EasyChair Preprint

№ 1390

Social Activism Analysis: An Application of Machine Learning in the World Values Survey

Francielle M. Nascimento, Dante A. C. Barone and
Henrique Carlos de Castro

August 9, 2019

Social Activism Analysis: An Application of Machine Learning in the World Values Survey

Francielle M. Nascimento[1], Dante A. C. Barone[1], and Henrique Carlos de Castro[2]

[1] Institute of Informatics, UFRGS, Porto Alegre - RS, Brazil
{francielle.nascimento,barone}@inf.ufrgs.br
[2] Political Science, UFRGS, Porto Alegre - RS, Brazil
Principal Investigator of the WVS and National Director for Brazil
henrique@ufrgs.br

Abstract. The study of social sciences is essential to understand different dimensions of human society. Different researches are done to understand human development and its relationships in the community. Given this, in this paper, we have developed a methodology to use typical social science metrics and resources in conjunction with Artificial Intelligence techniques. The goal is to collaborate with research and visualize patterns to help explain human behavior. In this way, we use the World Values Survey's fifth wave data to apply self-learning methods and contribute to the advancement of social science research. We use algorithms to perform classifications, such as Random Forest, Stochastic Gradient Descent and Support Vector Machine in data collected in 58 countries, to verify if there are social patterns that can explain political participation. Thus, we identified that there is a stronger relationship in the results found in the so-called advanced democracies (USA and Europe) compared to those in other societies. From this, we can consider that eventual adjustments in the theory underlying the WVS research or in the instruments of data collection could be made and that more studies are needed to analyze other dimensions.

Keywords: Social Sciences · WVS · political participation · Artificial Intelligence.

1 Introduction

With the advent of globalization, world society is undergoing a process of internationalization and thought reformulation. This evolution has triggered a globalized economy with new standards, practices and cultures, and a worldwide concern for solutions to the problems faced in global proportions. [8]. In this way, mapping and understanding these problems has become increasingly necessary.

Thus, Inglehart [12] developed a method of application of questionnaires in several countries to understand the changes of values and their implications over time. It was called World Values Survey (WVS), forming, then, a database for

studies of social scientists. In this sense, the possibility of applying Artificial Intelligence techniques [9] can generate information and assist in the understanding of world society. Also, it can help to understand the course of value changes regarding economic, cultural and democratic development, according to Inglehart [12] (2003). Knowing this, the use of machine learning to obtain a system to mine data, extract information and standards, from WVS data is proposed.

In this work, we proposed to use machine learning techniques to complement social science studies in the WVS database. This paper is organized as follows. Section 2 discusses the purpose of the WVS for the study of social sciences. Section 3 defines the research pipeline, from the data preparation to the evaluation of machine learning models applied to WVS data. Section 4 shows the analysis of results and the last section contains the conclusions.

2 World Values Survey

Since the advent of globalization, world society has undergone several transformations in the process of internationalization and reformulation of thoughts and actions in a global order. This evolution has triggered the emergence of a globalized economy, new patterns, practices, cultures, political processes, social hierarchies, and global governance [8].

Concurrently, existing problems have taken on global proportions, with new formations such as inequality, hunger, disease, wars, and terrorism among others. Thus, it becomes necessary to understand and map these problems, to seek solutions at global levels to be applied.

As a result, social scientists began to study the change in values and the impact on social and political life from a World Values Survey Association (WVS)[4] initiative to verify the hypothesis that economic and technological changes transform societys values. In 1981, this research became well known. Its main researcher Ronald Inglehart [4], had conducted a survey in many different countries, which became an important instrument in forming the World Values Survey(WVS). In Europe, the data are collected in collaboration with the European Values Study (EVS).

The WVS is a survey conducted in more than 100 countries with complex questions that map out societal characteristics such as economic development, democratization, religion, gender equality, and others. In this way, a WVS database was built and made available by the organization with the aim of scientists developing studies on world values. This database currently has questionnaires and six wave responses (1981-1984, 1990-1994, 1995-1998, 1999-2004, 2005-2009, 2010-2014), and wave 7 is in progress.

3 Data Preparation and Model Definition

3.1 WVS Database

The WVS Database is composed of questions and answers to various aspects to map the behavioral change of society as a whole over the years. For each

application of the questionnaire, there are questions of the economic, social and cultural core, applied using probabilistic population sampling stratified in each of the participating countries. The average population sample of the participating countries is 1500 respondents to the questionnaire, depending on population size.

