UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

FRANCIELLE MARQUES DO NASCIMENTO

Contributions of Machine Learning to knowledge acquisition in the field of Social Sciences

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof. Dr. Dante Augusto Couto Barone
Coadvisor: Prof. Dr. Henrique Carlos de Oliveira de Castro

Porto Alegre
January 2021

CIP — CATALOGING-IN-PUBLICATION

Marques do Nascimento, Francielle

Contributions of Machine Learning to knowledge acquisition in the field of Social Sciences / Francielle Marques do Nascimento. – Porto Alegre: PPGC da UFRGS, 2021.

85 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2021. Advisor: Dante Augusto Couto Barone; Coadvisor: Henrique Carlos de Oliveira de Castro.

1. World Values Survey. 2. Artificial Intelligence. 3. Social Sciences. 4. Political Participation. I. Barone, Dante Augusto Couto. II. de Castro, Henrique Carlos de Oliveira. III. Título.

“Doing your best is more important
than being the best.”
— Sir Zig Ziglar

ACKNOWLEDGMENT

À Deus que me iluminou nessa caminhada.

Ao Prof. Dante Barone e Prof. Henrique Carlos pela orientação, contribuindo para o meu crescimento pessoal e profissional.

Ao WVS Brasil pelo fornecimento da base de dados para essa pesquisa.

Ao grupo de pesquisa Pirandello e WVS Brasil por todo o suporte e amizade.

À Capes pelo financiamento do projeto.

À todos os professores que contribuiram para minha formação acadêmica.

Aos meus professores de Inglês Cathy Judge e George Lopez por todos os ensinamentos e suporte na minha caminhada.

Aos meus pais, Rosalina e Elizeu e as minhas irmãs Francislayne e Gleyciene, pelo amor e incentivo durante a minha jornada.

As minhas amigas, Mirele Carolina e Iasmyn Mazer por sempre estar ao meu lado me apoiando apesar da distância.

Aos meus companheiros de trabalho, Eduardo Cortes, Jaime Riascos, Vanessa Borba e Carolina Nery por todo o suporte e amizade.

Às minhas amigas Gabriela Cechinatto e Viviane Koch pela amizade e acolhida em Porto Alegre.

E a todos que contribuíram para minha formação.

ABSTRACT

The research in Social Sciences is fundamental to the study of human behavior. Beliefs and motivations play an important role in people's decision-making and choices. This relationship is relevant to explain the behavior in a population, and therefore, it allows for outlining social actions to improve the community. Knowing this, we proposed a way to discover meaningful patterns from a database of social studies using state-of-the-art techniques of Artificial Intelligence and Social Sciences. In this context, we selected Social Activism to perform classification using the extensive Word Values Survey (WVS) database. The algorithms applied in the task were Random Forest, Multilayer Perceptron , Stochastic Gradient Descent, and Support Vector Machine. Additionally, we use Recursive Feature Elimination to dimensionality reduction and analyze the selected features. The database used contain a survey applied in several countries, organized into Waves conducted in every five years. The Waves handled in this study were Wave 5 (2005-2009), Wave 6 (2010-2014), and Wave 7 (2018-2022). Thus, we discovered the patterns in the databases in the longitudinal view that make sense from the perspective of the Social Sciences. These patterns indicate that people are usually more concerned about moral-ethical issues than other aspects such as politics. This way, the results demonstrated that the use of the approach proposed contributed to discovering implicit knowledge in structured data.

Keywords: World Values Survey. Artificial Intelligence. Social Sciences. Political Participation.

Contribuições de Aprendizado de Máquina para a aquisição de conhecimento na área de Ciências Sociais

RESUMO

A pesquisa em Ciências Sociais é fundamental para o estudo do comportamento humano. Crenças e motivações desempenham um papel importante nas decisões e escolhas das pessoas. Essa relação é relevante para explicar o comportamento de uma população e, portanto, permite delinear ações sociais para a melhoria da comunidade. Sabendo disso, propusemos uma forma de descobrir padrões a partir de um banco de dados de estudos sociais usando técnicas de Inteligência Artificial e Ciências Sociais. Nesse contexto, selecionamos o Ativismo Social para realizar a classificação por meio do banco de dados Word Values Survey (WVS). Os algoritmos aplicados na tarefa foram Random Forest, Multilayer Perceptron, Stochastic Gradient Descent e Support Vector Machine. Além disso, usamos Recursive Feature Elimination para reduzir a dimensionalidade e analisar as features selecionadas. O dataset utilizado contém uma pesquisa aplicada em diversos países, organizada em Ondas realizadas a cada cinco anos. As ondas tratadas neste estudo foram Onda 5 (2005-2009), Onda 6 (2010-2014) e Onda 7 (2018-2022). Assim, descobrimos os padrões nas bases de dados na visão longitudinal que fazem sentido na perspectiva das Ciências Sociais. Esses padrões indicam que as pessoas geralmente estão mais engajadas com questões éticas morais do que com outros aspectos, como exemplo, aspectos políticos. Dessa forma, os resultados demonstraram que a utilização da abordagem proposta contribuiu para a descoberta do conhecimento implícito em dados estruturados.

