AI & Society

Movie films consumption in Brazil: an analysis of Support Vector Machine classification

Correspondent author: **Marislei Nishijima**, marislei@usp.br
Associate Professor at Institute of International Relations - University of Sao Paulo
Av. Prof. Lúcio Martins Rodrigues, s/n, travessas 4 e 5
Cidade Universitária- CEP 05508-020 - São Paulo -SP - Brazil
+55(11) 3091-0526
Orcid: 0000-0003-1162-7987

Nathalia Nieuwenhoff, nathalia.nieuwenhoff@gmail.com
School of Arts, Science and Humanities - University of Sao Paulo

Ricardo Pires, ricardopires29@gmail.com
Federal Institute of Sao Paulo

Patrícia R. Oliveira, proliveira@usp.br
School of Arts, Science and Humanities of University of Sao Paulo

Abstract: We employ the support vector machine (SVM) classifier, over different types of kernels, to investigate whether observable variables of individuals and their household information are able to describe their consumption decision of film at theaters in Brazil. Using a very big dataset of 340,000 individuals living in metropolitan areas of a whole large developing economy we performed a Knowledge Discovery in Databases (KDD) to classify the film consumers, which results in 80% instances correctly classified. To reduce the degrees of freedom for SVM and to learn the more important determinants of film consumption, we apply the Linear Discriminant Analysis (LDA), that allow us to identify the key determinants of this consumption. The main individual characteristics are age, education (that merges to be a student), income, and preferences for cultural goods. Regarding the main geographic characteristics, these are the timing of sample, population concentration, and supply of movie theaters. The results point to an ineffective policy for the sector at the time investigated.

Key words: *film at theaters, SVM, LDA, KDD, classification, consumers, individual data*

Movie films consumption in Brazil: an analysis of Support Vector Machine classification

1. Introduction

Continuous technological progress has changed the pattern of information goods consumption mainly due to the wide access provided by Internet broadband. Now it is possible to consume a film by means of several alternative media to the movie theaters, because films are also available for consumption on the Internet, for example, either as paid content or by means of piracy.

According to consumer theory (Jehle and Reny, 2000), expenditures in consumption unveil not only individual preferences, but social priority and personal value as well, given the scarcity of resources. In addition, the consumption of culture, which includes many types of information goods, expresses the economic, social and cultural capital situation of an individual, his or her family background, groups of influence and distribution and citizenship condition (Scott 2017).

As a consequence, consumption decisions depend on individual characteristics, their preferences, and their related demographic, cultural, economic and social background (Jehle and Reny 2000). Many of these variables, however, are not directly observable, for example, it is not possible to observe the individual preferences or their willing to pay for a film ticket. Given all these complex interactions, we use computational techniques to identify the main determinants of consumption decision and to classify individuals as film consumers or non-consumers.

To conduct this task, we follow Fayyad, Shapiro and Smyth (1996) in performing Knowledge Discovery in Databases (KDD), which consists on applying computational techniques for data analysis and knowledge generation.

We execute the first stage of our KDD process by defining the initial dataset to be investigated, as characteristics/determinants of film consumers. We base its conduction on consumer theory, as mentioned above. Then we apply Linear Discriminant Analysis (LDA) to assess whether a subset of the initial set of variables, or observed attributes, is more informative for classifying individuals as consumer or non-consumer in comparison to the initial set of variables (Fisher 1936). The classification task was performed by the Support Vector Machine (SVM), a data-mining technique, and explored different kernels. SVM is the most frequent classifier used recently in the literature, according to a systematic review we conducted, where results attest its best results.

Under the assumptions that i) consuming films at theaters, as cultural good, is expensive and ii) the inequality in Brazil is very high in all dimensions (i.e., regarding personal income, regional income and in other levels), we suppose, and later confirm, that only a very small part of the population goes to the movies. Our hypothesis is that the individual observable differences, as well as their social and economic background, allows us to classify them as either consumers or non-consumers of this good, even when the preferences are not directly observable.