The Database consists of variables (corresponding to the questionnaire questions) and the value labels for each issue (which correspond to the answers). For each instance, we have a set of questions and their answers.

In this work, we have chosen Wave 5 for our studies. In Wave 5, there are 58 participant countries, and 258 questions [1]. The issues between 4 and 233 for building the models, were selected, because these questions belong to the central core and don't allow bias. Also, we chose them with the objective of understanding human behavior in regards to political activism.

To complete this study effectively, we decided to start our analysis at wave five, because it had more participation from countries of different continents enabling an overall summary. Contained within this study, were a total of fifty-eight participating countries from around the world and an extensive survey of over two hundred and fifty questions related to individual opinions on different topics. For example, the importance of friends, politics, and child obedience, or issues related to the fitness of politicians who don't believe in God to serve in public office. For this study, we decided to select a limited number of questions that don't correspond to demographic items, based on the expectancy of finding some relationship between values and human behavior. The questionnaire can be found at the website of WVS [1].To satisfy the needs of all participants involved in this study, WVS systematically translated each survey question according to their specific language (if applicable). Due to the intense nature of the study, specific guidelines and a code of ethics have been instilled so that the WVS survey teams reduce any bias and further limitations throughout the questioning process.

To ensure an accurate national sampling, WVS has relied upon the stratified sampling method because it allows dividing into groups mutually exclusive and frequent, allows discriminating different behaviors within the population. Thus, the sampling can reflect the style of the general population of different places, sexes, genders, and ages among other things. It is important to note that the WVS database is public with free access and available for researchers to carry out studies on this basis.

3.2 Data Processing

In this step, some methods were used to organize the data and build the architecture of the model. We re-defined some questions of WVS for binary classification. We have chosen issues related to our principal problem about social activism. Hence, we considered nine questions from WVS, regarding the active participation of the interviewee in voluntary organizations from a list (see Figure 1).The belonging to the class was considered true when at least one of the questions

[1] http://www.worldvaluessurvey.org/WVSDocumentationWV5.jsp

about active membership was answered positively, and false, when none of the questions was answered positively.

Now I am going to read off a list of voluntary organizations. For each one, could you tell me whether you are an active member, an inactive member or not a member of that type of organization? (*Read out and code one answer for each organization*):

		Active member	Inactive member	Don't belong
V24.	Church or religious organization	2	1	0
V25.	Sport or recreational organization	2	1	0
V26.	Art, music or educational organization	2	1	0
V27.	Labor Union	2	1	0
V28.	Political party	2	1	0
V29.	Environmental organization	2	1	0
V30.	Professional association	2	1	0
V31.	Humanitarian or charitable organization	2	1	0
V32.	Consumer organization	2	1	0
V33.	Any other (*write in*):_____	2	1	0

Fig. 1. Questions of Wave five correspondents the variables for defining the class, with the list of voluntary organizations [1].

In Table 1 we can observe the results of this criterion used to obtain the classes for the task. The characteristics available for each country show us that there are countries that have a balanced data set of a positive and negative target, as South Africa, Rwanda, and Brazil, while most of the other countries are unbalanced. This way, we expect that the countries that have few samples in one of the classes, it will to present a low performance in the metrics of evaluate.

Distribution by Class

Country	Positive	Negative	Total
Andorra	108	895	1003
Argentina	171	831	1002
Australia	226	1195	1421
Brazil	772	728	1500
Bulgaria	18	983	1001
Canada	615	1549	2164
Chile	225	775	1000
China	55	1936	1991
Taiwan	91	1136	1227
Colombia	741	2284	3025
Cyprus	66	984	1050
Ethiopia	447	1053	1500
Finland	180	834	1014
France	45	956	1001
Georgia	47	1453	1500
Germany	268	1796	2064
Ghana	1105	429	1534
Hungary	72	935	1007
India	438	1563	2001

Distribution by Class

Country	Positive	Negative	Total
Indonesia	758	1257	2015
Iran	532	2135	2667
Italy	92	920	1012
Japan	47	1049	1096
Jordan	37	1163	1200
South Korea	228	972	1200
Malaysia	187	1014	1201
Mali	561	973	1534
Mexico	639	921	1560
Moldova	135	911	1046
Morocco	17	1183	1200
Netherlands	147	903	1050
New Zealand	155	799	954
Norway	85	940	1025
Peru	378	1122	1500
Poland	125	875	1000
Romania	95	1681	1776
Russia	46	1987	2033
Rwanda	795	712	1507

