Assessing and forecasting method of financial efficiency in a free industrial economic zone

Tomás José Fontalvo-Herrera*

Faculty of Economics Sciences,
University of Cartagena,
Cartagena, Colombia
Email: tfontalvoh@unicartagena.edu.co
*Corresponding author

Enrique Delahoz-Dominguez

Industrial Engineering Department,
Engineering Faculty,
Universidad Tecnológica de Bolívar,
Cartagena, Colombia
Email: edelahoz@utb.edu.co

Orianna Fontalvo-Echavez

Engineering Faculty,
Universidad del Norte,
Barranquilla, Colombia
Email: oriannaf@uninorte.edu.co

Abstract: Industrial free zones are key to the economic progress of developing countries, making the evaluation and forecast of efficiency in these organisations relevant. This research proposes a three-phase method to evaluate and forecast the financial efficiency of the business profiles of companies belonging to the free economic zone of Cartagena – Colombia. The first phase consisted of a cluster analysis to determine representative groups among the companies analysed. In the second phase, financial efficiency is measured for each of the clusters found in phase 1. Finally, in phase 3 a machine learning model is trained and validated to predict the belonging of a company to a category of financial efficiency – cluster. The results show the creation of two business clusters, with an average efficiency of 49.8% and 14.6% respectively. The random forest model has an accuracy of 95% in the validation phase.

Keywords: data envelope analysis; DEA; clustering; machine learning; random forest; efficiency.

Reference to this paper should be made as follows: Fontalvo-Herrera, T.J., Delahoz-Dominguez, E. and Fontalvo-Echavez, O. (2021) 'Assessing and forecasting method of financial efficiency in a free industrial economic zone', *Int. J. Productivity and Quality Management*, Vol. 33, No. 2, pp.253–270.

Biographical notes: Tomás José Fontalvo-Herrera has more than 25 years of experience as a Headline Professor, Lecturer, consultant and researcher in the fields of quality management, quality control, multivariate calculation, efficiency, productivity, machine learning and management. For over 15 years, he has held management positions. He has worked as a Professor in the Master's of Management of Organizations, Quality Engineering, Administration and Innovation and Integrated Management Systems. He has published more than 100 articles in national and international indexed journals and has been speaker at international level in different countries. He is an evaluator in high quality of the Department of Science and Technology and in the Ministry of National Education of Colombia. He had an internship as a Visiting Professor at the Polytechnic University of Valencia, Spain.

Enrique Delahoz-Dominguez is an industrial engineer with a Master in Operations Research from the Universitat de Barcelona. He is an author of more than 15 papers in Scopus and WOS journals. He is currently working at the Industrial Engineering Department, Technological University of Bolivar. He does his research in machine learning, educational data mining and information systems.

Orianna Fontalvo-Echavez is a junior researcher in the Quality and Integral Organizational Productivity Group, an author of multiples articles in the areas of machine learning, multivariate statistical control, efficiency analysis, big data for business development, service quality and Six Sigma. She is an international speaker in conferences in Spain and the USA.

1 Introduction

Worldwide, free economic zones are organisational entities that promote the exchange of products and services at an international level. What constitutes them in centres of reception, transformation and distribution of merchandise in different economic contexts. However, it is important to understand that the member companies of these free zones have different characteristics and structures to achieve efficiency. Under this concept, the efficiency criterion allows to evaluate the performance of an entity against its economic sector of reference, which constitutes a necessary criterion for the decision making of the operations of this type of companies. Likewise, it is necessary to characterise the types of efficiencies according to business groups, so that they can be organised for analysis, compression and forecasting. Therefore, the use of machine learning tools is essential to predict in the future the characteristics and membership of a new company that wants to be part of that free economic zone.

In accordance with this research, there are many studies that have been using multivariate data analysis, data envelope analysis (DEA) and machine learning, as tools for decision-making in the business sector, as indicated in their studies by different authors (Dia et al., 2019; Garg and Goyal, 2019) associated with data analysis and efficiency respectively. The above is consistent with other research that has developed articulated models for measuring business efficiency by applying different techniques, giving rise to what is known as multilevel or multi-stage studies, similar to that developed in this research. In the context of the variables analysed in the study, (Granadillo et al., 2019) have also integrated the analysis of financial items, performance levels and prognosis to understand the behaviour of productivity indicators in the business sector.

Consequently, Globerson and Vitner (2019) in their research show the relevance of cluster analysis to identify critical patterns or factors within a given context, associated with the explicit performance of financial and operational variables. Based on analysing business contexts, other authors have used machine learning and artificial intelligence to predict the behaviour of key business variables, such as those analysed in their research (Smitha and Rajkumar, 2019; Panjehfouladgaran and Shirouyehzad, 2018). However, Nazir (2019) makes a critique and proposes limitations to the machine learning processes to perform the forecasting process based on the comparability of the units studied, so it is important to determine the levels of accuracy and validity in the forecasting processes where these machine learning analysis are used and in which processes of homogenisation and grouping of the study units are contemplated.

Bailke and Patil (2019), have used the different classification algorithms to optimise the classification processes of key variables such as those analysed in this research. On the other hand, other studies (Sreenivasan and Sundaram, 2018) have used a machine learning model to predict operational performance in a business context, similar to the intentions of this research.