We use individual data from the Brazilian Consumer Expenditure Survey (POF) - carried out during a period of 12 months in the years 2002/2003 and 2008/ 2009, and made available by Brazilian Institute of Geography and Statistics (IBGE) - to classify Brazilians as either consumers or non-consumers of films at theaters. Our sample includes around 340,000 individuals.

From our knowledge, there are quite a few film-consumption studies based directly on individuals, or on the demand side. In general, the literature uses information from the box office data to study their consumption (Moretti 2011; Chen, Chen and Weinberg 2013). This

very rich dataset also allow us to investigate also who is not a consumer, differently from using marked data where you only observe who consumes the good.

Brazil is a very important case in point since it is an extremely unequal country and its pattern of films consumption might shed light on who is the representative consumer. This information may give guidelines to evaluate whether the policy of setting half-price tickets for students and elderly people, running at that time, is suitable for reducing the inequality of culture goods access.

The remainder of this paper is organized as follows: Section 2 presents theoretical aspects. In Section 3 we describe our dataset and methodological strategy. Results and discussion are in Section 4. Section 5 concludes with our final remarks.

2. Theoretical aspects

The traditional model that transforms data into knowledge consists of data processing by specialists in a way that make the data understandable. Currently, however, such practice has become a very complex task due to the process of automation, and mainly, due to the huge amount of data, frequently available from different sources and datasets. This leads to the emergence of the concept of Knowledge Discovery in Databases (KDD) (Fayyad, Shapiro and Smyth 1996). Although data mining and KDD are considered as synonymous in some literature (Amo and Rocha 2003), the most common approach considers data mining as a step of the KDD process.

We conduct a KDD focusing on data-mining stage, which is supported by the classification algorithm Support-Vector Machine (SVM). The stages of KDD process are: i) data cleaning, ii) data integration, iii) data selection, defining the relevant attributes and variables (in a first stage we select variables according the consumer theory, in the second stage we select a subsample among them using the Information Gain measure), iv) data

transformation, v) data mining, application of machine learning algorithms for extraction of patterns and data classification, vi) evaluation or post-processing analysis of results regarding the variables that determine the data classification and pattern recognition, and vii) visualization of results.

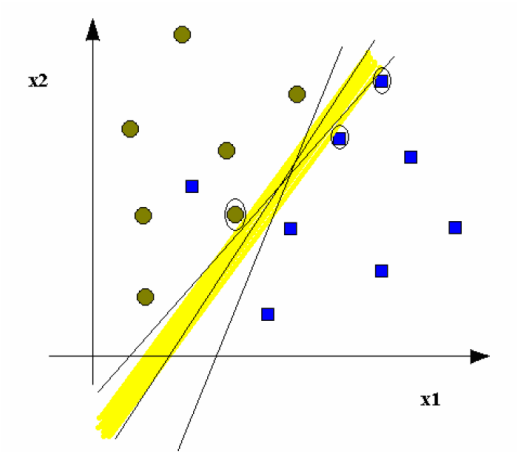
According to Mitchell (1997), machine learning (ML) is a subfield of the computer science area that focuses on the development of computational methods in order to extract knowledge from data. Several ML algorithms are used to generate classifiers for a set of records. We follow Russel and Norvig (1995), defining classification as a process of assigning information to the class to which it belongs.

We apply the Linear Discriminant Analysis (LDA) to assess whether a subset of the initial set of variables, or observed attributes, is more informative for classifying individuals as consumer or non-consumer in comparison to the initial set of variables (Fisher 1936). When each element to be classified is described as a vector containing a set of parameters values and some parameters are not relevant to the classification or are redundant, LDA allows the identification of the parameters that best discriminate elements as belonging to their classes (McLachlan 2004). This makes it possible a reduction in the number of parameters to be used without sacrificing the classification results, which reduces the computational effort of machine-learning algorithms.

