continuous dependent variable, while discriminant analysis must have a discrete dependent variable. The methodology used to complete a discriminant analysis is similar to regression analysis. Each independent variable is plotted versus the group variable. Often, the next step is to go through a variable selection phase to determine which independent variables are beneficial. A residual analysis must then be conducted to determine the accuracy of the discriminant equations.

DA is most often used to help to predict the group or category to which a subject belongs. In terms of an equation, the expected relationship between the response variable and the independent variables can be explained by the equation $d=v1*X1+v2*X2+\ldots+vn*Xn+c$, where d is the discriminate function, v represents the discriminant coefficients, X is the respondent's score for that variable, and c is the constant (error). $n-1$ discriminant equations are obtained, where n is the number of groups that the dependent variable has.

3.3.6 Artificial Neural Network (ANN)

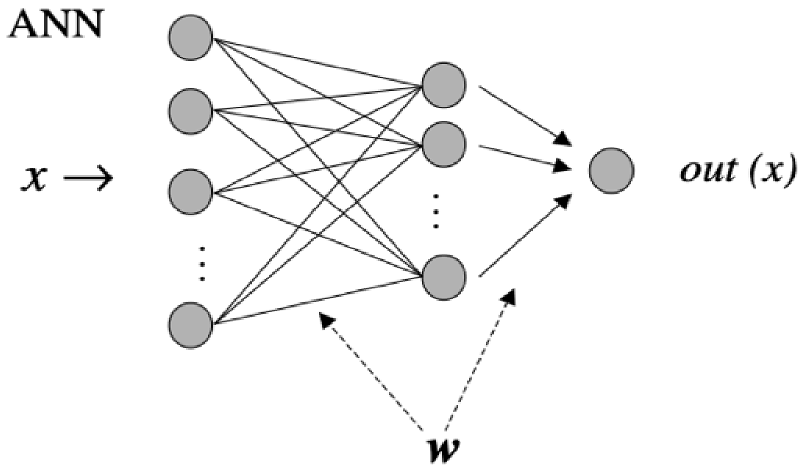
An Artificial Neural Network is the piece of a computing system designed to simulate the way the human brain analyses and processes information. It is the foundation of artificial intelligence (AI) and solves problems that would prove impossible or difficult by human

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

or statistical standards. The development of artificial neural networks arose from the attempt to simulate biological nervous systems by combining many simple computing elements (neurons) into a highly interconnected system and hoping that complex phenomena such as “intelligence” would emerge as the result of self-organization or learning (Sarle, 1994). The connections between these nodes have associated degrees of relative effect, which can be adjusted to match the actual relationships among the data (Griffith, 2010).

Neural networks are widely used in many fields of study. This could be attributed to the fact that these networks are attempts to model the capabilities of human brains (Paliwal & Kumar, 2009). The task of learning from data has been the object of two independent but converging traditions: machine learning, which emphasizes algorithmic approaches; and statistical modelling, which emphasizes the choice of a model for the probability distribution of the observed data (Ciampi & Lechevallier, 2007). Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted connections. The hidden layers then link to an 'output layer' where the answer is output, as shown in Figure 1.

Figure 1: ANN visualisation form



A specific issue of ANN is that in a sense they are ultimate 'black boxes'. The final product of this analysis process is a trained network that provides no equations or coefficients defining a relationship (as in regression). However, ANNs provide an analytical alternative to conventional techniques which are often limited by strict assumptions of normality, linearity, variable independence, and other cases. Because an ANN can capture many kinds of relationships, phenomena can be modelled, which otherwise may have been very difficult to explain.

The goal of creating artificial intelligence has led to some fundamental differences in philosophy between neural engineers and statisticians (Sarle, 1994). The major differentiation between machine learning and statistics is their purpose. Machine learning models are designed to make the most accurate predictions possible. On the other hand, statistical models are designed for inference about the relationships between variables. Nevertheless, when there is the possibility of cooperation, there can usually be cooperation, since independent decisions can reduce the rate of more realistic results. The

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

actions of researchers must be aimed at ensuring the solution to problems cooperatively without coming into conflict with nature (Gupta & Kumar, 2013).

4. RESEARCH METHODOLOGIES

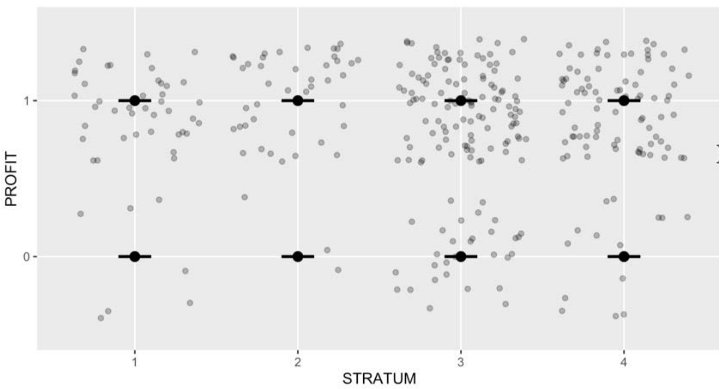
From the database of neighbourhood stores in the city of Bucaramanga in 2014, 24 variables were selected as the most relevant. The dichotomous variable, PROFIT, is identified with 1 for businesses that generate profit and 0 for those that do not generate profit. Although the initial data did not include the PROFIT variable, it was created as a rank among neighbourhood stores according to their sales per square meter, as shown in Table 4 (Bravo, 2019; Giraldo, 2009; Uribe, 2015).

Table 4: Profitability by area in neighbourhood stores according to stratum in Bucaramanga

STRATUM	ADSAREACS (X1.000 COP$/m^2)
1	>85
2	>110
3	>140
4	>160

These values have been generated based on the general characteristics of each establishment, the sales per square meter and the value of the legal minimum wage in Colombia for 2014. For example, stores in stratum 1 with sales over $ 85/m^2 per month will be considered profitable, otherwise non profitable. The same process was continued for stratum 2 to 4 as shown in Table 4 and Figure 2.

Figure 2: Visualisation of profitable and non-profitable neighbourhood stores by stratum 1 to 4 in Bucaramanga



4.1 Results and Interpretations

This analysis was undertaken by carrying out the different data analysis tests considered for the project:

4.1.1 The independent t-Test

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

With the purpose of answering to RQ1, the following hypotheses were raised:
The null hypothesis H_0 for the independent t-Test is that the population means from the two unrelated groups are equal. The opposite for alternative hypothesis H_A.
H_0: $u_1 = u_2$
H_A: $u_1 \neq u_2$

If p<0.05, H_0 is rejected and H_A accepted. This means that the variances are significantly different and cannot be assumed they are equal.
If p>0.05, H_0 is accepted. This means that variances are not significantly different and can be assumed they are equal.