In accordance with the analysis of one of the purposes of this investigation, it is worth noting the relevance of the creation of representative groups through cluster analysis, to calculate productivity profiles in the chemical sector in Colombia as presented in Gomez et al. (2018). Likewise, other research at the international level have studied efficiency through data envelopment analysis in large business sectors. For example, Bhavani et al. (2018) use the techniques of cluster analysis and data envelopment analysis to analyse efficiency in specific contexts, supported by algorithms that allow operating at different levels, as analysed and proposed in this investigation.

In front of all the previous approaches analysed, it is necessary to organise a three-phase method that allows answering the following questions: How to group the business groups taking into account variables and patterns of financial statements that allow business clusters to be established? What is the financial efficiency of calculated business clusters? How to predict through machine learning the efficiency or not of a company that is part of these business clusters?

All of the above, led to consider in this investigation as the main objective and intentionality the design of a three-phase method to evaluate and forecast the financial efficiency of companies in the Mamonal Industrial Zone, as a generalisable proposal with which criteria of analysis for objective decision making and improvement in industrial areas of this type are provided. Considering the exposed elements, the main contributions of this research, are associated with being able to define business clusters in the free zone, through representative companies, which subsequently allow the calculation of the efficiencies of the clusters and as an added value, to contribute to the free zone with the forecast of a new company or one that is part of the free zone already, to which cluster it belongs and if it is efficient or not. With what is provided a structured third level method to classify, evaluate and forecast the belonging of a company.

2 Theorical framework

2.1 Cluster analysis

Cluster techniques are within the domain of unsupervised learning models in machine learning. These models are characterised by their ability to identify patterns of similarity and difference between the observations studied, so they are able to create homogeneous groups, where the average distance between all member elements of the same cluster is minimal and the distance between the different clusters is maximum. Algorithms of the non-hierarchical type will be used for this investigation. K-medoid algorithms iteratively choose k observations as representative medoids, since the medoids are based on a real observation of the dataset, it is less sensitive to the effect of outliers compared to K-means. For the present investigation, the algorithm based on PAM K-medoids was used, the main justification of their choice is to respond to the objective of characterising the profiles of each group found, therefore, it is preferable to have a real element of the dataset as object of analysis and interpretation of the group.

2.1.1 Distance and clustering (PAM)

Euclidean distance was used as a measure of dissimilarity. Thus, the distance between an object i and an object j is given by the equation (1).

$$d_{ij} = \sqrt{\sum_{p=1}^{k} \left(x_{ip} - x_{jp}\right)^2} \tag{1}$$

For the development of the first stage of the proposed method, a non-hierarchical cluster analysis was performed, using a partition algorithm around the medoids (PAM) (Kaufman and Rousseeuw, 2009), which defines a medoid as the representative element of a cluster from which the average dissimilarity towards each of the other members of the cluster is minimal. Consequently, a Silhouette test was developed as it is developed in (Menardi, 2011), to evaluate the quality of the belonging of the observations to its group. This test delivers for each observation a weighting that ranges between values of –1 and 1; the –1 being the assessment for observations that would be better represented in another cluster; the value 0 for observations that is in the border between two clusters; and 1 for observations well coupled to the current cluster.

2.2 DEA efficiency measurement

The main concept of DEA is the evaluation of the efficiency of decision-making units (companies) that interact within a common competition and development station, as is the case of the business sectors. The DEA analysis is also known as border analysis, it has become the standard for the development of efficiency comparison, measurement and evaluation processes in productive organisations (Pawsey et al., 2018). Different approaches can be taken from the point of view of analysis DEA, for example Cook et al. (2019) assess organisational performance in the specific context of incentive plans based on performance (pay-for-performance incentive plans). On the other hand, Ohsato and Takahashi (2015) propose the concept of management efficiency, implementing a network-based DEA model. Ghiyasi (2018) points out that through the model of data envelopment analysis, efficiency and resource estimation can be improved in particular contexts. Similarly, Kumar and Suganthi (2019) show the use and articulation of the enveloping analysis and a forecasting method, to assess efficiency in other contexts.

The DEA, also called border analysis, is a mathematical optimisation model consisting of an objective function h_0 [equation (2)] that represents an efficiency index, and a set of constraints formed by equations and/or inequalities that express limiting conditions for the equation (3) system. The objective function is established by the ratio of the output variables or results and the input variables or resources. The optimisation process involves determining the values of the variables to achieve the maximum or minimum value of the objective function (Mardani et al., 2017). The DEA analysis is a non-parametric technique that determines a relative efficiency frontier from the treatment of several input variables (resources) and several output variables (products or results).

The DEA CCR model, known in the literature as technical efficiency, is the ratio of the weighted sum of the outputs to the weighted sum of the inputs. The intent of the CCR model is to maximise the efficiency of a decision-making unit, within a group of reference organisations, through the optimal weights related to the input and output variables (Benicio and de Mello, 2015). The optimisation model associated with the DEA CCR model, endogenously calculates the weighting of the performance criteria and the result of the variables to reach the maximum or minimum value of the objective function (Sinuany-Stern et al., 2000).

$$Max\left(h_0\right) = \frac{\sum_{r=1}^{s} U_r^*}{\sum_{i=1}^{m} V_i^*} \tag{2}$$

Subject to

$$\frac{\sum_{r=1}^{s} U_r * Y_{rj}}{\sum_{i=1}^{m} V_i * X_{io}} \tag{3}$$

h_0 efficiency index of the observed unit Índice de eficiencia de la unidad observada

s number of output variables (result)

m number of input variables (resources)

U_r relative weight of the nth output variable (positive and unknown)

Y_{ro} value of the nth output variable at observation o

V_i relative weight of the nth input variable (positive and unknown)

x_{io} value of the nth output variable at observation o

n number of observations studied.