The SVM approach is an usual approach for binary classification (two classes), which is based on a learning algorithm from statistical theory, even though it is possible to generalize for more, but a finite, number of classes (Kinto 2011). The non-probabilistic standard SVM application includes a set of variables such as inputs and two possible classes as outputs. SVM identifies patterns in a vector space by finding an optimal decision surface to separate the training patterns belonging to each class. In this context, this “optimal” choice is specified as the largest-margin separation hyperplane, where the term “margin” accounts for the distance

from this hyperplane to the nearest points of each class. Since each instance in the dataset is treated as a vector, the critical instances at the margin are referred as "support vectors". In order to achieve a good generalization capacity, the objective of the algorithm is to choose the best discriminating function from the classes of all the possible ones. Figure 1 illustrates the support vectors at separation margins between classes.

Figure 1: Support-vectors at the separation margin between classes: these special instances are highlighted with a circle around them



Source: Kinto (2011)

3. Data and Methodological strategy

Regarding data, we use individual information from the 2002/2003-2008/2009 Consumer Expenditure Survey (POF) performed by the Brazilian Institute of Geography and Statistics (IBGE), and get a sample of 340,000 individuals. Following our procedure of data integration, we connect each individual to his/her household and family to track his characteristics in order to find out if he/she is a consumer or not of films at theaters, and any other information regarding related goods consumption. The main variables considered in our study can be seen in Table 1.

The early stages of KDD, including dataset cleaning, integration, selection, and treatment of available data was conducted by using Stata - Data Analysis and Statistical

Software. Data-mining activities, such as classification experiments accomplished by SVM was carried out by using the libSVM library (Chang and Lin, 2011) and Linear Discriminant Analysis with the software Octave (Eaton et al., 2014).

Following Jehle and Reny (2000), we define which variables are relevant for the individual consumer decision and then conduct the data cleaning, integration, and data selection activities. As consumption is a consequence of the consumer decision-making process, which relies on individual preferences, income, the supply of products, etc., we use all related variables available in the POF dataset as relevant variables to determine films consumption. According to consumer theory, individuals make consumption choices because they have income restrictions and many necessities of goods. Thus, personal characteristics of individuals, as well as their psychological, social, and cultural influences affect their decision-making process.

We also follow Diniz and Machado (2011) - who use microdata from POF 2002-2003 and studies cultural goods and services consumption in the Brazilian metropolitan regions - to select the relevant economic variables for our study. The authors conclude that cultural consumption, which includes information goods, such as film at theaters, is affected by socioeconomic, educational and socio-demographic variables of the head of household, as well as by variables that indicate the locality and region of the households, which is connected with the supply of cultural goods. The authors adopt a microeconomic approach of human capital, where the consumption of culture is strongly determined by previous individual exposure to such goods and services, and identifies some individual characteristics that influence this exposure.

Accordingly, for this paper, we select the group of variables to be investigated as relevant inputs by considering that consumer choice depends on preferences, observable attributes related to personal characteristics, such as gender, age, religion, skin color and race,

medical aid and education level, if a person is a student, and number of hours usually worked per week.