Palavras-chave: Pesquisa Mundial de Valores, Inteligência Artificial, Ciências Sociais, Participação Política.

LIST OF ABBREVIATIONS AND ACRONYMS

AI *Artificial Intelligence*

WVS *Word Values Survey*

SVM *Support Vector Machine*

SGD *Stochastic Gradient Descent Classifier*

RFC *Random Forest Classifier*

MLP *Multilayer Perceptron*

ML *Machine Learning*

RFE *Recursive Feature Elimination*

LIST OF FIGURES

LIST OF TABLES

CONTENTS

–

1 INTRODUCTION

The increasing demand to deal with a large scale of data raises the need to process data to make it useful. In this way, information starts to support decision-making systems, expanding the potential of computational tools that offer different possibilities of solutions. Artificial Intelligence (AI) is the leading utility to study theses adaptive mechanisms using intelligent behavior to solve complex problems that traditional approaches do not have as much efficiency.

Thus, to search for solutions in finance, credit analysis, medical diagnostics and prognosis, disease detection, and the most varied needs for information and knowledge, computer systems and data science are fundamental for improving efficiency and effectiveness in diagnostics and decision-making. In this work, the study object is not a common area of using AI methods to understand these patterns. We are proposing the application of AI in the field of Social Sciences.

The study of Social Science is essential to map and understand problems, practices, cultures, searching for comprehensive solutions to the issues faced in global proportions. In this way, to find a path to look for standards not seen or evidenced by traditional methods it becomes necessary to explore other mechanisms.

Social scientists collect data in very different ways in order to understand and explain human society. One of such techniques is applying surveys, which that can be defined as a research method used for collecting data from samples of respondents in the population with the objective of obtaining information and insights on different topics of interest. However, most of the time, social scientists use traditional techniques, such as descriptive and inferential statistics, (BHATTACHERJEE, 2012), both to collect data and to analyze them.

In 1986, Inglehart (WELZEL; INGLEHART; KLIGEMANN, 2003) has organized an international survey in several countries with the aim of understanding the changes in values and their implications over time, called World Values Survey (WVS) (INGLEHART et al., 2014). Moreover, it can help to understand the values changes regarding economic, cultural and democratic development (WELZEL; INGLEHART; KLIGEMANN, 2003). The WVS consists of surveys conducted in almost 100 countries, comprising about 90 percent of the world's population, using a common questionnaire. The WVS is the largest academic research regarding different aspects and number of countries. To date, about 400,000 people have been interviewed, and its database is widely used by so-

cial scientists around the globe. The WVS is organized in waves conducted in every five years. This dataset can be understood as an example of a data source in which Artificial Intelligence techniques can be applied in order to generate new information and to help the understanding of human behavior and values.

The use of a structured database based on social theories allows the underlying theories to be tested or even improved. Thus, interdisciplinary research (AI together with Social Sciences) can imply the theoretical advancement of both fields of knowledge. Thereby, the objective of this dissertation is to use Machine Learning tools in order to obtain a system to mine data, extract information and standards, from WVS database.

Postmaterialism concepts defined by Inglehart show that society is looking for activities that can contribute to the improvement of society. One of these movements is Social Activism, evidenced by participation in voluntary projects. This research chooses to study the human values around Social Activism to complement Inglehart studies (WELZEL; INGLEHART; KLIGEMANN, 2003).

Specifically, we propose the use of Machine Learning techniques to complement social science studies in the WVS database, focusing on understanding the relationships between the selected features in the model construction process. Using Social Activism as a target, we developed this work based on the relationships between Social Activism and human values around this topic. The analyzed period comprises the last three waves of the WVS, Wave 5, Wave 6 and Wave 7.

This work's general objective is to apply machine learning techniques in a structured database in Social Science research to understand social activism. This way, the contributions resulting from this research are listed below:

- A survey of research carried out using Artificial Intelligence techniques on an unstructured basis for society's study.
- A survey of research carried out in the WVS database by Social Scientists.
- A survey of research that used the base of WVS and Artificial Intelligence.
- The development of a methodology for a hybrid approach to analyzing Artificial Intelligence applications in a structured database in the Social Science field.
- Analysis of features selected by RFE from the perspective of social sciences.
- An application of the methodology, with four different machine learning models, SGD, RFC, SVM, and MLP.

In addition, we conclude these studies showing the performance of four models for

the task. First, we discuss the background, showing the purpose of the WVS for the study of Social Science, Social Activism. Also, the concepts of Machine Learning applied in this work.Next, following with related works and then we define the research line, from the preparation of the data to the evaluation of the Machine Learning models applied to the WVS data.Finally, the conclusions.