Using the SPSS software, an independent sample t-Test was run on the data set of the neighbourhood stores with a 95% confidence interval (CI) for the mean difference (see Table 5).

The results showed the variability between profitable and non-profitable neighbourhood stores variables. On the one hand, according to the Levene's Test, there was statistically significant difference between the mean of 17 variables with p>0.05. On the other hand, independently, the test of equality of mean for MaAGE, CAGE41to50 and ADSAREACS showed a p<0.05. This means there is a significantly different value for the mean of the three variables.

Therefore, the H_o was rejected for the following three variables:
- MaAGE, t (324) = 2.529, p = 0.012
- CAGE41to50, t (80.7) = 2.427, p = 0.017
- ADSAREACS, t (289.6) = -13.047, p = 0.000

In conclusion, according to the t-Test results MaAGE, CAGE41to50 and ADSAREACS must be considered as variables of high relevance when identifying profitable and non-profitable neighbourhood stores.

4.1.2 Logistic Regression (LR)

A logistic regression analysis was performed to assess the effects of 12 factors on the likelihood that neighbourhood stores in Bucaramanga are profitable. As shown in Table 6, the logistic regression model was statistically significant, X^2 (12, N = 326) = 208.889, p < .001, indicating that the model was able to distinguish between profitable and non-profitable neighbourhood stores. The model as a whole explained between 47.3% (Cox and Snell R square) and 79.8% (Nagelkerke R squared) of the variance in profitability status, and correctly classified 93.9% of cases.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Table 5: t-Test for Equality of Means

		Levene's Test for Equality of Variances		t-Test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
STRATUM	Eva	4.767	0.03	0.663	324	0.508	0.099	0.149	-0.194	0.392
	Evna			0.709	80.838	0.481	0.099	0.139	-0.179	0.376
MaAGE	Eva	3.012	0.084	2.529	324	0.012	0.376	0.149	0.084	0.668
	Evna			2.668	79.718	0.009	0.376	0.141	0.096	0.656
EDUCATION	Eva	1.456	0.228	-0.021	324	0.984	-0.002	0.119	-0.237	0.233
	Evna			-0.023	87.774	0.981	-0.002	0.104	-0.21	0.205
DHOURS	Eva	0.006	0.94	-1.779	324	0.076	-0.60505	0.34018	-1.27429	0.06419
	Evna			-1.628	70.063	0.108	-0.60505	0.37166	-1.34628	0.13618
LEGAL	Eva	10.928	0.001	1.791	324	0.074	0.094	0.052	-0.009	0.196
	Evna			1.543	67.09	0.128	0.094	0.061	-0.027	0.215
YEARSCS	Eva	2.13	0.145	-0.728	324	0.467	-1.16325	1.59743	-4.3059	1.9794
	Evna			-0.767	79.614	0.445	-1.16325	1.51608	-4.18057	1.85406
CSTYPE	Eva	3.773	0.053	-0.932	324	0.352	-0.044	0.047	-0.136	0.048
	Evna			-1.064	87.52	0.29	-0.044	0.041	-0.125	0.038
SERVTYPE	Eva	0.35	0.554	-1.204	324	0.229	-0.175	0.145	-0.46	0.111
	Evna			-1.207	75.72	0.231	-0.175	0.145	-0.463	0.113
DIRPUR	Eva	0.33	0.566	-0.726	324	0.468	-0.03214	0.04428	-0.11926	0.05497
	Evna			-0.697	72.841	0.488	-0.03214	0.04612	-0.12406	0.05977
VISPUR	Eva	0.33	0.566	0.726	324	0.468	0.03214	0.04428	-0.05497	0.11926
	Evna			0.697	72.841	0.488	0.03214	0.04612	-0.05977	0.12406
MACUS	Eva	2.659	0.104	-0.257	324	0.797	-0.00575	0.02236	-0.04974	0.03825
	Evna			-0.284	83.93	0.777	-0.00575	0.02025	-0.04602	0.03453
FECUS	Eva	2.659	0.104	0.257	324	0.797	0.00575	0.02236	-0.03825	0.04974
	Evna			0.284	83.93	0.777	0.00575	0.02025	-0.03453	0.04602
CAGE15orLESS	Eva	7.185	0.008	-1.061	324	0.289	-0.01139	0.01073	-0.03249	0.00972
	Evna			-1.161	83.015	0.249	-0.01139	0.00981	-0.03089	0.00812
CAGE15to20	Eva	5.819	0.016	-1.234	324	0.218	-0.01308	0.0106	-0.03394	0.00777
	Evna			-1.443	90.313	0.152	-0.01308	0.00906	-0.03109	0.00492
CAGE21to30	Eva	0.872	0.351	-0.194	324	0.846	-0.00234	0.01207	-0.02608	0.0214
	Evna			-0.213	83.305	0.832	-0.00234	0.011	-0.02422	0.01953
CAGE31to40	Eva	8.328	0.004	-1.101	324	0.272	-0.01521	0.01382	-0.04239	0.01197
	Evna			-1.496	115.712	0.137	-0.01521	0.01017	-0.03535	0.00493
CAGE41to50	Eva	9.091	0.003	2.274	324	0.024	0.02523	0.0111	0.0034	0.04706
	Evna			2.427	80.727	0.017	0.02523	0.0104	0.00454	0.04592
CAGE51to60	Eva	0.941	0.333	1.192	324	0.234	0.00982	0.00824	-0.00639	0.02603
	Evna			1.348	86.458	0.181	0.00982	0.00728	-0.00466	0.0243
CAGE61orOLDER	Eva	2.12	0.146	1.15	324	0.251	0.00697	0.00606	-0.00495	0.0189
	Evna			0.996	67.326	0.323	0.00697	0.007	-0.00699	0.02094
SERVFIN	Eva	0.121	0.729	-0.169	324	0.866	-0.012	0.072	-0.153	0.129
	Evna			-0.169	75.55	0.866	-0.012	0.072	-0.155	0.131
ADSAREACS	Eva	21.037	0	-5.91	324	0	-265.86113	44.98699	-354.36462	-177.35764
	Evna			-13.047	289.654	0	-265.86113	20.37768	-305.96823	-225.75403

*Equal variances assumed (Eva)

**Equal variances not assumed (Evna)

As shown in Table 7, only four of the independent variables made a unique statistically significant contribution to the model (EDUCATION, DHOURS, ADSAREACS and CAGE21to30). High levels of schooling (EDUCATION) showed association with reduction in profit. In contrast, having the establishment open for more hours (DHOURS) and generating more sales per square meter (ADSAREACS) was associated with an increase in profit. Therefore, the derived estimated equation model for profitability prediction is:

Z = -5.129 - 11.695 (CAGE21to30) - 1.108 (EDUCATION) + 0.363 (DHOURS) + 0.074 (ADSAREACS)

As can be seen in Table 8, the model showed that there were 263 true positives out of the 272 cases reported. The positive predictive value was 96.7%, indicating that of the

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

neighbourhood stores predicted to be profitable the model accurately selected 96.7% of them. The true negatives were 43 of 54. The negative predicted value is 79.6%, indicating that of the neighbourhood stores predicted to be non-profitable the model accurately selected 79.6% of them. The overall predictability accuracy of the logistic regression model was 93.9%.