2.3 *Random forest model*

The random forest model is an assembly-type method, based on the recurring and growing construction of multiple decision trees through a bootstrapping aggregation process (Breiman, 2001). In other words, multiple decision trees are created, of different variable compositions, so that each tree yields an independent result, to then carry out a process of democracy where a category is assigned according to the most voted resulting class in general. This characteristic of generating separate responses for each decision tree and then joining them in a general prediction produces robust models, less susceptible to extreme values and the problem of overfitting than a simple decision tree, thus improving the predictability and classification of the model. The RF model presents a variable selection technique, in this way it can handle datasets with a large number of variables without using prior processes for reducing dimensions. In addition, the model allows to identify the importance of each variable for the correct classification of the observations, through a permutations test. It is important to note that other research shows the importance of models or recommendation systems for forecasting supported by data mining (Khodabandehlou, 2019). Similarly, other studies (Varma and Padma, 2019) have used different algorithms and models to forecast financial variables, such as those analysed in this research. Other authors, in similar studies (Globerson and Vitner, 2019) propose the articulation of different multivariate techniques to predict scenarios in business sectors.

2.3.1 Performance metrics

The success of the classification process occurs by minimising the difference between the predicted value and the actual value. This relationship is described by the positive true (VP), true negative (VN), false positive (FP), and false negative (FN) metrics. The metrics used to assess performance will be the correct classification rate or accuracy (A), positive predictive value (PPV), negative predictive value (PPN), sensitivity (S) and specificity (E) and the area under the curve (AUC). The AUC represents the rate of TP and FP at various discrimination thresholds. A model with a perfect classification will have an AUC = 1. On the other hand, a totally random model would yield an AUC value = 0.5.

$$A = \frac{TP + TN}{n} \tag{4}$$

$$S = \frac{TP}{TP + FN} \tag{5}$$

$$E = \frac{TN}{TN + FP} \tag{6}$$

3 Methodology

For the development of this research, 145 companies from the free economic zone of Mamonal, Cartagena, which presented their financial statements in 2017 at the financial superintendence of Colombia, were taken as the object of study, from these the financial variables for the investigation were taken. A rational analysis was previously carried out that allowed identifying the variables and items required for the cluster analysis and financial efficiency of the companies in the free economic zone. From the above, the input variables were identified: total assets, total liabilities, total equity, and as output variables: revenue from ordinary activity and operating income. This to perform the data envelopment analysis, which allowed to establish efficient and non-efficient companies of the clusters found.

With the information previously purified and selected, the following steps were carried out.

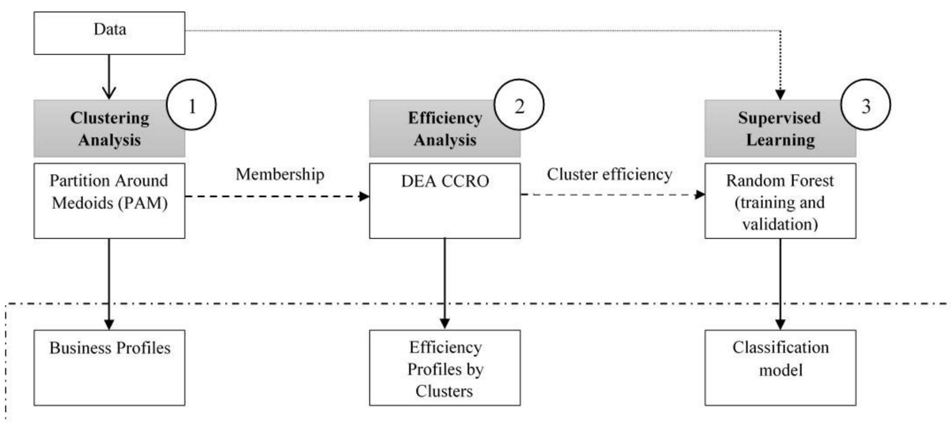
1 An analysis of PAM conglomerate or partition around the medoids, in order to identify the different business groups, associated with their characteristics and performance patterns, which led to define the business profiles of the sector under study.

2 A financial efficiency analysis was subsequently carried out with an outflow optimisation approach, which in turn allowed to establish financial efficiency profiles and finally.

3 A machine learning classification model, based on random forest in order to predict the efficiency or lack of it of all previously defined business financial efficiency profiles.

The above, integrated into the method of evaluation and forecasting of the financial efficiency profiles of the free zone of Cartagena, as can be seen in Figure 1.

Figure 1 Assessment and forecasting method of financial efficiency profiles



In this research as an epistemological conception, the origin of science was generated through a rational analysis that allowed to articulate and structure a three-phase method that integrated the different theoretical, variable, and cluster models, DEA and machine learning. In Figure 1, through a multiple empirical analysis supported by cluster analysis, data envelopment analysis and random forest technique, with the intention of grouping, evaluating and forecasting business efficiency profiles. From the above, it can be asserted that the essence of the science generated is combined whenever it is part of the researchers to understand and propose the object of study, or business efficiency profiles. The truth criteria in this investigation are considered mixed, considering that part of a rational analysis to integrate and structure the proposed evaluation method. But, it also has an empirical component, taking into account that science originates from the observation of the variables associated with business groups, with which a statistical inference is made and once analysed, the level of validity of these inferences or projections is verified and calculated. The logic of the method of this study is on one side inductive, since it starts from empirical information, but it is also deductive, considering that the articulation and structuring of the method required a rational construction for its design.