2 BACKGROUND

This chapter will address concepts and methods that are essential for understanding this research. We divided into three subsections: Social Sciences: Theory and Methodology, Social Activism and Machine Learning.

2.1 Social Sciences: Theory and Methodology

Social Sciences involves disciplines that study human behavior and society. It focuses on people - individuals and groups, and their relationships with the world - economics, politics, culture, well-being, and safety, for example.

To acquire this knowledge or data, social scientists use qualitative or quantitative methods. Qualitative methods aim to discover this information by collecting data, for instance, from focus groups or in depth interviewers. The main objective is to understand the motivations of certain behaviors. By contrast, quantitative methods aim for causal explanations using quantifiable data to conduct statistical analyses to find correlation or causation with certain variables (PORTA; KEATING, 2008).

Also, we can investigate the data collected by a quantitative or qualitative method using positive analysis or normative analysis. The positive analysis consists of factual statements, known as positive statements, where it is possible to find testable conclusions (DAVIS, 1998). The normative analysis consists of normative statements where we use factual evidence as support, but include opinions, underlying values, and moral judgments (LIPSEY, 1968).

In fact, the model of the normative social sciences is an assumption of our society. If we act following the normative model, we can expect our behavior rules reflect our society. In contrast, positive analysis sustain the empirical model. The empirical model explains our civilization with clarity using verifiable evidence of all insights and understanding (SATO, 2014). For instance, laws observed in our society have high acceptance because procedures are carried out in a constitutionally competent body to represent citizens' will.

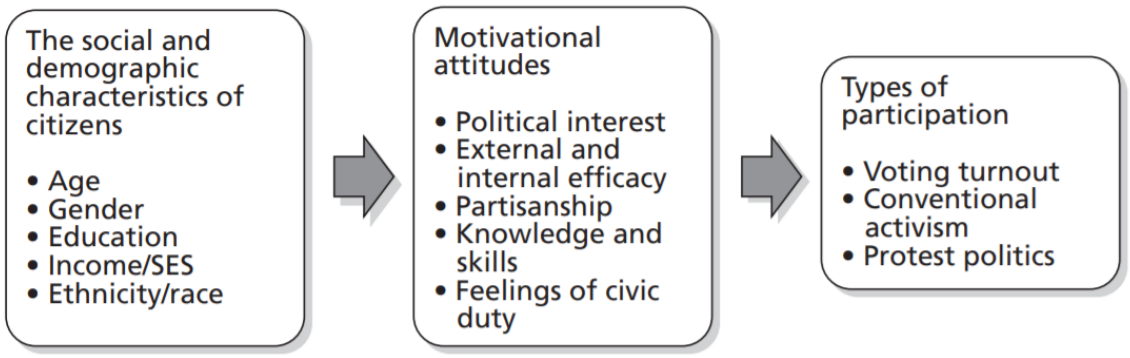
Using different methodologies and techniques, collecting and analyzing data, the purpose of Social Sciences research is to provide comprehension of shared values, improving social awareness in a longitudinal view (REALE et al., 2017). Including resources to understand social phenomena to propose new solutions in order to understand-

ing or explaining the society.

2.2 Social Activism

In different aspects, social activism can be manifested in the sphere of society. Historically, voting participation illustrates the most common way of revealing political activism expression. Afterward, we can see civic activism that remains voluntary organizations, community associations, besides new social movements. Finally, protest activism that can be massive protests organized to demonstrate dissatisfaction around events (INGLEHART et al., 2003).

Figure 2.1: The standard theory of political activism by(JONES, 2011).



In Figure 2.1, we can see these relationships in society and the theory patterns of social activism. In this context, social scientists are responsible for studying and understanding these dimensions.

Kende explains that activism is regarding change or preserving the social order being contrary most of the moment to political or economic issues. After all, the importance of diversity of the topics that demand understanding activism in a different context, for example, social, economic, and political matters. Including methodological plurality allows analyzing different domains of political activism in Social Sciences (KENDE, 2016).