Through the above model, it is therefore possible to predict when a neighbourhood store is profitable in Colombia. The information required to develop the formula can be defined by the owners and / or managers of the neighbourhood stores. The probability Z can be determined by replacing the values EDUCATION, DHOURS, ADSAREACS and CAGE21to30. If this probability is higher than the cut value 0.5, there is a 95% of confidence than the neighbourhood will be successful. With $Z<0.5$, there is a 95% of confidence that the neighbourhood store is not profitable.

Table 6: Model Summary LR

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	83.802a	0.473	0.798
a Estimation terminated at iteration number 11 because parameter estimates changed by less than .001.			

Table 7: Variables in the Equation LR

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1a	MaAGE	-0.283	0.343	0.681	1	0.409	0.753	0.384	1.476
	EDUCATION	-1.108	0.521	4.532	1	0.033	0.33	0.119	0.916
	DHOURS	0.363	0.177	4.221	1	0.04	1.438	1.017	2.033
	SERVFIN	1.198	0.623	3.699	1	0.054	3.315	0.977	11.245
	ADSAREACS	0.074	0.013	30.702	1	0	1.077	1.049	1.105
	DIRPUR	-1.05	1.153	0.828	1	0.363	0.35	0.037	3.357
	CAGE21to30	-11.695	5.796	4.071	1	0.044	0	0	0.716
	CAGE31to40	-7.636	5.803	1.732	1	0.188	0	0	41.973
	CAGE41to50	-4.468	4.561	0.96	1	0.327	0.011	0	87.393
	CAGE51to60	-3.977	6.979	0.325	1	0.569	0.019	0	16333.055
	CAGE61orOLDER	-12.955	9.53	1.848	1	0.174	0	0	305.652
	MACUS	1.059	2.339	0.205	1	0.651	2.883	0.029	282.272
	Constant	-5.129	4.08	1.58	1	0.209	0.006		
a Variable(s) entered on step 1: MaAGE, EDUCATION, DHOURS, SERVFIN, ADSAREACS, DIRPUR, CAGE21to30, CAGE31to40, CAGE41to50, CAGE51to60, CAGE61orOLDER, MACUS.									

Table 8: Training Classification Table LR

	Observed		Predicted		
			PROFIT		Percentage Correct
			0	1	
Step 1	PROFIT	0	43	11	79.6%
		1	9	263	96.7%
	Overall Percentage				93.9%
a The cut value is .500					

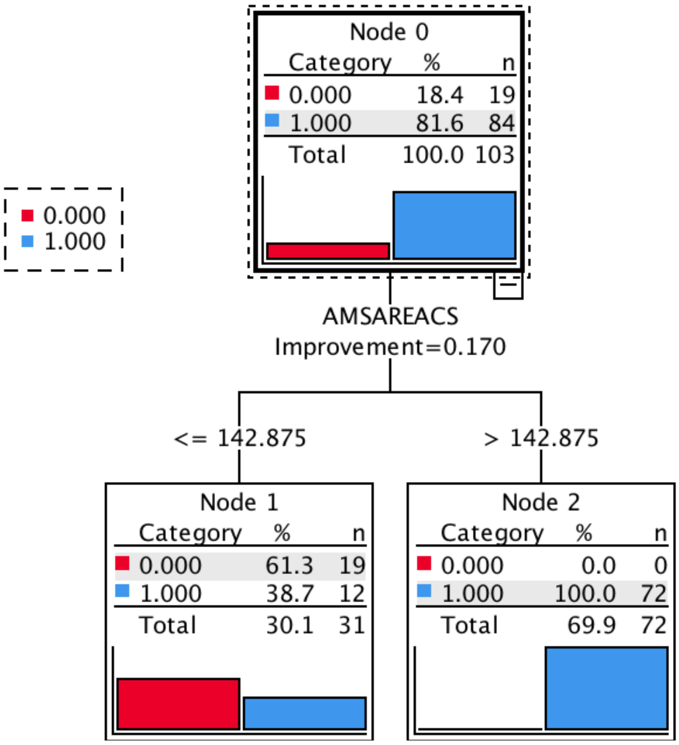
4.1.3 Decision Tree

Using SPSS software, decision trees were built using Classification and Regression Trees (CART) as the growing method on the data set. The ratio for training to testing was taken as 7:3. Twenty-one independent variables were specified for CART, although

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

ADSAREACS was the only variable with significant contribution to the model. The other variables were excluded from the first model.

Figure 3. CART Decision Tree



As can be observed in Figure 3 of the CART tree, ADSAREACS is the terminal node in the tree diagram for predicting when a neighbourhood store is profitable or not. ADSAREACS specifies that if the average monthly sales per square meter is lower than COP\$142.875/m^2, there is a 61.3% probability that the neighbourhood store is not making a profit and a 38.7% probability that the neighbourhood store is making a profit. If ADSAREACS is greater than COP\$142.875/m^2, there is a probability of 100% that the store is making a profit.

As shown in Table 9, the constructed decision tree was able to correctly classify 72 of the observed cases in the test sample of 84, and 12 cases incorrectly. The model presents approximately 100% of sensitivity and 85,7% of specificity when predicting if a neighbourhood store is profitable, with an overall correct classification of 88.30% achieved.

Table 9. Classification Table DT

Sample	Observed	Predicted		Percent Correct
		0	1	
Training	0	35	0	100.00%
	1	15	173	92.00%
	Overall Percentage	22.40%	77.60%	93.30%
Test	0	19	0	100.00%
	1	12	72	85.70%
	Overall Percentage	30.10%	69.90%	88.30%
Growing Method: CRT				
Dependent Variable: PROFIT				

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

4.1.4 Random Forest

Through the software Salford System, a RF was modelled with the neighbourhood store data set. The statistics model shown in Figure 4 correspond to the resultant RF, which in this case is a group of 500 trees. However, the lowest balanced error rate is 0.056 with 19 trees. RF generally have superior predictive performance versus CART trees because RF have lower variance than a single CART tree. Nevertheless, random forests are more likely to be effective when the model has many features where the importance across each is more balanced. In the case of the neighbourhood store data set, the importance of the variable ADSAREACS is far more important than the other variables. As a result, the overall outcomes of DT with 88.3% (see Table 9) had a higher performance than the RF which had an overall percentage of 79.41% (see Table 10).