3.1 Data

The data used in this investigation correspond to 145 companies belonging to the free zone of Mamonal. The source of the data corresponds to the Chamber of Commerce of Cartagena – Colombia. Entity responsible for managing and publishing the financial information of the city's companies. The real names of the companies studied are not shown to preserve their anonymity. In this way, the names of the companies or decision-making units are represented by numbers (1, 2, 3 …, 145).

3.2 Analysis of the information

For the analysis of the information, the first phase was carried out with the support of the R software, which allowed defining the conglomerates or financial efficiency profiles of the companies under study. Subsequently, the financial efficiency of the business profiles was assessed, using the DEA technique with an output optimisation approach and the CCR-O model. Finally, the random forest machine learning algorithm was used to train and predict the four outputs of the two profiles of financial efficiency of enterprises in the free economic zone in Cartagena. In phases two and three, the R software was also used.

4 Results

4.1 First stage results

For the development of the proposed method, initially clusters were created through the PAM algorithm by varying the parameter of number of groups k from 2 to 10 to identify the formation with the best adjustment of membership of the companies studied. The highest value of the test is obtained for an analysis with formation of two clusters and a Silhouette value of 0.65 (see Figure 2).

Using two groups as a parameter of inputs to the PAM analysis, we proceeded to the analysis of the representative elements of each group (see Figure 3), where it can be seen how the companies belonging to cluster 2 have consistently higher values for each of the financial items analysed. In this context, the general analysis of the studied companies allows characterising the companies in cluster 1 as support and backing companies for the operation of the Mamonal industrial free economic zone, within this group are companies dedicated to the logistics services sector, transportation, customs services, temporary employees, restaurants and insurance agencies.

In relation to cluster 2, the analysis shows the membership of this group of large industrial and logistic companies, which are the companies that give rise to the free zone due to its size, production capacity, number of employees local economic impact. It is important to highlight in this group the belonging of an oil refinery, which presents quite different financial results to the rest of the companies, in addition to companies in the petrochemical, shipbuilding and metalworking sector.

Figure 2 Silhouette test analysis for cluster of size 1 to 10 (see online version for colours)

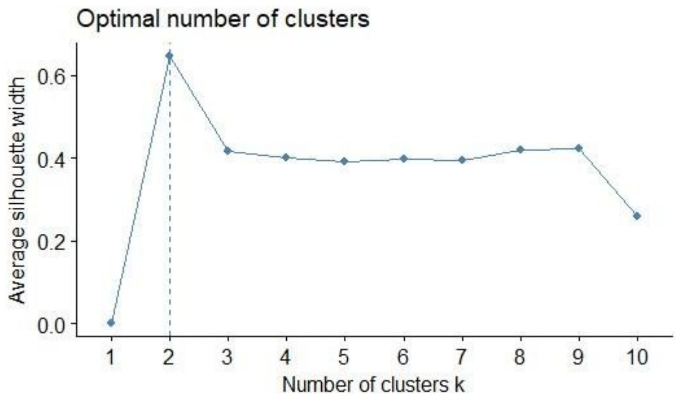
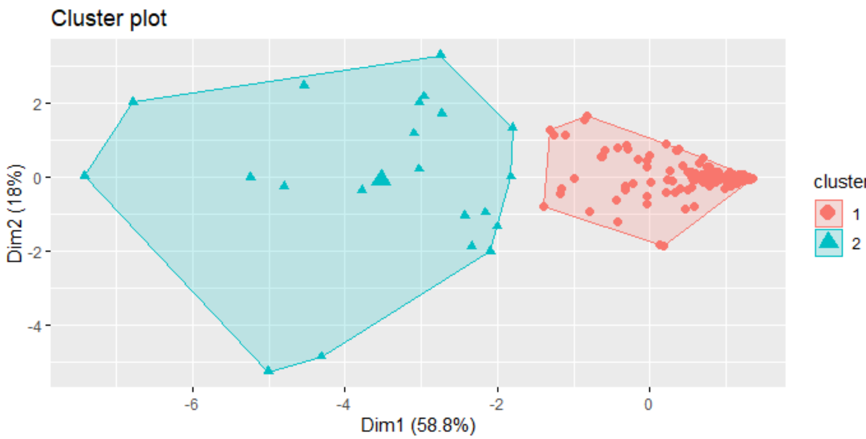


Figure 3 Medoids comparison by cluster (see online version for colours)

Additionally, it can be seen in Figure 4, the two-dimensional visual representation of the clusters found using the PAM algorithm, here the two groups are clearly separated. Cluster 1 is located on the right side of the plan, representing a compact and homogeneous group of companies, endorsing the concept of these companies as support and support companies for the industrial processes of the companies belonging to cluster 2. Similarly, on the side on the left of the plane in Figure 4, cluster 2 of companies is observed, as a slightly dispersed group, characterised by the variety of industrial sectors represented by the companies. Which serves as significant evidence to justify the clustering process proposed in the three-phase method presented in this research, given that the diversity of sectors present in the free zone is a vital factor to consider in order to develop a subsequent analysis and forecast of efficiency. In summary, from the first phase two business clusters are created and validated within the industrial free zone studied. Cluster 1 will be labelled as support companies and cluster 2 as dynamic industrial companies. The groups found here will serve as inputs for the second phase of the proposed method.