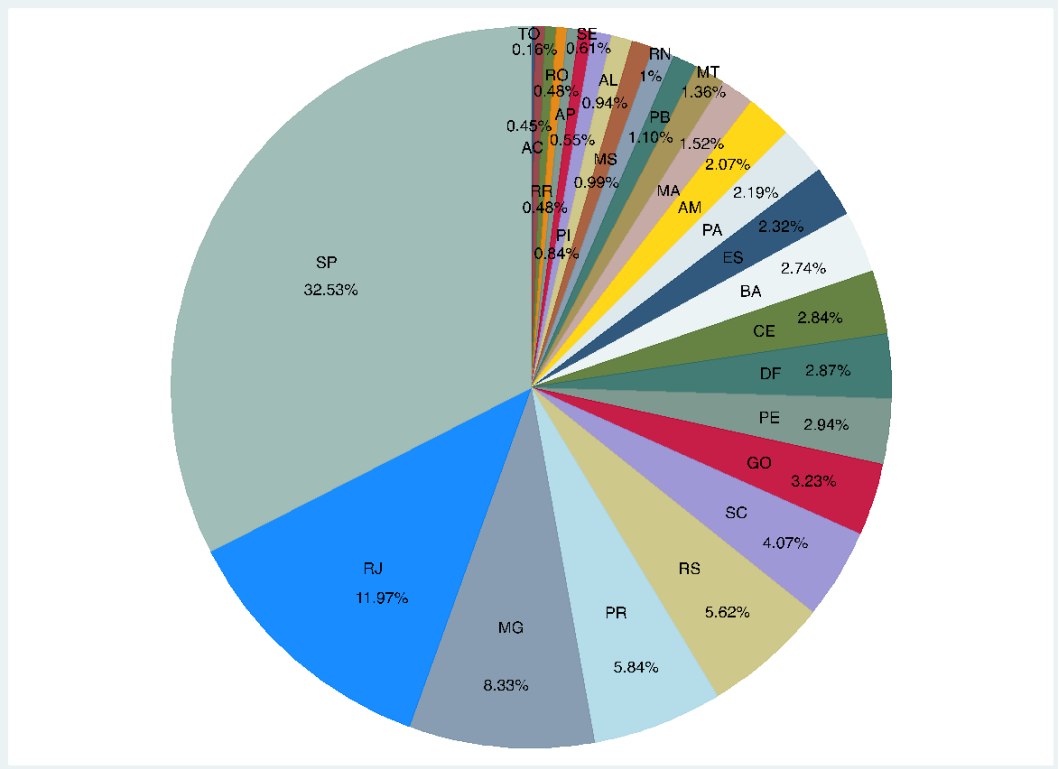
Considering individual socioeconomic factors, as well as the household factors to which the individual belongs, we select attributes to measure these aspects, such as type of household (own or rented house), number of bathrooms, number of rooms, number of families, number of house dwellers, and number of consumption units, as well as the family monetary income.

In addition, regarding product supply factors, we consider the attributes related to individual socio-demographic aspects, such as the individual weights in the POF research sample that represents their relative importance in the local population, as well as states where they are located and types of regions they live in, such as rural and urban areas.

Finally, because there is no information in the POF dataset regarding prices, we utilize information about income to control this variable, since the ticket prices are determined by income and other variables. We consider the following individual attributes as the indirect price attributes: income per capita, gross monetary income, and employment status. In addition, attributes related to age, as well as to whether the consumers are students, can also be considered price indicators, since these factors affect ticket prices.

To control for the supply of movie theaters, since the number of theaters are very different among the Brazilian regions as a consequence of a very unequal regional wealthy pattern, we use the weight of each individual sample projection from POF as a measure of population agglomeration and higher supply of theaters. We also control for rural and urban areas, as there is no supply of theaters in the Brazilian rural areas. Furthermore, to control for higher preferences for cultural goods, we use the amount spent in cultural goods, excluding spent with films at theaters. Figure 2 shows the movie theaters distribution according to states. As it can be seen, the richest states, SP (São Paulo), RJ (Rio de Janeiro), MG (Minas Gerais), PR (Paraná) and RS (Rio Grande do Sul) have more movie theaters.

Figure 2 – Percentage of movie theaters by state in 2009



Source: ANCINE, the Brazilian movies Agency

We end up with 57 variables as indicated by consumer theory, and other available variables, as indicated by economic theory, at POF. The variables are described in the Appendix, Table 1. Note that skin color, water supply and state are variables transformed into categorical variables in a set of dummy (or binary) variables.