Figure 4. Random Forest: Number of Trees

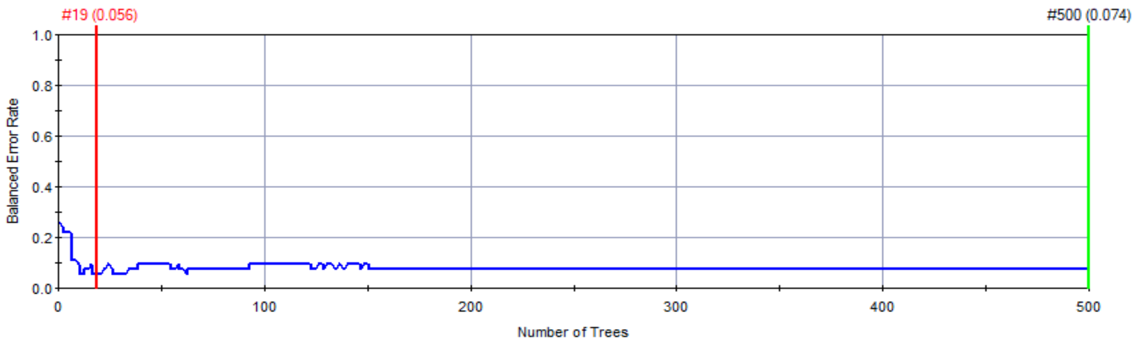


Table 10 shows that the optimal RF correctly classified 43 of the observed cases in the test sample of 57, and 14 cases incorrectly. The model presented approximately 100% of sensitivity and 75.44% of specificity when predicting if a neighbourhood store was profitable. An overall correct classification of 79.41% was achieved.

Table 10. Classification Table RF

Sample	Observed	Predicted		Percent Correct
		0	1	
Training	0	43	0	100.00%
	1	62	229	78.69%
	Overall Percentage	31.44%	68.56%	81.44%
Test	0	11	0	100.00%
	1	14	43	75.44%
	Overall Percentage	36.76%	63.24%	79.41%

4.1.5 Discriminant Analysis

In Test of Equality of Group Means, out of the 21 variables, MaAGE, CAGE41to50 and ADSAREACS made a significant contribution to the model (see Table 11). Of these, ADSAREACS was the most important independent variable to discriminate the function.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Table 11. Test of Equality of Group Means DA

	Wilks' Lambda	F	df1	df2	Sig.
MaAGE	0.981	6.396	1	324	0.012
EDUCATION	1	0	1	324	0.984
DHOURS	0.99	3.164	1	324	0.076
LEGAL	0.99	3.207	1	324	0.074
YEARSCS	0.998	0.53	1	324	0.467
CSTYPE	0.997	0.868	1	324	0.352
SERVTYPE	0.996	1.45	1	324	0.229
DIRPUR	0.998	0.527	1	324	0.468
VISPUR	0.998	0.527	1	324	0.468
MACUS	1	0.066	1	324	0.797
FECUS	1	0.066	1	324	0.797
CAGE15orLESS	0.997	1.127	1	324	0.289
CAGE15to20	0.995	1.523	1	324	0.218
CAGE21to30	1	0.038	1	324	0.846
CAGE31to40	0.996	1.212	1	324	0.272
CAGE41to50	0.984	5.17	1	324	0.024
CAGE51to60	0.996	1.42	1	324	0.234
CAGE61orOLDER	0.996	1.323	1	324	0.251
SERVFIN	1	0.028	1	324	0.866
ADSAREACS	0.903	34.925	1	324	0
STRATUM	0.999	0.44	1	324	0.508

The Eigenvalue is a measure of how much of the variance of the observed variables a factor explains. In table 12, Eigenvalue is .203, which means that 100% of the variance of the observed variables is explained by the factor. The canonical correlation associated to the function is 0.411. The square of this correlation indicates that 16.89% of the dependent variable (PROFIT) can be explained by group differences.

Table 12. Eigenvalues Table

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.203a	100	100	0.411
a First 1 canonical discriminant functions were used in the analysis.				

The Wilks' Lambda coefficient is defined as the proportion of the total variance in the discriminant scores not explained by differences among the groups, here 0.831>0.05 (See Table 13). This confirms that the functions are statistically significant.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Table 13. Wilks' Lambda DA

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	0.831	58.181	18	0

Table 14. Classification Function Coefficients DA

	PROFIT		
	0	1	Dif
MaAGE	2.356	1.974	0.382
EDUCATION	3.368	3.194	0.174
DHOURS	4.453	4.592	-0.139
LEGAL	25.742	24.923	0.819
YEARSCS	0.477	0.482	-0.005
CSTYPE	7.515	8.073	-0.558
SERVTYPE	3.175	3.366	-0.191
DIRPUR	-4.931	-4.682	-0.249
MACUS	14.954	14.568	0.386
CAGE15orLE	637.13	644.273	-7.143
CAGE15to20	743.866	750.653	-6.787
CAGE21to30	602.404	605.352	-2.948
CAGE31to40	643.572	649.948	-6.376
CAGE41to50	660.731	661.97	-1.239
CAGE51to60	762.763	766.085	-3.322
SERVFIN	6.174	6.597	-0.423
AMSAREACS	0	0.003	-0.003
STRATUM	7.894	7.625	0.269
(Constant)	-389.009	-392.088	3.079
Fisher's linear discriminant functions			

Through the results in Table 14 Classification Function Coefficients, it is possible to define Fisher's linear discriminant function for profitability of neighbourhood stores as follows:

F = 3.079 + 0.382(MaAGE) + 0.174(EDUCATION) - 0.139(DHOURS) + 0.819(LEGAL) - 0.005(YEARSCS) - 0.558(CSTYPE) - 0.191(SERVTYPE) - 0.249(DIRPUR) + 0.386(MACUS) - 7.143(CAGEorLESS) - 6.787(CAGE15to20) - 2.948(CAGE21to30) - 6.376(CAGE31to40) - 1.239(CAGE41to50) - 3.322(CAGE51to60) - 0.423(SERVFIN) - 0.003(ADSAREACS) + 0.269(STRATUM)

Table 15. Functions at Group Centroids

PROFIT	Function
	1
0	-1.008
1	0.2
Unstandardized canonical discriminant functions evaluated at group means	

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

The discriminant scores are centered so that they have a sample mean zero. These scores can be compared with the average of their group means in Table 15 Functions at Group Centroids, to allocate the neighbourhood store into the correspondent group of profitable or non-profitable. Here the threshold against which a store discriminant score is evaluated is -0.404 = 1/2 x (-1.008 + 0.2). Therefore, the neighbourhood stores with discriminant scores above -0.404 are likely to be profitable; otherwise, they would be classified as non-profitable.