Distribution by Class

Country	Positive	Negative	Total
Serbia	46	1174	1220
Vietnam	89	1406	1495
Slovenia	129	908	1037
South Africa	1561	1427	2988
Spain	109	1091	1200
Sweden	67	936	1003
Switzerland	243	998	1241
Thailand	297	1237	1534
Trinidad and Tobago	432	570	1002
Turkey	19	1327	1346
Ukraine	53	947	1000
Egypit	24	3027	3051
United Kingdom	189	852	1041
United States	466	783	1249
Burkina Faso	365	1169	1534
Uruguay	146	854	1000
Zambia	932	568	1500

Table 1. Distribution of the obtained classes by Country

For feature selection, we use the method called recursive feature elimination (RFE), where resources in each interaction according to coefficient obtained from the estimator weights to features, are removed [3]. This approach evaluates the performance of attributes set for making predictions, and how each features influences in a group of sets in the final model. Lastly, the results represent the best collection of features for the model. We expect that the set of selected features could explain some differences between the types of the countries, with respect to the culture, politics, and economy, which we plan to evaluate in our future work.

Also, we removed some variables and cleaned the data. We evaluate different numbers of variables using RFE the criterion to decide them to composing the model, and we choose the best according to tests. This way, for each country 40 variables, were selected. Furthermore, to prevent data leakage [5] the process of data normalization during cross-validation was deployed. Above all, it seems pertinent to remember that this method has been implemented to all the countries studied in Wave 5.

3.3 Models and Evaluation

With the data ready to be used, we elaborate some models of machine learning for building the classifications and evaluate results. Due to our previous knowledge in the area, we have considered that four models are enough for testing. We have chosen the following Machine Learning methods: Support Vector Machine (SVM), Random Forest Classifier (RFC), linear models with Stochastic Gradient Descent (SGD), and a neural network Multi-layer Perceptron (MLP). Significantly, these four models were applied to each country.

The validation of the model has been done using Stratified K-Folds cross-validator, in which ten different training and test sets were separated computing the evaluation metrics of each group to obtain the mean and the standard deviation (STD) of the results.

We also used measures to evaluate the predictive and classification models, from the F1-Score and Matthews correlation coefficient (MCC) [6]. The MCC is a measure of quality, which analyzes (binary) classifications even when there is an imbalance between occurrence and non-occurrence classes. The coefficient assumes values ranging from -1 and +1, where +1 coefficients relate to a perfect prediction, 0 random predictions and -1 imperfect prediction (total disagreement). This measure is relevant in this work since it makes a global analysis of the predictions and indicates the quality of the binary classifications in a context of the confusion matrix.

4 Analysis of Results

In the evaluation phase of the models, have built different perspectives for analyzing the results. For each country, we choose the model that presents the best F1-Score. In this research we have decided to calculate the metrics F1-Score and

MCC, where the F1-Score is the harmonic mean of precision and recall measures and the MCC assumes values that range between [-1 and +1], where +1 coefficients correspond to a perfect prediction, 0 to random prediction and -1 to imperfect prediction, respectively.

The TreeMap [10] in Figure 2 shows the distribution of countries according to F1-Score . The best result among countries can be seen at the extreme left, and the worst is at the extreme right of the map. We can observe that Australia got the 0.8811 F1-Score representing the best performance, and other countries like the United States, New Zealand, Canada, Netherlands, South Korea, United Kingdom, also presented satisfactory F1-Score [2] in the range of 0.80. The countries that offered the worst performance are Morocco, Turkey, Bulgaria, and Jordan, with F1-Score on average of 0.49.