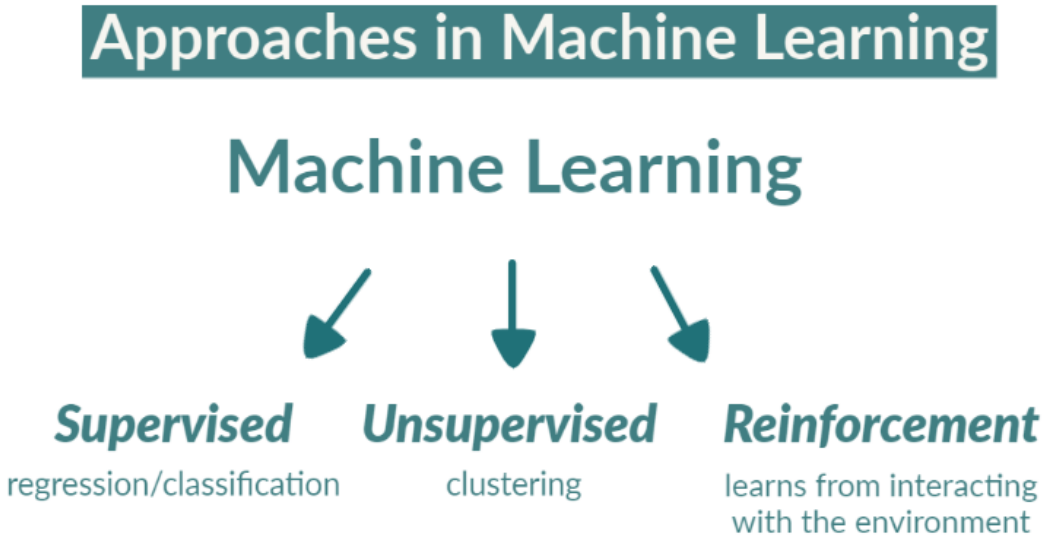
Understanding how people engage in activism and the values related to this is an ongoing interest in social science literature. This research proposal brings new ways to discover relationships using Artificial Intelligence and Social Sciences in the search for these associations.

2.3 Machine Learning

Machine Learning (ML) refers to computational resources to develop knowledge of an automated way. The main goal is to deal with large datasets aiming to earn it useful as information (SHALEV-SHWARTZ; BEN-DAVID, 2014).

We use ML systems to recognize faces, predict diseases, financial analysis in the real world, and the most varied needs of information and knowledge. To acquire this learning, we have three approaches to train algorithms in Figure 2.2.

Figure 2.2: Types of machine learning. Source: The author



In supervised learning, we have a collection of examples with labels that a supervisor informs. This way, the algorithm learns to find the data standards according to these labels, seeking to minimize the errors of predictions.

In unsupervised learning, we do not have the figure of a supervisor or labels. It means that the algorithm needs to find similarities among the set of data to reveal associations and anomalies.

In reinforcement learning different from the other approaches needs to learn to interact with the environment learning based on trial and error.

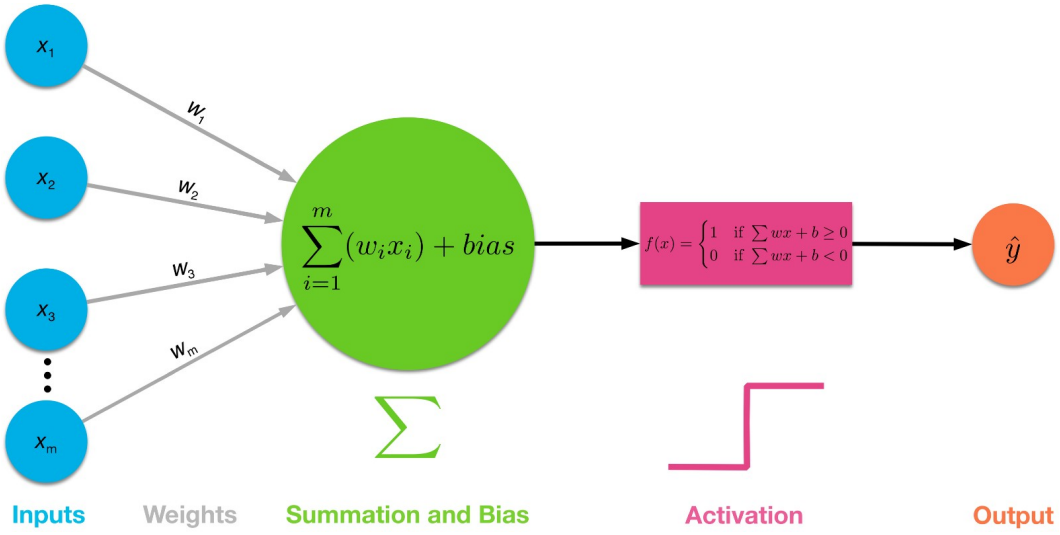
The following sections show machine learning models that use the supervised approach to acquire knowledge.

2.3.1 Multi-layer Perceptron Classifier

Multilayer perceptron (MLP) is a machine learning technique that is inspired by the functioning of neurons of the human brain. An artificial neuron receives signals of the environment or other neurons that associate those signals and put forward to other artificial neurons. Mathematically the structure of MLP is capable of mapping complex functions seeking a useful approximation algorithm.