Table 16. Classification results a,c

		PROFIT	Predicted Group 0	Predicted Group 1	Total
Original	Count	0	44	10	54
		1	78	194	272
	%	0	81.5	18.5	100
		1	28.7	71.3	100
Cross-validated b	Count	0	38	16	54
		1	87	185	272
	%	0	70.4	29.6	100
		1	32	68	100
a 73.0% of original grouped cases correctly classified.					
b Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.					
c 68.4% of cross-validated grouped cases correctly classified.					

According to table 16, the linear discriminant function reported that in 73% of the cases neighbourhood stores can be correctly classified as profitable or non-profitable. Cross validation drops to 68.4%, which is a considerable low success rate.

4.1.6 Artificial Neural Network (ANN)

Artificial Neural Network analysis was performed on the neighbourhood stores data set using the Multilayer Perceptron in order to produce a predictive model based on the values of the independent variables. The case processing summary in Table 17, shows that 231 cases were assigned to the training sample and 42 to the testing sample.

Table 17. Case Processing Summary ANN

		N	Percent
Sample	Training	231	84.60%
	Testing	42	15.40%
Valid		273	100.00%
Excluded		53	
Total		326	

As shown in Table 18, 182 of 186 are classified correctly in the training sample, corresponding to 97.8% of the training cases. Overall, this is considered a good model as it classifies more than 85% of the cases correctly.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Table 18. Classification Table ANN

	Observed	Predicted		
Sample		0	1	Percent Correct
Training	0	3	42	6.70%
	1	4	182	97.80%
	Overall Percent	3.00%	97.00%	80.10%
Testing	0	0	6	0.00%
	1	0	36	100.00%
	Overall Percent	0.00%	100.00%	85.70%
Dependent Variable: PROFIT				

According to the Receiver Operating Characteristic Curve (ROC), plotting the true positive rate against the false positive rate, represents a visual display of sensibility and specificity for all possible cut-offs. The values in Table 19 represent the probability that the predicted model was correct. In the case a profitable neighbourhood store and another non-profitable are randomly selected, there is 72.6% probability of success in the prediction. The area under the curve gives an idea about the benefit of the test (See Figure 5).

Figure 5. Receiver Operating Characteristic Curve (ROC) ANN

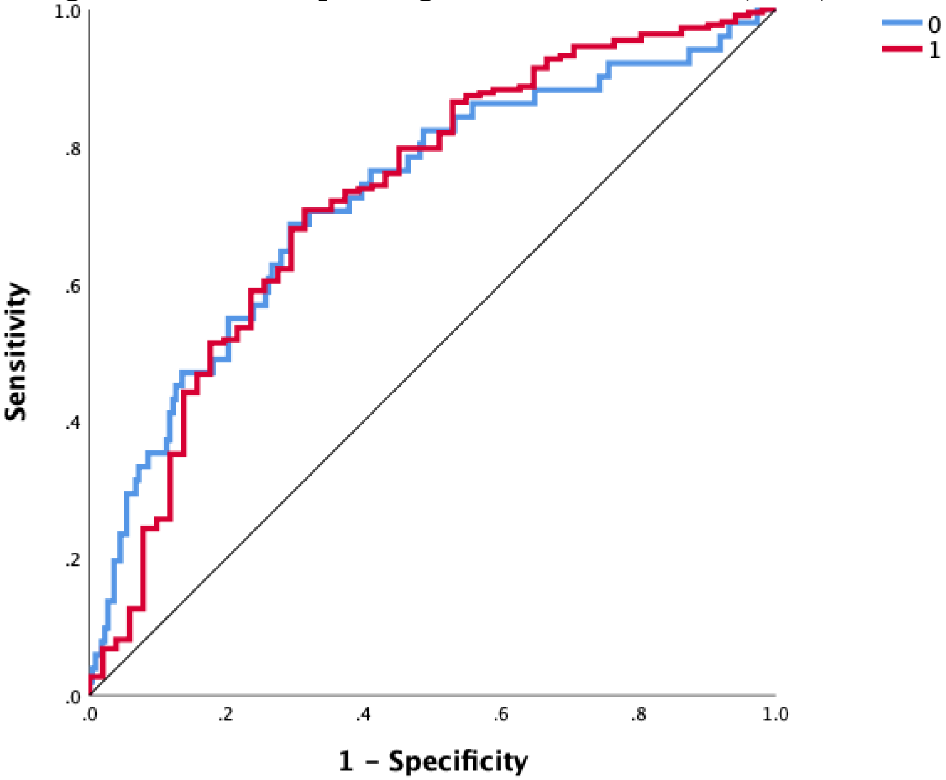


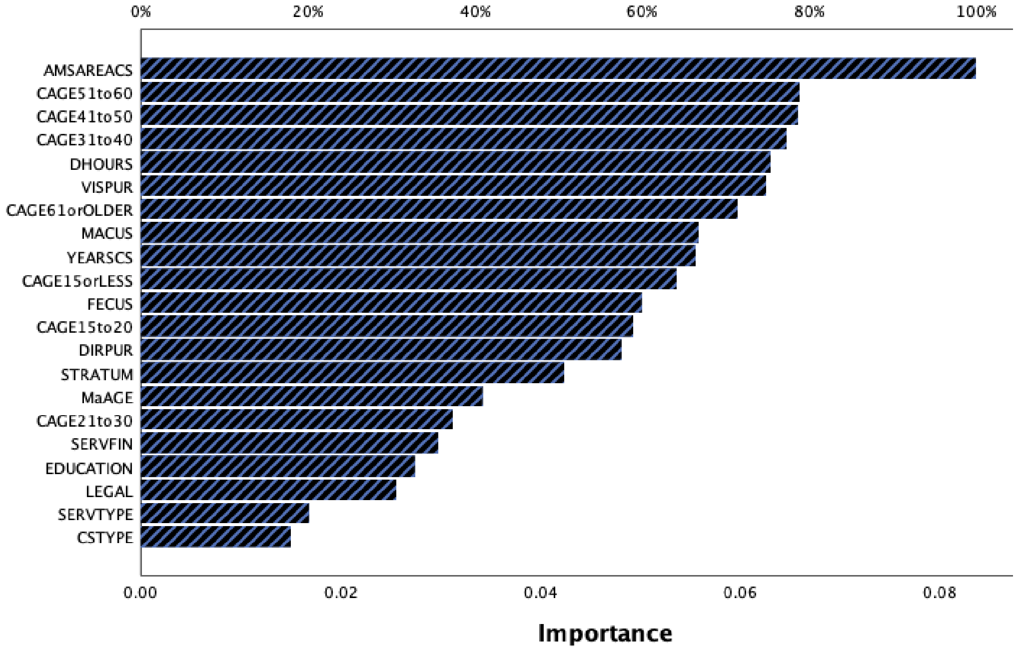
Table 19. Area under the curve ANN

		Area
PROFIT	0	0.726
	1	0.726

The most important variables in predicting the profitability of neighbourhood stores as per normalised importance chart, were the average daily sales per square meter, followed by the age of the customers from 31 to 60 years old (See Figure 6).