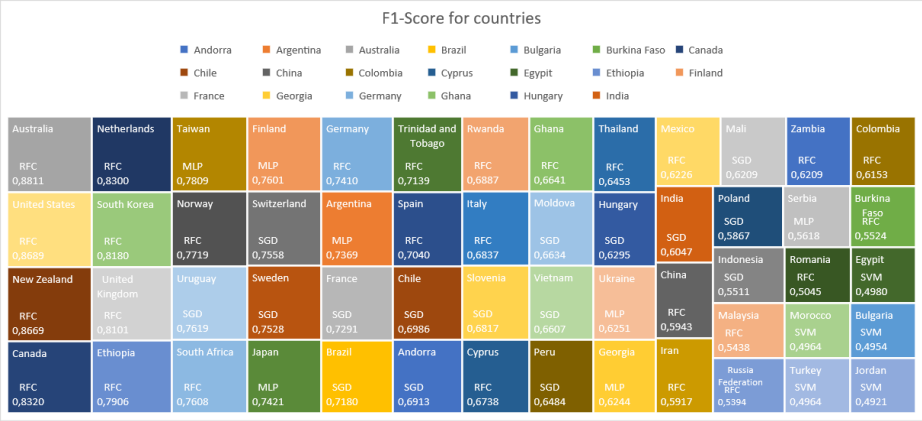

Fig. 2. Tree Map of F1-Score for countries

In this perspective, noting the distribution of TreeMap Figure 3 with hierarchy by continents, we can check that at Wave five, that Europe there is more representative, followed by Asia, Africa, South America, North America and finally Oceania.

We calculate the average and STD for continents, being that Oceania attains the best result with 0.874 of F1-Score and 0.0100 of STD. North America with 0.7594 and 0.1126, and, South America with 0.6965 and 0.0552. The worst performance is at Asia with 0.6265 and 0.1023, and Africa, with 0.6325 and 0.1054, F1-Score and STD, respectively.

The graph of Figure 4 shows the average of the F1-Score and the standard deviation by country. We can note that the vast majority of federations introduce an STD relatively small at the folds of cross-validation, varying between 0.0004 and 0.09. The region that shows the worst STD is Thailand, Japan, Hungary, Cyprus, Ukraine, Argentina, Sweden, with values between 0.10 and 0.15.

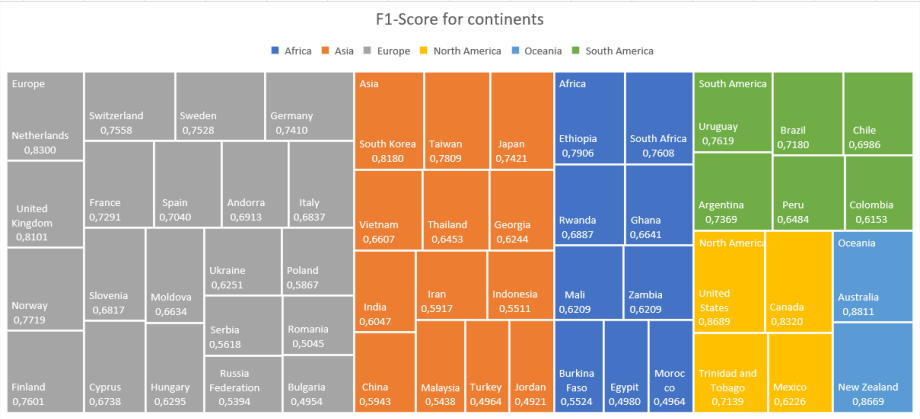

Fig. 3. TreeMap of F1-Score for continents

Analyzing the Pareto diagram in the Figure 5, we can see that 64% of the countries presented F1-Score between 0.5901 and 0.7861, which shows that the models were able to learn a satisfactory pattern during the classifications, in addition, 15% of the countries appeared with an excellent performance, with values varying between 0.7861 and 0.8841, and the remaining 21% with F1-Score of 0.4921 and 0.5901.