Table 1 – Variable classified and variables used to classify

Variables	N. Obs.	Mean	Std. Dev.	Min	Max	Main attributes order
		Outcome				
Movie at theaters consumption	344536	0.0194871	0.1382295	0	1	
Variables used to classify						
Year	344536	2005.077	2.999007	2002/2003	2008/2009	1
Expansion factor as proxy for population concentration	344536	957.0172	1264.463	3.5	21015.15	3
Urban area	344536	0.7614792	0.4261797	0	1	11
Number of household residents	344536	4.488231	2.096934	1	20	7
Number of bathrooms in household	344536	1.194723	0.7374931	0	13	9

Number of bedrooms in household	344536	2.231395	0.8908965	1	9	8
Number of rooms	344536	5.984449	2.248017	1	30	
Age in years	344536	29.47331	20.24922	0	110	2
Female	344536	0.4977303	0.4999956	0	1	
Years of schooling	344536	5.142313	4.48028	0	17	4
Student	344536	0.338432	0.4731769	0	1	6
Student at a private school	344536	0.0611982	0.239694	0	1	5
Student at a public school	344536	0.2772337	0.4476335	0	1	
Catholic religion	344536	0.7341961	0.4417609	0	1	
Credit card ownership	344536	0.1478859	0.3549873	0	1	
Private Health insurance	344532	0.1452521	0.3523554	0	1	
Income per capita in monetary values of 2002/2003	344536	441.7918	842.3349	0	62450.54	13
Family expenditures in culture goods except film at movies	344536	45.47498	164.2528	0	16800	10
Books consumption	344536	0.0375752	0.1901667	0	1	12
Internet access	344536	0.1454478	0.3525523	0	1	
DVD consumption	344536	0.3387542	0.4732868	0	1	14
TV cable consumption	344536	0.0286095	0.1667065	0	1	
Type of water supply (3 types)	344536	-	-	1	3	
Skin color/race (5 types)	349212	-	-	1	4	
States of household (27 dummies)	344536	-	-	11	53	

Source: POF 2002/2003 and 2008/2009

In order to identify relevant attributes for film consumption decision-making by the individuals, the computational selection involves restricting attributes in those pre-selected from consumer theory. According to Witten, Frank and Hall (2011), it is necessary to reduce the number of similar attributes to avoid irrelevant information and improve the classifier accuracy, that is, its ability to predict instances correctly. For this purpose, we apply the Linear Discriminant Analysis (LDA). Fisher (1936) argues that the LDA reduces the number of attributes preserving a minimum that best separate one class from another, eliminating unnecessary or redundant attributes.

The data-mining stage shows us a huge imbalance in the binary attribute classification between film consumers and the non-consumers as expected, because of Brazil's inequalities. We face a skewed data problem, where one class has a huge quantity of records as compared

to the other; while 342,328 (or 98%) records belong to the "No film" class; only the remaining 6,893 (or 2%) belongs to the "Film" class.

Bruzzone and Serpico (1997) and Akbani, Kwek and Japkowicz (2004) observes that the existence of skewed data tends to difficult the mining process, since the evaluation criterion considered in the learning phase can lead to ignore the instances in the minority class, handling them as noise or falsely detecting them therefore, a classifier built under such conditions (where one class outnumbers the instances of the other one) might lose its classification ability (Galar et al., 2013). To overcome this potential problem, we employ 1000 random samples for each class in the training stage and other 1000 random samples of each class for the test stage.

4. Results and discussion

Aiming both to reduce the computational effort in the SVM application and to minimize the interference from irrelevant parameters, the LDA has been applied, using 1000 examples randomly chosen from each class (cinema and not cinema consumption).

Accordingly, in the original 57-dimensions space, the LDA has identified the principal direction leading from one class to the other. The eigenvalue corresponding to the principal direction was orders of magnitude larger than the eigenvalue corresponding to the second principal direction, so only the principal direction has been considered in the analysis. This direction is described in the original space by means of a 57 components vector. But it has been observed that 14 of these components were much larger than the remaining ones, which are not statistically significant. This indicates that in the original space, the spatial variations from one class to the other occur mainly in those 14 directions and are almost orthogonal to the other directions, Table 1. Thus the classification problem might be carried out in a new 14 dimensions space instead of in the original 57 dimensions space, by preserving only those 14 corresponding parameters, or a subset of the whole sample. Hence, all the variables used in this

work have been transformed to that 14 dimensions space by simply ignoring the remaining parameters in each vector.