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Figure 6. Normalized importance of variables ANN



5. DISCUSSION AND CONCLUSIONS

Research in the field of retail, in the area of neighbourhood stores, has had very little relevance in the academy, especially in developing countries. This is partly because of the difficulty in data collection and because neighbourhood stores are also part of a kind a business model that has been disappearing in recent decades around the world. In the current research, through the analysis of 21 variables related to the normal operation of neighbourhood stores in Colombia, it was possible to propose parameters to define when a neighbourhood store is profitable or not. Although the data was obtained in the city of Bucaramanga, these results could have applicability in other main cities of Colombia. This is due to the nature of the data, obtained in different stratums, which are similar throughout the country.

As per comparison of all the techniques used in this research, it was found that the most successful prediction model was Logistic Regression. In Table 20, the comparison between the prediction models is showed.

Although LR outperformed all other models as per overall model accuracy, other particular characteristics can be rescued from the information set. It can be seen in Table 20 that all the models confirm the sales per square meter (ADSAREACS) as the most relevant variable. This suggests that the best use of space is vital for neighbourhood stores, in other words, more products in the same area can provide greater opportunities to achieve greater profit. MaAGE and CAGEs (Customer Age) are also presented as relevant variables in other models. According to LR's results, younger clients negatively affect the equation for profitability. Consistently, t-Test and DA and ANN, consider mature clients (Older tan 31 years) as important variables for the achievement of profit, in addition to the age of managers variable (MaAGE) in the neighbourhood stores. Thus,

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

this indicates that neighbourhood stores with mature customers and mature managers have a greater chance of being profitable.

Table 20. Prediction Models Comparison

Model	Most Important Variables	Overall Model Accuracy
Logistic Regression	ADSAREACS, CAGE21to30, EDUCATION, DHOURS	Sensitivity: 79.6% Specificity: 96.7% Average: 93.9%
Decision Tree	ADSAREACS	Sensitivity: 100% Specificity: 85.7% Average: 88.3%
Random Forest	ADSAREACS	Sensitivity: 100% Specificity: 75.44% Average: 79.41%
Discriminant Analysis (Cross-Validated)	ADSAREACS CAGE41to50 MaAGE	Sensitivity: 5.56% Specificity: 98.16% Average: 82.82%
Artificial Neural Network	ADSAREACS, CAGE51to60, CAGE41to50, CAGE31to40	Sensitivity: 0% Specificity: 100% Average: 85.7%

The variable EDUCATION is mentioned in LR, and although it might be thought that the higher the academic level, the greater the profit, the opposite happens. This is because shopkeepers have a low level of academic preparation, this has been replaced with their work experience at managing their neighbourhood stores. This suggests that their preparation could be strengthened with training related to the professionalization of their commercial activity. It should also be noted that customers of neighbourhood stores have similar characteristics of school preparation than shopkeepers, which leads to have similarities that strengthen ties with the needs and desires in their local microcosm.

5.1 Future Work

Research in the business field has generally been focused on large companies. However, there are countless opportunities to explore small and micro businesses. Despite being small from the individual point of view, when considered as a whole, the micro business presents a wealth of information relevant to the economy of any country, especially those in a developing stage. In the case of neighbourhood stores, there is an opportunity to obtain more information relevant to financial, administrative, and marketing variables that can generate more informative data. With more information, more accurate predictions could be achieved in future studies. Integration of statistical models could also be achieved and compared with the results of the present study. For example, for future research in the field of neighbourhood stores, the number of product references and the planimetric display area could be considered as very informative variables. Likewise, financial, inventory and cost indicators, among others, could generate a greater opportunity in the search for knowledge.

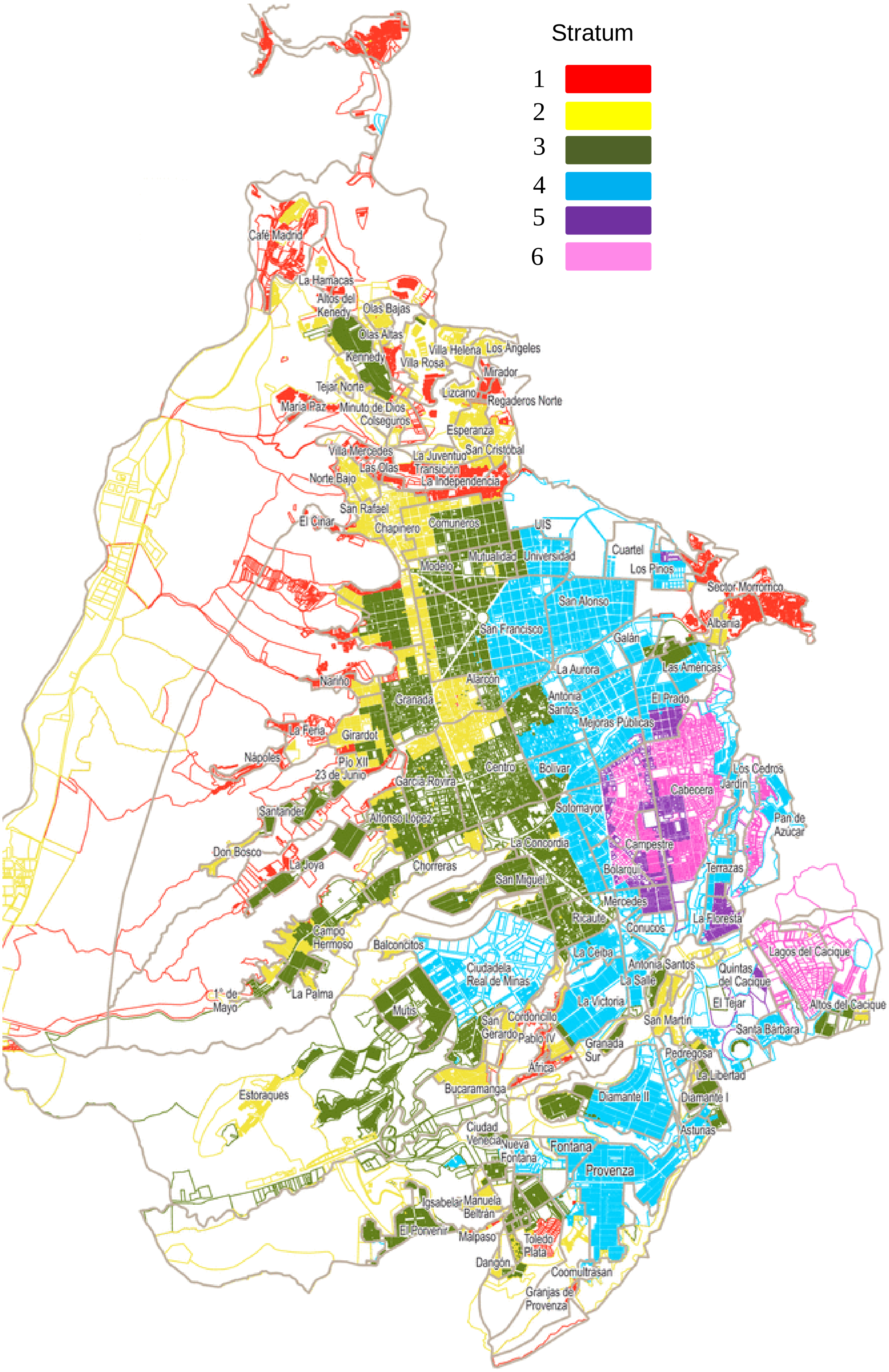
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Appendix A – Variable definitions