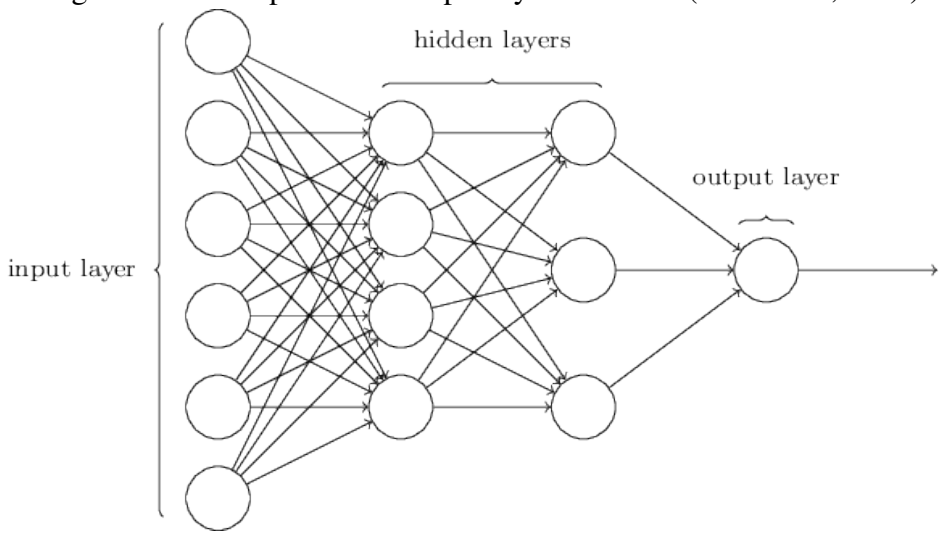
Figure 2.3 shows the mathematical model of a Single Layer Perceptron. A weight x_i is the input of a synapse i connected by a neuron j, where the links characterize synapses, multiplied by a synaptic weight $w_{i,j}$. There is also an adductor to add the input signals. They are weighted by the neuron's synapses, which results in a linear combination. The activation function aims to limit the amplitude of a neuron's output to some finite value and introduces non-linearity in the output. (HAYKIN, 2004).

Figure 2.3: Procedures of a Single-layer Perceptron Network (KANG, 2017)



Multi-layer Perceptron is a network of layers of artificial neurons. This typical neural network consists of an input layer, hidden layers, and an output layer represented in Figure 2.4. The learning capacity of a model is in adjustments applied iteratively to synaptic weights and biases. The internal parameters of a neural network model play a significant role in obtaining optimal and fast results. For this reason, the model use strategies to calculate and optimize these values through algorithms. Optimization algorithms help to minimize (or maximize) an objective function. Weights and bias are parameters that can be updated to minimize the loss in the training process through optimizers.

Figure 2.4: Example of a Multiple layer networks (NIELSEN, 2015)