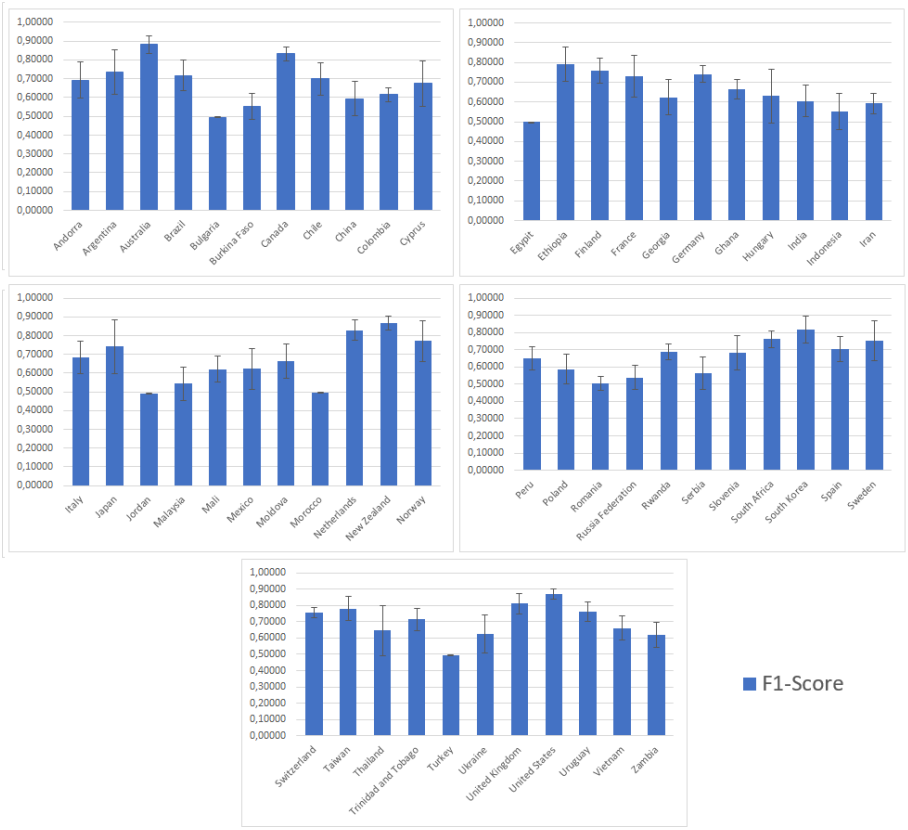

Fig. 4. Mean and Standard deviation by countries

Analyzing Figure 6 we can observe that the RFC Model showed greater variability, besides that according to superior limits the boxplot, most of the countries ranked with this model were better. And the worst model was SVM, with less variability and F1-Score of 0.49 on average. The SGD and MLP behaved results similarly, compared to the RFC. We can understand why the RFC shows the best performance because of the task uses the Wisdom of the Crowd, which significantly improves the results, making the model more accurate and reliable than the other [11]. Also, the data provide a framework that allows generalizations made by the RFC, because it is a survey structure.

Nyman & Ormerod applied machine learning models in research to predict economic recessions. The author used algorithms such as ordinary least squares regression and RFC, in which the latter presented better results. Thus, we can see that, for this task, the RFC is more appropriate and perform well in other problems like this available in the literature. [7]

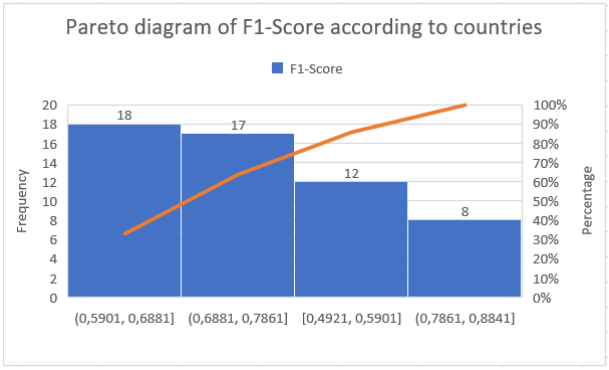

Fig. 5. Country distribution by F1-Score range

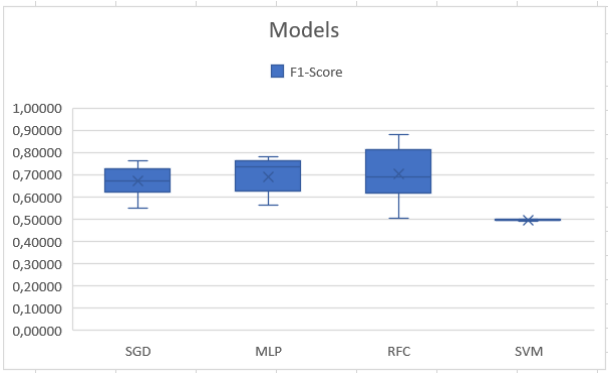

Fig. 6. Comparison between the models

The coefficient of Matthew's can demonstrate like the task of classification behaved in the general aspect at the models and like the predictions were satisfactory for each country. Figure 7 Shows this concept, and we can observe that some countries presented an MCC of zero, like, Morocco, Jordan, Egypt, Bulgaria, and Turkey, that is to say, the predictions were random, which allows us to verify that in these countries the model failed to identify a pattern about the social activism task,for not having expressive differences that allows to separate the classes. On the other hand, the other countries presented positive MCC, of these eighteen federations presented MCC with values above 0.50, and some of them very close to 0.80.