Figure 4 Two dimensional representation of PAM clustering (see online version for colours)



4.2 *Second stage results*

To guarantee the integration of the method, we proceeded to calculate the efficiency of the two previously established clusters. The general results of the DEA model implemented in the clusters obtained in the first phase are found in Table 1. This shows how the average efficiency of the group of support companies (cluster 1) is equal to 0.14 with a standard deviation of 0.25. Of the total of 115 companies that make up this cluster, only 7 (6%) are efficient in relation to the variables studied, that is, these are at the frontier of efficiency and the clearance of their variables is equal to zero. The results of the efficiency score for DMUs corresponding to support companies are shown in Table 2.

Table 1 DEA model summary

	Support companies	*Dynamic industrial companies*
Mean efficiency	0.14	0.49
Minimum efficiency	0.00005	0.008
Maximum efficiency	1	1
Standard deviation	0.25	0.34
Number of companies	115	30
Efficient companies	7 (6%)	5 (17%)
Reference companies	6	4

In turn, the group of dynamic industrial companies (cluster 2) has an average efficiency value of 0.49 and a standard deviation of 0.34. Of the total of 30 companies that make up this group 5 (17%) are efficient. The efficiency results by DMU can be seen in Table 3. From the empirical evidence it can be noted that these industrial and dynamic companies have stronger financial structures, which is reflected in their higher average efficiency. Tables 2 and 3 show the efficiencies of the two previously calculated business profiles.

Table 2 DMU's efficiency results for dynamic industrial companies

Rank	*DMU*	*Score*	*Rank*	*DMU*	*Score*
1	*1*	1	16	16	0.3833
1	*2*	1	17	17	0.3815
1	*3*	1	18	18	0.3812
1	*4*	1	19	19	0.3030
1	*5*	1	20	20	0.2821
6	*6*	0.9931	21	21	0.2614
7	*7*	0.9481	22	22	0.2487
8	*8*	0.8526	23	23	0.1836
9	*9*	0.7655	24	24	0.1633
10	*10*	0.7653	25	25	0.1346
11	*11*	0.6897	26	26	0.1108
12	*12*	0.5618	27	27	0.1005
13	*13*	0.4810	28	28	0.0958
14	*14*	0.4161	29	29	0.0650
15	*15*	0.3890	30	30	0.0083

4.3 Third stage results

4.3.1 Random forest results

With the four outputs of the efficient and non-efficient companies of the two previously calculated business profiles, we proceeded to forecast using the machine learning algorithm. The random forest model with the best performance, yielded a mean accuracy (0.89) and AUC = 0.95 during the training phase based on ten-fold cross validation (Table 4). In the test phase, the model identified the efficient companies of the dynamic group with (100%) of sensitivity and the efficient companies of the support cluster with (66.7%) of sensitivity. The AUC of the receiver operating characteristic (ROC) was equal to 94.5% for the predictions of the RF model (Table 6). The 95% accuracy results achieved in this investigation are significantly good, considering that when similar tools associated with logistic activity have been used in similar contexts, other research has shown precision results below 52.2% and 60.6 (Herrera, 2014; Fontalvo Herrera et al., 2012). Similarly, other studies have shown the relevance of using tools in the same category to establish prognosis or recommendation processes similar to those developed in this research (Khodabandehlou, 2019). It is important to note that other similar investigations of application of several similar tools in multistages that have articulated cluster analysis tools and forecasting processes show the effectiveness to group and forecast financial variables with those used in this research (Mahjoub and Afsar, 2019).

Table 3 DMU's Efficiency results for support companies

Rank	*DMU*	*Score*	*Rank*	*DMU*	*Score*	*Rank*	*DMU*	*Score*	*Rank*	*DMU*	*Score*
1	1	1	34	34	0.11275	67	67	0.0334	100	100	0.00669
1	2	1	35	35	0.11163	68	68	0.03327	101	101	0.00519
1	3	1	36	36	0.11002	69	69	0.03059	102	102	0.00408
1	4	1	37	37	0.1061	70	70	0.02883	103	103	0.00403
1	5	1	38	38	0.10101	71	71	0.02843	104	104	0.00394
1	6	1	39	39	0.09977	72	72	0.02338	105	105	0.00353
1	7	1	40	40	0.07896	73	73	0.02264	106	106	0.00204
8	8	0.76643	41	41	0.07708	74	74	0.0204	107	107	0.00199
9	9	0.72996	42	42	0.07692	75	75	0.01941	108	108	0.00136
10	10	0.63512	43	43	0.07172	76	76	0.01862	109	109	0.00087
11	11	0.44861	44	44	0.06709	77	77	0.01839	110	110	0.00086
12	12	0.41055	45	45	0.06631	78	78	0.0177	111	111	0.00037
13	13	0.39596	46	46	0.06575	79	79	0.01679	112	112	0.0003
14	14	0.37763	47	47	0.06457	80	80	0.01637	113	113	0.00014
15	15	0.27699	48	48	0.06129	81	81	0.01626	114	114	0.00005
16	16	0.27047	49	49	0.0599	82	82	0.01547	115	115	0.00005
17	17	0.23694	50	50	0.05964	83	83	0.01531			
18	18	0.23462	51	51	0.05674	84	84	0.0151			
19	19	0.21139	52	52	0.0556	85	85	0.0147			
20	20	0.20206	53	53	0.05014	86	86	0.01465			
21	21	0.20006	54	54	0.04978	87	87	0.01418			