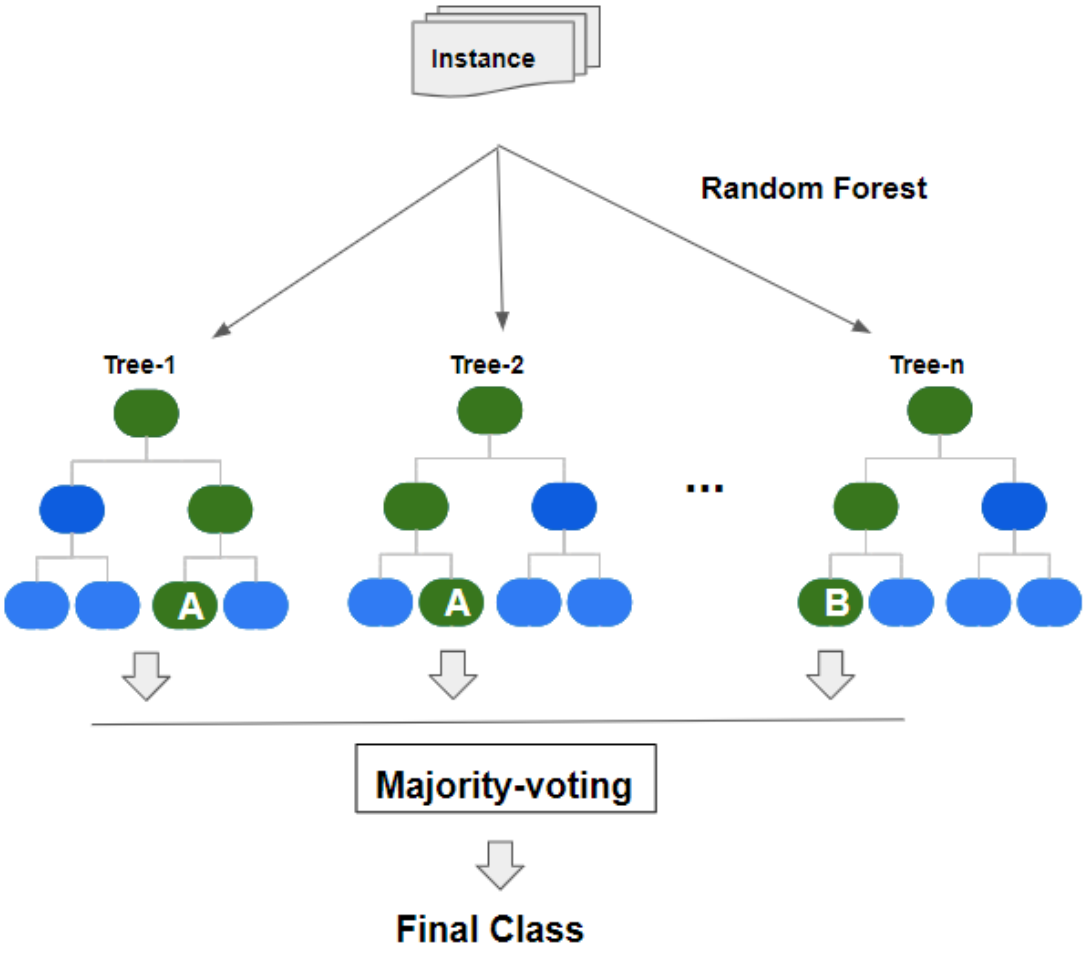
2.3.2 Random Forest Classifier

The Random Forest Classifier is a model containing a large number of individual decision trees that consists of combining those decisions to decide the final classifier with majority-vote predictions.

The process of constructing the decision tree follows a top-down approach using statistical tests to determine if the attribute is the most relevant to be the root node of the tree. After the decision, a descendant of the root node is created for each attribute's possible value. This process is performed recursively until the tree is fully defined (MITCHELL, 1997).

A representation of a random forest is shows in Figure 2.5.

Figure 2.5: Example of Random Forest Classifier.Source: The author

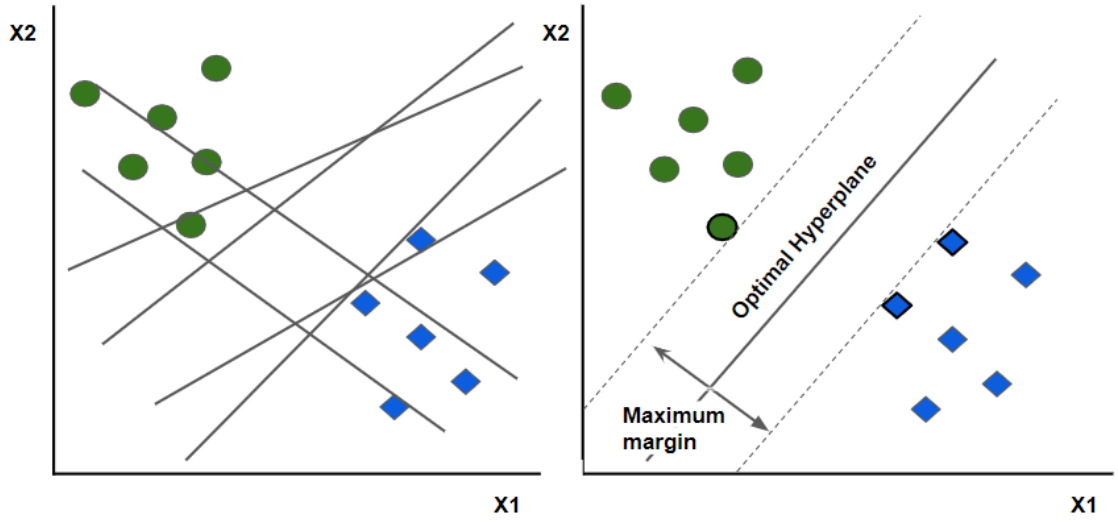


2.3.3 Support Vector Machine

Support Vector Machine (SVM) algorithm has the goal to find an optimal hyperplane in an N-dimensional space capable of categorizing new examples. The method combines the principle of generalization in dimensional spaces, controlling the margin where the separation is ample for both classes (VAPNIK, 1998).

On the left side of Figure 2.6, we can see different hyperplanes draw in the space, but they do not have the optimal margin or separate these points— the SVM search for a hyperplane that maximizes the margin between the classes. On the right side of the Cartesian plane, we can see the hyperplane that maximizes the margin resulting in properly fits the training data.

Figure 2.6: Example of Support Vector Machine.Source: The author



2.3.4 Stochastic Gradient Descent Classifier

Stochastic gradient descent is a linear classifier which implements regularized linear models with stochastic gradient descent (SGD) learning. This model supports different loss functions and penalties for classification (PEDREGOSA et al., 2011).

The stochastic gradient methods minimize the loss of a model computing the gradient of a loss function on a single or batch training examples (HARDT; RECHT; SINGER, 2016).

2.3.5 Recursive Feature Elimination

In large datasets is expected the presence of several features distributed in a high-dimensional space. These features often can be unnecessary or not correlated with the goal — for example, features that can overfit or provide noise in the data set. An ideal solution to the problem is reducing dimensionality, that is, to transform data from a high-dimensional space into a low-dimensional space. Therefore the transformation can retain meaningful characteristics to the task.

Recursive Feature Elimination (RFE) is an algorithm capable to reduce dimensionality by selecting features. The main goal of the RFE is to recursively train a model and compute the importance for all features to find one of the best feature groups for the task. The feature that has the smallest ranking criterion is removed from the set. This

process is recursively repeated until achieving the number of features desired.