ADS	Average Daily Sales (ADS): This variable shows the average sales value of a neighbourhood store per month.
ADSAREACS	Average Daily Sales in Colombian pesos divided by the Area of the Neighbourhood Store ($/m^2): Sales per square meter is a metric commonly used by retail companies to determine the amount of revenue generated per square foot of retail space. Sales per square meter can be used to determine the sales efficiency of retail stores. To consider the real value of the Colombian pesos (COP$), the ADSAREACS value must be multiplied by 1,000. For instance, 123$/m2 x 1,000 = 123.000 $/m2, which means one hundred twenty-three thousand Colombian pesos per square meter.
AREACS	Area in square meters (m^2) of the commercial space of the neighbourhood store.
CAGE15orLESS	Percentage (%) of customers under the age of 16 who make purchases at the neighbourhood store.
CAGE16to20	Percentage (%) of customers between 16 and 20 years old, who make purchases in the neighbourhood store
CAGE21to30	Percentage (%) of customers between 21 and 30 years old, who make purchases in the neighbourhood store
CAGE31to40	Percentage (%) of customers between 31 and 40 years old, who make purchases in the neighbourhood store
CAGE41to50	Percentage (%) of customers between 41 and 50 years old, who make purchases in the neighbourhood store
CAGE51to60	Percentage (%) of customers between 51 and 60 years old, who make purchases in the neighbourhood store
CAGE61orOLDER	Percentage (%) of customers over 60 who make purchases at the neighbourhood store.
CSTYPE	This variable refers to whether the location of the neighbourhood store is in a commercial area close to other businesses or whether it is located in a residential area.
DHOURS	Average time of hours the neighbourhood store remains open each day.
DIRPUR	This variable shows the percentage (%) of products that are purchased by the store directly at selected purchase points.
EDUCATION	Level of education of the owner or administrator of the neighbourhood store. This classification goes from number 1 to 5, where the person in charge: 1 does not have any kind of studies, 2 has Primary School, 3 Secondary School, 4 Certificate and/or Technique School, and 5 University Degree.
FECUS	Female Customers: The percentage of female customers who shop at the neighbourhood store.
LEGAL	This variable denotes whether or not the neighbourhood stores are legally registered as a company with the national tax office and chamber of commerce. 1 = yes and 0 = no.
MaAGE	Age of the owner and/or administrator of the neighbourhood store. The range of years is identified from 21 to 30 = 1, 31 to 40 = 2, 41 to 50 = 3, 51 to 60 = 4 and 61 or older = 5.
MACUS	Male Customers: The percentage of male customers who shop at the neighbourhood store.
PROFIT	The degree to which the convenience store yields profit or financial gain. This indicator is denoted by 1 for neighbourhood stores that are considered to generate profit and 0 for those that are not.
SERFIN	Financing service: In this practice, the shopkeeper writes down in a paper notebook the products that the customer purchases to be paid later. In Colombian slang, this is known as “fiar”.
SERVTYPE	This variable refers to the type of service that the neighbourhood store provides. 1 denotes personalised service, where all the products are behind a fence or counter and the customer needs to indicate what he needs to be reached by the shopkeeper. 2 Self-service, where customers have the possibility of grabbing the products they are going to buy. 3 Mix, which represents that part of the products are behind the counter (especially the highest value) and others can be grabbed by customers.
STRATUM	In Colombia, houses, apartments, and other buildings are given a number on a scale of 1 to 6, with 6 being the highest on the socio-economic scale and 1 being the lowest. A stratum is based on housing conditions and infrastructure. For more information see Table 1.
VISPUR	This variable shows the percentage of products that are purchased by the store through supplier companies that visit them.
YEARSCS	Time in years that the neighbourhood store has existed.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Appendix B – Map of Bucaramanga divided by socioeconomic stratum (Uribe, 2015)



ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

REFERENCES

[1] Acosta, C. (2014). Aproximación a los hábitos de consumo de contenidos mediáticos por parte de los niños y niñas en Colombia. In *MinCultura Colombia* (Vol. 1, Issue 1). https://doi.org/10.16309/j.cnki.issn.1007-1776.2003.03.004

[2] Arboleda, M. (2001). *The neighborhood stores: The retail trade and its challenges in the market*. 1–10.

[3] Arrieta, R. (2018, November 7). Living in Bogota : how the strata system in Colombia works. *Colture*, 2018.

[4] Bender, R., Ziegler, A., & Lange, S. (2007). NCSS Multiple regression. In *Ncss, Llc* (pp. 1–63). https://doi.org/10.1055/s-2007-959036

[5] Bravo, S. (2019). *Experiencias de campo del proyecto de investigación para tenderos cuadernos del semillero de investigación "La Vitrina."* Fundación Universitaria del Área Andina.

[6] Bullard, B. (2016). *Style & Statistics: The art of retail analytics* (Vol. 21, Issue 1). Wiley.

[7] Cardona, J., & Garcia, W. (2017). *Impacto de ARA Y D1 sobre la dinámica de las tiendas de barrio en Villa Verde, Villa Del Prado y El Poblado de la ciudad de Pereira*. Universidad Libre de Colombia.