Table 3 DMU's Efficiency results for support companies (continued)

Rank	*DMU*	*Score*	*Rank*	*DMU*	*Score*	*Rank*	*DMU*	*Score*	*Rank*	*DMU*	*Score*
22	22	0.19206	55	55	0.04937	88	88	0.01346			
23	23	0.18312	56	56	0.04714	89	89	0.01344			
24	24	0.15981	57	57	0.04664	90	90	0.01266			
25	25	0.14602	58	58	0.04545	91	91	0.01033			
26	26	0.14584	59	59	0.04391	92	92	0.00956			
27	27	0.14464	60	60	0.04268	93	93	0.00954			
28	28	0.14385	61	61	0.04104	94	94	0.00851			
29	29	0.13227	62	62	0.04047	95	95	0.00828			
30	30	0.12625	63	63	0.03907	96	96	0.00787			
31	31	0.12428	64	64	0.03862	97	97	0.00723			
32	32	0.11987	65	65	0.03571	98	98	0.00686			
33	33	0.11546	66	66	0.03548	99	99	0.00681			

Table 4 Predictive performance metric of the RF model

	Accuracy				*Roc*			
	Min	*Mean*	*Max*	*sd*	*Min*	*Mean*	*Max*	*sd*
RF model	0.72	0.87	1	0.08	0.73	0.92	1	0.02

The cross-validation process allowed to determine the consistency of the model. The results show a reduced standard deviation for the results of the Accuracy and AUC metrics with values of 0.08 and 0.02 respectively (Table 4).

Table 5 Variables importance for the RF model

Variable	*Importancia (%)*
Ordinary_incomes	100
Total_asset	84.67
Total_liabilities	48.17
Operational_profit	11.52
Total_equity	0

The representative values of importance of the variables in the model were calculated in order to determine the key variables that allow companies to be classified as efficient according to their role within the industrial zone. In Table 5, it is observed that the variable 'ordinary incomes' is the most important factor, followed by 'total asset', total liabilities and operational profit. It is observed that the total equity variable does not contribute to the classification model.

Table 6 Metrics summary for the validation process

Metric	*Efficient_G1*	*Not Efficient G1*	*Efficient_G2*	*Not_Efficient_G2*
Sensitivity	1.000	0.857	0.667	1.000
Specificity	1.000	1.000	1.000	0.833
Pos Pred value	1.000	1.000	1.000	0.941
Neg Pred value	1.000	0.974	0.976	1.000
Prevalence	0.045	0.159	0.068	0.727
Detection rate	0.045	0.136	0.045	0.727
Detection prevalence	0.045	0.136	0.045	0.773
Balanced accuracy	1.000	0.929	0.833	0.917

From Table 6 it is important and significant to analyse the great capacity of the machine learning model of forecasting, to determine the specificity of the four variables associated with the forecast of efficiency. In this sense, it is important to contrast these results with other research where machine learning and artificial intelligence have been used to predict belonging to a business cluster and/or type of efficiencies in business sectors, which is consistent with the results found in this research (Fontalvo et al., 2019; De La Hoz et al., 2019a; Fontalvo et al., 2018). However, as a differential element that generates value in this research, there is the fact that a more structured third level method is used for the assessment, classification and forecasting, compared to other investigations that show smaller applications that use simpler methods, and just a second level. Similarly, the authors (Lin et al., 2012), use DEA and subsequently forecast with machine Learning, showing the relevance through a second level method. Other research uses the machine learning (De La Hoz et al., 2019b) to develop forecasting processes with structures similar to those used in this research. All these investigations present levels of precision similar to those found in this study.

5 Conclusions

This article fully evaluated the financial efficiency for the 145 companies in the Mamonal Industrial Free Zone in Cartagena – Colombia. For this, a three-stage method was developed to clarify the effect that large companies have on the overall efficiency results of the sector. The greatest contribution of this research was the implementation of a three-phase method to evaluate and forecast the financial efficiency of a business sector, which allows through a cluster analysis (first stage) the grouping of companies with similar financial characteristics in clearly defined clusters. In this way, the DEA analysis is carried out in a fair manner, comparing homogeneous companies in their financial dimensions. It is clear to note how previous studies have developed DEA models with multiple stages, using the analysis of main PCA components as a stage for reducing the number of variables, or also deep learning models for predicting efficient companies. However, this research integrates the three concepts into a method; clustering, efficiency and forecast articulated in a method. Which constitutes a scientific contribution and a tool for decision-making in free zones where such methods are implemented.

From the empirical evidence, the following criteria can be identified as investigative findings. The results of the first stage show the conformation of two groups; the first formed by manufacturing companies and logistics operators and a second group made up of support and support companies, such as transporters, catering services, legal services, maintenance and general services. In the second stage, there is a significant difference between the average efficiency of the 49% dynamics cluster and the 14% support companies, endorsing the initial purpose of this research. Finally, in the third phase the random forest model was trained and validated, which obtained a high percentage of success for the prediction of a free zone company to a category of financial efficiency – cluster. In addition to the random forest model identified the ordinary Incomes variable as the most important in the process of classifying companies as efficient or inefficient.