2.4 Final Considerations

This chapter presents Social Sciences concepts and the methods applied in social scientists, such as positive analyzes and normative analyzes. We used the normative analysis to interpret the results obtained from the feature selection algorithm, and the positive analysis for comparison with statistical results.

In addition, we discussed social activism, the object of study of this work, and the characteristics of social activists and the movements around different spheres of our society.

Finally, we present the machine learning models chosen for this work. The criteria for choosing the models were state-of-the-art classics that use different training and test paths to define the starting point. In order to contribute in the social and computational perspective with the approach developed in this dissertation.

3 RELATED WORK

This chapter will address the related works that were the basis for the development of this research. We divided this section into three research categories: Artificial Intelligence and Social Sciences, World Values Survey and Social Sciences, World Values Survey, and Artificial Intelligence. The first presents researches that apply Artificial Intelligence techniques in Social Sciences. The second presents research carried out from the WVS database. The third presents research that applied AI to the WVS database.

3.1 Artificial Intelligence and Social Sciences

Many Social scientists collect data in very different ways in order to understand and explain human society. However, most of the time, social scientists use traditional techniques both to collect data and to analyze them. In this thesis, the focus is to think about alternative uses of both data collection and analysis, especially with the use of AI.

Lohuizen (LOHUIZEN; SAMOHYL, 2011) introduced a study about the use of automated telephone interviewing techniques (robo-polls) to discover Presidents' job approvals and compares with the Internet, and live operator methods. The authors used a time series regression model to control and analyze the data collected from the interviewed and find that robo-polls method has a lower rate for undecided people.

In another perspective, Kulesza et al. present a study using a model of Naive Bayes to text classification to increase the performance in intelligent assistants. The idea is to reduce the incorrect behavior made by the predictor, using techniques to discover why they made such a choice. Knowing why an assistant produces a certain behavior, the user, has the possibility to answer with the objective of correct action predictors. The authors consider end-user debugging in intelligent assistants is essential for advance and adaptation in this tool based on the experience and knowledge provided by users (KULESZA et al., 2011).

Robila examined applications of the field in Social Sciences, including human behavior, focused on studies on children and family. The research covered three main areas of the use of AI, firstly related to the effectiveness of diagnosis plus prediction of different conditions (for example, the influence of parenting norms), secondly data management in social and human service. Thirdly and most related to this work is an understanding of human development and functioning. The key principle most related to

this is the understanding of human development and functioning. Regarding the last point, the author analyzes the use of AI to learn behaviors regarding the application of techniques to identify themes in online postings around family topics (ROBILA; ROBILA, 2019).

Miller examines the field of Explainable Artificial Intelligence (Došilović; Brčić; Hlupić, 2018). The author discusses the importance of inclusion of other specialists in different fields like Philosophy, Psychology, and Social Science in the body of research into explainable AI. The author focused on explaining the process decisions among humans, considering all the modes, from the questions to the answers and evaluation. Miller emphasizes methods already used by scientists as a complement and closure of the explainable AI cycle. Also, the author highlighted the relevant insights from Social Science research on human behavior, especially to the application into this topic (MILLER, 2017).

Khartica et.al analyzed the behavior of terrorist profiles of the 9/11 attack. The authors used an unsupervised approach to judge the involvement of a person in a specific activity. The object of study was a network structure to identify the key player that influences other peoples automatically. Khartica represented the network structure among these relations in the scheme of a directional semantic graph. As a result, the authors build an unsupervised called SoNMine system using selection strategies to determine the node's importance in a profile based on the contribution amount (KARTHIKA; KIRUTHIGA; BOSE, 2012).

Scientists commonly carry out analysis of behaviors and attitudes using the social network, such as unstructured data from twitter. McCormick et al. (2017) developed an approach to using demographic data from twitter, such as age, race, gender. Increment textual data from Twitter with additional information can open new paths to Social Scientists, analyze behaviors, investigate social problems and events. The authors propose a data-processing infrastructure allowing to gather demographic information not available within Twitter users. The evaluations of Twitter user profile pictures to predict demographic information using tweets to examine intention not to vote the 2012 U.S. presidential election (MCCORMICK et al., 2017).

3.2 World Values Survey and Social Sciences

The World Values Survey addresses many areas of our society. According to the WVS Survey findings, Inglehart highlights human empowerment advances in three levels, socio-economic, socio-cultural, and legal-institutional levels. First, socio-economic

empowerment improves with action resources to increase people's capabilities to exercise freedoms on the socio-economic level. Second, socio-cultural emancipative values improve aspirations to exercise autonomy. Finally, at the juridical-institutional level, with increasing democratic rights, people's right to exercise freedom increases (WELZEL; INGLEHART; KLIGEMANN, 2003).

The study of human values broad into two dimensions. On the first dimension, religiosity, national pride, respect for authority, obedience, and marriage highlight traditional values. On the opposite side, secular-rational values do not emphasize religion or traditional family, while divorce, abortion is relatively tolerable.