Thus, applying the LDA to the selected attributes from the dataset, Table 1, results in a rank where the fourteen most relevant attributes (Column of main attributes), according to the average merit and average rank metrics, are: 1) year of the sample showing that different samples in timing matter to determine patterns; 2) age of individuals measure in years, suggesting that going to the movies is connected to a specific range of age; 3) the sample expansion factor, that we used both as a proxy for the population concentration and for the supply of movie theaters, indicating both are important to determine cinema consumption; the individual degree of education evaluated by three variables, 4) years of schooling, 5) whether the individual is student at private school, and 6) whether an student in general; 7) number of people leaving in the household, that is connected to income per capita; 8) number of bedrooms in the household, and 9) number of bathrooms at the household that is also connected to income per capita; 10) the total expenditure on recreation and culture, our proxy for revealed preference for consuming cultural goods; 11) urban area; 12) book consumption; 13) income per capita; 14) and TV cable consumption. Summarizing, the main determinants are year sample, age, education, income, supply of movie theaters, and preferences for culture.

The SVM has been trained on the training set and executed on the test set for several kernel configurations. Independently of the kernel chosen, the libsvm C parameter is the regularization parameter defined by Cortes and Vapnik (1995). There is no special parameter for the linear kernel. The polynomial kernel has as parameter the polynomial degree. The RBF kernel has the gamma parameter, as presented in Chang and Lin (2011).

The results of the execution of libsvm on the test data are presented in Table 2, for several kernels and parameters values. One may notice that the best results for the accuracy are approximately 80%. For the RBF kernel, the grid tool present in the libsvm library has been

activated, in order to automatically search the best C and gamma parameters values. For the polynomial kernel, a degradation in the accuracy has been observed when the polynomial degree has been increased.

Table 2 - Accuracy in classifying the test data according to the kernel

kernel	gamma	C	degree	accuracy (%)
Linear	n.a.	1	n.a.	79.85
Linear	n.a.	64	n.a.	80
Polynomial	n.a.	1	2	79.5
Polynomial	n.a.	64	2	79.9
Polynomial	n.a.	512	2	80.35
Polynomial	n.a.	1	3	76.75
Polynomial	n.a.	1	4	66.4
RBF	0.125	512	n.a.	80.45
RBF	0.25	64	n.a.	80.45
RBF	0.5	32	n.a.	80

(n.a. = not applicable)

Thus, the observed variables are able to describe about 80% of film consumption decision, where the most relevant variables are described above. Note that, to find the main determinants by means of the LDA, means also eliminated determinants with similar information brought different variables.

5. Final Remarks

We perform an experimental study applying KDD process, which includes data- mining techniques performing Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) classification algorithm aiming to classify Brazilian individuals who consume film at theaters and that do not consume film at theaters, according to their observable characteristics. In the country, only 2% of the Brazilians at the sample go to the movies in 2002-2003 and 2008-2009, reflecting in part the huge inequality of the country, even if we consider that film can be consumed in other bundles of goods, as in Internet or home TV 3D for instance. The

algorithm was applied to a large individual micro-data level, available at 2002/2003 and 2008/2009, the Brazilian Consumer Expenditure Survey (POF).

We were able to classify the instances correctly in 80% of cases, which is a high degree of success since to predict film consumption depend on non-observable individual characteristics as individual past experiences or any individual innate characteristics or even an effect connected to his social network.

In addition, the early phase of reducing information, using LDA to assess whether a subset of the initial set of variables, is more informative for classifying individuals as consumer or non-consumer in comparison to the initial set of variables, suggested that the individual characteristics such as age, education (that merges to be a student), income, and preferences for cultural goods, and the geographic characteristics as the timing of sample, population concentration, and supply of movie theaters are the main determinants of film consumption in Brazil.

Finally, these results suggest that the Brazilian program, that enforces movie theaters to charge half-price of tickets for students, at the time of our sample, mattered but was not effective. This because be a student was a significant determinant, especially for the students of private schools, but since just 2% of the population at our sample went to the movies in 2002-2003 and 2008-2009, the policy was not fostering great access to this cultural good. Our results point to the need for new policy design in the subsidies towards wider population access to this good at that time. Brazil, however, is still under the same policy of tickets' half-price nowadays, thus it is important to replicate our methodology for the new POF survey data as soon as it became available in order to know the current performance of the policy.