In general, a structured method for analysing, measuring and forecasting the efficiency of a business sector is presented to the scientific community and similar business sectors internationally. The above allows the proposed method to be replicable and reproducible. What facilitates the objective decision making for the generation of value in other business contexts.

The main limitation in this investigation was the difficulty of using other variables of the companies in the free zone, such as the number of workers or variables associated with rationality that contributed to efficiency. Therefore, it is proposed to use the three-level method in future research using other variables and compare with other machine learning algorithms apart from random forest, which allows to contrast the research results.

References

Bailke, P.A. and Patil, S.T. (2019) 'Distributed algorithms for improved associative multilabel document classification considering reoccurrence of features and handling minority classes', *International Journal of Business Intelligence and Data Mining*, Vol. 14, No. 3, pp.299–321, DOI: 10.1504/IJBIDM.2019.098843 (accessed 10 September 2019).

Benicio, J. and de Mello, J.C.S. (2015) 'Productivity analysis and variable returns of scale: DEA efficiency frontier interpretation', *Procedia Computer Science, 3rd International Conference on Information Technology and Quantitative Management, ITQM 2015*, Vol. 55, pp.341–349, DOI: 10.1016/j.procs.2015.07.059 (accessed 29 July 2019).

Bhavani, R., Prakash, V. and Chitra, K. (2018) 'An efficient clustering approach for fair semantic web content retrieval via tri-level ontology construction model with hybrid dragonfly algorithm', *International Journal of Business Intelligence and Data Mining*, Vol. 14, Nos. 1–2, pp.62–88, DOI: 10.1504/IJBIDM.2019.096836 (accessed 10 September 2019).

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32, DOI: 10.1023/A:1010933404324 (accessed 25 April 2019).

Cook, W.D., Ramón, N., Ruiz, J.L., Sirvent, I. and Zhu, J. (2019) 'DEA-based benchmarking for performance evaluation in pay-for-performance incentive plans', *Omega*, Vol. 84, No. C, pp.45–54 [online] https://econpapers.repec.org/article/eeejomega/v_3a84_3ay_3a2019_3ai_3ac_3ap_3a45-54.htm (accessed 29 July 2019).

De La Hoz, E., De La Hoz, E. and Fontalvo, T. (2019a) 'Metodología de aprendizaje automático para la clasificación y predicción de usuarios en ambientes virtuales de educación', *Informacion Tecnologica*, Vol. 30, No. 1, pp.247–254 [online] http://dx.doi.org/10.4067/S0718-07642019000100247.

De La Hoz, E., Fontalvo, T. and Lopez, L. (2019b) 'Data envelopment analysis and multivariate calculus to assess, classify and predict the productive efficiency and innovation of companies in the chemical sector', *Informacion Tecnologica*, Vol. 30, No. 5, pp.213–220 [online] http://dx.doi.org/10.4067/S0718-07642019000500213.

Dia, M., Abukari, K., Takouda, P.M. and Assaidi, A. (2019) 'Relative efficiency measurement of Canadian mining companies', *International Journal of Applied Management Science*, Vol. 11, No. 3, pp.224–242, DOI: 10.1504/IJAMS.2019.101002 (accessed 10 September 2019).

Fontalvo Herrera, T., De la Hoz Granadillo, E. and Vergara, J. C. (2012) 'Aplicación de análisis discriminante para evaluar el mejoramiento de los indicadores financieros en las empresas del sector alimento de Barranquilla-Colombia', *Ingeniare. Revista chilena de ingeniería*, Vol. 20, No. 3, pp.320–330.

Fontalvo, T., De La Hoz, E. and Morelos, J. (2018) 'Combined method of conglomerate analysis and multivariate discriminant analysis to identify and evaluate financial efficiency profiles in exporting companies', *Información Tecnologica*, Vol. 29, No. 5, pp.227–234 [online] http://dx.doi.org/10.4067/S0718-07642018000500227.

Fontalvo, T., De La Hoz, E. and Olivos, S. (2019) 'Methodology of data envelopment analysis (DEA) – GLMNET for assessment and forecasting of financial efficiency in a free trade zone – Colombia', *Información Tecnologica*, Vol. 30, No. 5, pp.263–270 [online] http://dx.doi.org/10.4067/S0718-07642019000500263.

Garg, A. and Goyal, D.p. (2019) 'Sustained business competitive advantage with data analytics', *International Journal of Business and Data Analytics*, Vol. 1, No. 1, pp.4–15, DOI: 10.1504/IJBDA.2019.098829 (accessed 10 September 2019).

Ghiyasi, M. (2018) 'Efficiency improvement and resource estimation: a tradeoff analysis', *International Journal of Productivity and Quality Management*, Vol. 25, No. 2, pp.151–169, DOI: 10.1504/IJPQM.2018.094758 (accessed 10 September 2019).

Globerson, S. and Vitner, G. (2019) 'Measuring productivity in multi-stage, multi-product environment', *International Journal of Productivity and Quality Management*, Vol. 26, No. 3, pp.290–304, DOI: 10.1504/IJPQM.2019.098365 (accessed 10 September 2019).