Above all, the aspects analyzed exhibit that the model's performance was satisfactory for this task, and we can perceive that there is a pattern in most countries that can explain why people choose social activism, according to results to F1-Score and MCC. This way, we observed that countries that have zero in MCC are some that have unbalanced data, but despite it, France has few samples in positive class and present a significant performance in F1-Score (0,7291), and

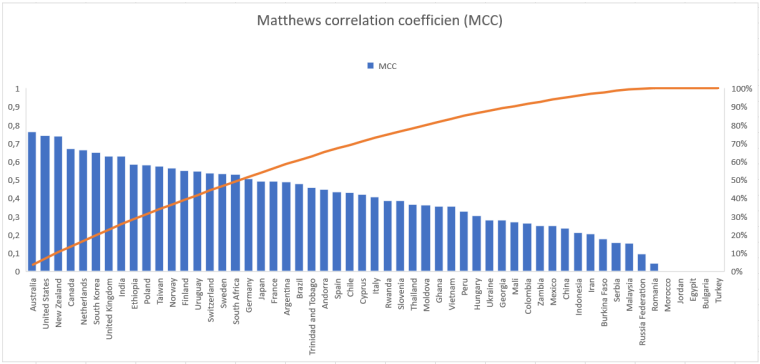

Fig. 7. Matthews correlation coefficient by countries

MCC (0,4939), Japan, Cyprus, Sweden, Norway, Vietnam are some examples with same behavior. Therefore, we can notice the models has clues that could differentiate and realize the classify, even with situations of unbalanced data.

5 Conclusions

The WVS research was built based on theories of the political and social behaviors found in the so-called advanced democracies (Europe and USA). Thus, the variables of the WVS dataset here analyzed seem to fit better to the values existing in those societies (as the results suggested). In societies with different social constructs and histories, the WVS research data may not reflect, at least in relation to the dimension analyzed (political participation), the behavior of the individuals. Given this, the present paper indicates, in a preliminary way, that it may be necessary to review the subjacent theory of certain dimensions of the WVS research, or the data collection instruments (questionnaire). Finally, new data treatments are needed to affirm whether this pattern is found in other dimensions of analysis or whether it is only a characteristic of the political participation dimension.

*Acknowledgements:*This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nıvel Superior - Brasil (CAPES) - Finance Code 001

References

1. Association, W.V.S., et al.: World values survey. Wave **5**, 2005–2008 (2005)
2. Frakes, W.B., Baeza-Yates, R.: Information retrieval: Data structures & algorithms, vol. 331. prentice Hall Englewood Cliffs, NJ (1992)
3. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1-3), 389–422 (2002)
4. Inglehart, Haerpfer, R.C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., Puranen, B.: World values survey: Round one - country-pooled datafile (2014), www.worldvaluessurvey.org/WVSDocumentationWV1.jsp.
5. Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD) **6**(4), 15 (2012)
6. Matthews, B.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure **405**(2), 442 – 451 (1975). https://doi.org/https://doi.org/10.1016/0005-2795(75)90109-9, http://www.sciencedirect.com/science/article/pii/0005279575901099
7. Nyman, R., Ormerod, P.: Predicting economic recessions using machine learning algorithms. arXiv preprint arXiv:1701.01428 (2017)
8. Robinson, W.I.: Theories of Globalization, chap. 6, pp. 125–143. Wiley-Blackwell (2008). https://doi.org/10.1002/9780470691939.ch6, https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470691939.ch6
9. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson Education, 2 edn. (2003)
10. Schreck, T., Keim, D., Mansmann, F.: Regular treemap layouts for visual analysis of hierarchical data. In: Proceedings of the 22Nd Spring Conference on Computer Graphics. pp. 183–190. SCCG '06, ACM, New York, NY, USA (2006). https://doi.org/10.1145/2602161.2602183, http://doi.acm.org/10.1145/2602161.2602183
11. Wang, L., Michoel, T.: Wisdom of the crowd from unsupervised dimension reduction. arXiv preprint arXiv:1711.11034 (2017)
12. Welzel, C., Inglehart, R., Kligemann, H.D.: The theory of human development: A cross-cultural analysis. European Journal of Political Research **42**(3), 341–379 (2003). https://doi.org/10.1111/1475-6765.00086, https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6765.00086