[8] Ciampi, A., & Lechevallier, Y. (2007). Statistical Models and Artificial Neural Networks: Supervised Classification and Prediction Via Soft Trees. *Advances in Statistical Methods for the Health Sciences*, 239–261. https://doi.org/10.1007/978-0-8176-4542-7_16

[9] Coakes, S. J., Steed, L., & Price, J. (2008). SPSS: Analysis Without Anguish; version 15.0 for Windows. In *SPSS for Windows* (p. 270).

[10] Coen, S. E., Ross, N. A., & Turner, S. (2008). "Without tiendas it's a dead neighbourhood": The socio-economic importance of small trade stores in Cochabamba, Bolivia. *Cities*, *25*(6), 327–339. https://doi.org/10.1016/j.cities.2008.06.003

[11] Cordoba, J., & Cano, L. (2009). Study of consumer habits and value in the neighborhood Stores channel in Cali. *Entramado*, *5*(2), 44–63.

[12] Craig, J., & Moores, K. (2010). Strategically aligning family and business systems using the Balanced Scorecard. *Journal of Family Business Strategy*, *1*(2), 78–87. https://doi.org/10.1016/j.jfbs.2010.04.003

[13] de Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448–455. https://doi.org/10.1002/wics.1278

[14] Elias, R. P. (2008). Why haven't small retail grocery shops failed in Colombia with the arrival of the big chains of hypermarkets? *Entornos*, *1*(21), 37–50.

[15] Gaitan, D. (2010). Caraterización de las tiendas de barrio en Cartagena. *Panorama*, *8*, 59–71.

[16] Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting*, *29*(6), 536–555. https://doi.org/10.1002/for.1153

[17] Giraldo, J. (2009). Impacto de la entrada de las grandes superficies de venta sobre las tiendas y micromercados de barrio de la ciudad de Bucaramanga y su área metropolitana. *Revista Colombiana de Marketing*, *8*(13), 7–27.

[18] Goyeneche, D. (2014). *Formulación de estrategias para el mejoramiento de la administración de las tiendas de barrio de Bucaramanga*. Saint Thomas Aquinas University.

[19] Griffith, A. (2010). *SPSS for Dummies* (2nd ed.). Wiley Publishing.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

[20] Guerra, H. S. (2012). Panorama del negocio minorista en Colombia. *Pensamiento & Gestión*, *32*, 115–141.
[21] Gupta, S. N., & Kumar, K. (2013). A game theoretic approach for optimum strategy in life. *Journal of Interdisciplinary Mathematics*, *16*(1), 49–58. https://doi.org/10.1080/09720502.2013.778509
[22] Ives, B., Cossick, K., & Adams, D. (2019). Amazon Go: Disrupting retail? *Journal of Information Technology Teaching Cases*, *9*(1), 2–12. https://doi.org/10.1177/2043886918819092
[23] Jin, Y., & Ji, S. (2018). Mapping hotspots and emerging trends of business model innovation under networking in Internet of Things. *Eurasip Journal on Wireless Communications and Networking*, *2018*(1). https://doi.org/10.1186/s13638-018-1115-4
[24] Kaplan, Robert; Norton, D. (2004). How strategy maps frame an organization's objectives. *Financial Executive*, *April*, 40–46.
[25] Khaled, H. (2019). *Topics on Financial Distress Prediction Modelling*. BOND University.
[26] Kraiwanit, T. (2017). Social Network Sanctions to against the Retail Business Monopolies. *Review of Integrative Business & Economics Research*, *6*(1), 217–228.
[27] Landau, S., & Everitt, B. (2004). A Handbook of Statistical Analyses Using SPSS. In CRC Press (Ed.), *The Statistician* (2nd Editio, Vol. 44, Issue 3). Chapman & Hall/CRC. https://doi.org/10.2307/2348714
[28] Martinsson, P., Villegas-Palacio, C., & Wollbrant, C. (2015). Cooperation and social classes: evidence from Colombia. *Social Choice and Welfare*, *45*(4), 829–848. https://doi.org/10.1007/s00355-015-0886-3
[29] Mcgurr, P. T., & Devaney, S. A. (1998). A retail failure prediction model. *International Journal of Phytoremediation*, *21*(1), 259–276. https://doi.org/10.1080/095939698342779
[30] Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, *36*(1), 2–17. https://doi.org/10.1016/j.eswa.2007.10.005
[31] Pallant, J. (2001). *SPSS Survival manual* (1st ed.). Open University Press, McGraw-Hill.
[32] Polacco, A. (2018). *JBM-2401-04-full*. *24*(March), 79–92. https://doi.org/10.6347/JBM.201803
[33] Poon, S., & Swatman, P. M. C. (1999). An exploratory study of small business Internet commerce issues. *Information and Management*, *35*(1), 9–18. https://doi.org/10.1016/S0378-7206(98)00079-2
[34] Quinlan, J. R. (1987). Simplifying Decision Trees. In *International Journal of Man-Machine Studies* (Vol. 27, Issue 3, pp. 221–234).
[35] Ramirez, L. (2013). La exhibición y el merchandising en las tiendas de barrio de Bogotá: un canal que cobra fuerza como oportunidad para la construcción de marca. *Revista Civilizar de Empresa y Economía*, *4*(8), 78–96. https://doi.org/10.22518/2462909x.260
[36] Revista Dinero. (2003). La tienda no se rinde. *184*, 6238.
[37] Sarle, W. (1994). Neural networks and statistical models. *Neural Networks and Statistical Learning*, 13. https://doi.org/10.1007/978-1-4471-5571-3
[38] Silva, H. (2011). Comportamiento de las superficies retail en Colombia. *Pensamiento y Gestion*, *30*(1657), 10–20.
[39] Tabachnick, B. (2019). *Using multivariate statistics* (7th ed.). Pearson Education Limited. https://doi.org/10.5455/jmood.20161230045344

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

[40] Takahashi, S., & Inoue, I. (2016). *The Manga Guide to Regression Analysis*. No Starch Press.
[41] Uribe, J. (2015). ICER Informe de coyuntura económica regional 2010. In *Banco de la republica de Colombia* (Vol. 53, Issue 9).
[42] Vargas, J. (2016). Strategic Analysis: Success of Convenience Stores of Mexican Economic and Trade Promotion. *International Journal of Banking Finance & Digital Marketing*, *May*, 11–31.
[43] Widaningrum, D. L. (2015). A GIS - Based Approach for Catchment Area Analysis of Convenience Store. *Procedia Computer Science*, *72*, 511–518. https://doi.org/10.1016/j.procs.2015.12.158

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)