Gomez, J.M., Herrera, T.J.F. and Granadillo, E.D.L.H. (2018) 'Behaviour of productivity indicators and financial resources in the field of extraction and exploitation of minerals in Colombia', *International Journal of Productivity and Quality Management*, Vol. 25, No. 3, pp.349–367, DOI: 10.1504/IJPQM.2018.095651 (accessed 10 September 2019).

Granadillo, E.D.L.H., Gomez, J.M. and Herrera, T.J.F. (2019) 'Methodology with multivariate calculation to define and evaluate financial productivity profiles of the chemical sector in Colombia', *International Journal of Productivity and Quality Management*, Vol. 27, No. 2, pp.144–160, DOI: 10.1504/IJPQM.2019.100141 (accessed 10 September 2019).

Herrera, T.J.F. (2014) 'Aplicación de análisis discriminante para evaluar la productividad como resultado de la certificación BASC en las empresas de la ciudad de Cartagena', *Contaduría y administración*, Vol. 59, No. 1, pp.43–62.

Kaufman, L. and Rousseeuw, P.J. (2009) *Finding Groups in Data: An Introduction to Cluster Analysis*, Vol. 344, John Wiley & Sons.

Khodabandehlou, S. (2019) 'Designing an e-commerce recommender system based on collaborative filtering using a data mining approach', *International Journal of Business Information Systems*, Vol. 31, No. 4, pp.455–478, DOI: 10.1504/IJBIS.2019.101582 (accessed 10 September 2019).

Kumar, V.R. and Suganthi, L. (2019) 'Relative efficiency of social CRM software: a hybrid fuzzy AHP/DEA approach', *International Journal of Business Information Systems*, Vol. 31, No. 1, pp.27–44, DOI: 10.1504/IJBIS.2019.099525 (accessed 10 September 2019).

Lin, W.Y., Hu, Y. and Tsai, C. (2012) 'Machine learning in financial crisis prediction: a survey', *IEEE Transactions on Systems, Man, and Cybernetics, Part C, Applications and Reviews*, Vol. 42, No. 4, pp.421–436, DOI: 10.1109/TSMCC.2011.2170420.

Mahjoub, R.H. and Afsar, A. (2019) 'A hybrid model for customer credit scoring in stock brokerages using data mining approach', *International Journal of Business Information Systems*, Vol. 31, No. 2, pp.195–214 [o]. DOI: 10.1504/IJBIS.2019.100279 (accessed 26 February 2020).

Mardani, A., Zavadskas, E.K., Streimikiene, D., Jusoh, A. and Khoshnoudi, M. (2017) 'A comprehensive review of data envelopment analysis (DEA) approach in energy efficiency', *Renewable and Sustainable Energy Reviews*, Vol. 70, pp.1298–1322, DOI: 10.1016/j.rser.2016.12.030 (accessed 26 July 2019).

Menardi, G. (2011) 'Density-based Silhouette diagnostics for clustering methods', *Statistics and Computing*, Vol. 21, No. 3, pp.295–308, DOI: 10.1007/s11222-010-9169-0 (accessed 28 September 2018).

Nazir, A. (2019) 'A critique of imbalanced data learning approaches for big data analytics', *International Journal of Business Intelligence and Data Mining*, Vol. 14, No. 4, pp.419–457, DOI: 10.1504/IJBIDM.2019.099961 (accessed 10 September 2019).

Ohsato, S. and Takahashi, M. (2015) 'Management Efficiency in Japanese Regional Banks: a network DEA', *Procedia - Social and Behavioral Sciences, Contemporary Issues in Management and Social Science Research*, Vol. 172, pp.511–518 [online] DOI: 10.1016/j.sbspro.2015.01.394 (accessed 29 July 2019).

Panjehfouladgaran, H. and Shirouyehzad, H. (2018) 'Classification of critical success factors for reverse logistics implementation based on importance-performance analysis', *International Journal of Productivity and Quality Management*, Vol. 25, No. 2, pp.139–150, DOI: 10.1504/IJPQM.2018.094757 (accessed 10 September 2019).

Pawsey, N., Ananda, J. and Hoque, Z. (2018) 'Rationality, accounting and benchmarking water businesses', *International Journal of Public Sector Management*, DOI: 10.1108/IJPSM-04-2017-0124 (accessed 29 July 2019).

Sinuany-Stern, Z., Mehrez, A. and Hadad, Y. (2000) 'An AHP/DEA methodology for ranking decision making units', *International Transactions in Operational Research*, Vol. 7, No. 2, pp.109–124, DOI: 10.1016/S0969- 6016(00)00013-7 (accessed 29 July 2019).

Smitha, J.a. and Rajkumar, N. (2019) 'Efficient moving vehicle detection for intelligent traffic surveillance system using optimal probabilistic neural network', *International Journal of Business Intelligence and Data Mining*, Vol. 15, No. 1, pp.22–48, DOI: 10.1504/IJBIDM.2019.100466 (accessed 10 September 2019).

Sreenivasan, S. and Sundaram, M. (2018) 'A probabilistic model for predicting service level adherence of application support projects', *International Journal of Productivity and Quality Management*, Vol. 25, No. 3, pp.305–330, DOI: 10.1504/IJPQM.2018.095648 (accessed 10 September 2019).

Varma, G.N. and Padma, K. (2019) 'Forecasting agricultural commodity pricing using neural network-based approach', *International Journal of Business Information Systems*, Vol. 31, No. 4, pp.517–529, DOI: 10.1504/IJBIS.2019.101584 (accessed 10 September 2019).