References

Akbani R., Kwek S. and Japkowicz N. (2004). Applying Support Vector Machines to Imbalanced Datasets. In: Boulicaut J. F., Esposito F., Giannotti F., Pedreschi D. (eds) Machine

Learning: ECML 2004. ECML 2004. Lecture Notes in Computer Science, vol 3201. Springer, Berlin, Heidelberg.

Amo, S. and Rocha, A.R. (2003). Mining Sequential Patterns using Genetic Programming, International Conference on Artificial Intelligence, Las Vegas, USA, pp. 451-456.

Bruzzone, L. and Serpico, S.B. (1997). Classification of imbalanced remote-sensing data by neural networks. Pattern Recognition Letters, v.18(11-13), 1323–1328

Chang, C.-C. and Lin, C.-J., LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen, X., Chen, Y. and Weinberg, C.B. (2013). Learning about movies: the impact of movie release types on the Nationwide. Journal of Cultural Economics, v37, pp. 359-386.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. Machine Learning, v.20, pp. 273-297.

Diniz, S.C. and Machado, A. F. (2011). Analysis of the consumption of artistic-cultural goods and services in Brazil. Journal of Cultural Economics, v. 35, Issue 1, pp. 1-18.

Eaton, J. W.; Bateman, D.; Hauberg, S.; Wehbring, R. (2014). GNU Octave version 3.8.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. ISBN 1441413006, URL http://www.gnu.org/software/octave/doc/interpreter/

Galar, M and Fernandez, A; Barrenechea, B; Bustince, H; Herrera, F. (2013) A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C. v. 42, pp. 463-484.

Gallagher, C, Madden, M.G. and D'Arcy, B. (2015). A Bayesian Classification Approach to Improving Performance for a Real-World Sales Forecasting Application, 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Florida, USA.

Fayyad, U. and Shapiro, G.P. and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, AI Magazine, vol. 17, Issue 3.

Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." *Annals of eugenics* 7.2 (1936): 179-188.

Jehle, G. A. and Reny, P.J. (2000). Advanced Microeconomic Theory. 2nd Ed., Addison Wesley.

Kinto, E.A. (2011). Otimização e análise das máquinas de vetores de suporte aplicadas à classificação de documentos, Sao Paulo: University of Sao Paulo, Brazil, p. 145. PhD Dissertation.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. [S.I.], International Biometric Society 33(1), pp. 159-174.

McLachlan, Geoffrey. Discriminant analysis and statistical pattern recognition. Vol. 544. John Wiley & Sons, 2004.

Mitchell, T.M. (1997). Machine Learning. 1st Ed., [S.1]: McGraw-Hill Science/Engineering/Math.

Moretti, E. (2011). Social learning and peer effects in consumption: Evidence from movie sales, The Review of Economic Studies 78(1), pp. 356–393.

Rogati, M. and Yang, Y. (2002). High-performing feature selection for text classification, In: Proceedings of the Eleventh International Conference on Information and Knowledge Management. New York, NY, USA: ACM (CIKM ’02), pp. 659–661.

Russel, S. J. and Norvig, P. (1995). Artificial Intelligence - a Modern Approach. Prentice Hall.

Scott, A. J. (2017). Creative cities: The role of culture, Revue d'économie politique 120(1), (Janvier-Février 2010), pp. 181-204.

Segaram, T. (2007). "Advance classification: kernel methods and SVMs.” In: Programming Collective Intelligence: Build Smart Web 2.0 Applications, O'Reilly.

Siegel S. and Castellan N. (1988). Nonparametric Statistics for the Behavioral Sciences. 2nd Ed., New York: McGraw-Hill, USA, pp. 284-285.

Witten, I.H., Frank, E. and Hall, A. M. (2011). Data Mining Practical Machine Learning Tools and Techniques. 3rd Ed., Elsevier.