On the second dimension, survival values oppose self-expression values. The survival values include security, lack of acceptance of homosexuality. In contrast, self-expression values prioritize environmental protection, gender equality, growing tolerance of foreigners, and rising demands for participation in decision-making in an economic and political level (INGLEHART; WELZEL, 2005).

An analysis by WVS Brazil members studied the changes in political participation in Brazil in the 21st Century. Castro and Reis (2012) evaluated data referred to as two survey applications, 1990 and 2006. The authors seek the importance of democracy, belonging, participation in social institutions, and citizen confidence in Brazil's institutions. The main findings have shown citizens' tendency to participate more in institutions not linked to political society, such as social institutions as family and religion. The authors found that the growth of individuals who belong to and participate in institutions unrelated to the state apparatus is superior to those related to formal institutions (CASTRO; REIS, 2012).

Castro, Ranincheski, and Capistrano (2015) compared political values and points of view of Latin Americans on globalization issues, from a longitudinal perspective. The authors analyzed the period from 1990 to 2014, encompassing six Waves. The countries analyzed (Brazil, Argentina, Chile, Mexico, Peru, Uruguay, Argentina, Ecuador) highlight that local and national sentiment is still strong. Also, the individual values material survival, such as employment, and values of defense of the environment are considered essential for Latin Americans (CASTRO; RANINCHESKI; CAPISTRANO, 2015).

3.3 World Values Survey and Artificial Intelligence

Li (LI et al., 2018) used the database of WVS to study the distributions of the values regarding the environmental protection values in different cultural districts. The authors analyzed data from 1994 to 2014 using variables in this domain.The authors divided the work into two main tasks; one subjectively classifies the variables in action values and attitude values and then applied hierarchical clustering analysis to classify the variables and confirm the results previously. The main findings show that the differences between the countries are based not only on objective factors (for example, economic development) but also, cultural and religious factors. In perspective, to understand these aspects, this work shows the necessity to know the aspects of the actions and attitude based on data to apply interventions to improve the system of environmental protection.

Nelson et al. implemented decision trees in the Wave 6 (2010– 2014) of the World Value Survey in four specific nationalities, Germany, India, Morocco, and the U.S. The authors searched for individual's values and their relationship with trust in other people. The question chosen was "people can be trusted" as a target variable. The root node in each country shows particularities. In the U.S., the question overall secular values-4 evidence that religious (traditional) tend to devalue the secular. In Morocco, the root of the tree that concerns respect for individual human rights, a value to trust is likely high. India captures xenophobia elements or reflects a high-value on traditionalism because the root of the tree concerns immigrants' respect, and those who disagree are trusting. Finally, Germany's root node was about satisfaction with life. In general, the more satisfaction, the higher propensity to trust (NELSON; KENNEDY; KRUEGER, 2016).

Tiago Vier proposed integrating artificial intelligence (AI) in the Social Sciences, applying machine learning techniques. The strategy applied was comparing traditional data analysis results using machine learning from seven waves of WVS. A traditional theme from political sciences to evaluate an inductive and deductive moment was chosen, in this case, Patriotism. In the deductive moment, the traditional technique selected to test was binary logistic regression that serves to study associations between a response of this type and other explanatory variables. The author used hypothetical-deductive logic to explain why a particular individual or a society is more or less patriotic. In the inductive moment, the authors defined a binary classification problem to construct the machine learning model, using the Patriotism variables as a target. The author evidenced that the conclusions drawn based on inductive research are different from research that takes the

hypothetical-deductive path. The relations found were among traditional values, religious family, and beliefs in social organization forms, which are frequently sustained by authoritarian and hierarchical institutions (VIER, 2020).

Although the use of Artificial Intelligence in Social Sciences is becoming more common, especially with unstructured data sets, such as social networks (Facebook, Etc.), as presented here, few studies are applying Artificial Intelligence to structured social science databases. That is that it has been built based on some social theory. We understand that using a structured database based on social theories allows the underlying theories to be tested or improved.

In this thesis, we applied the concepts around all of these works. From the WVS Survey, we analyzed classifications by different models of Artificial Intelligence, and use techniques of Social Science to understand the features selected of the model and the results generated.

4 PROPOSED WORK AND METHODOLOGY

We developed in this work studies of machine learning methods aiming the development of systems to discover patterns in a typical database on Social Sciences. Therefore, the research was carried out in stages, namely: an overview of the process, specify of data processing, models, and evaluation applied. The following sections detail the steps.

4.1 Word Values Survey

The WVS database is composed of a set of variables related to different aspects of the human values aiming to understand their change over time. In each wave a standard set of questions concerning economic, social, and cultural values is present. The WVS researchers collect data with stratified probabilistic sampling in all participating countries. The average population sample is 1500 respondents to the questionnaire, depending on the population size.

The Database consists of variables (corresponding to questions) and value labels for each issue (which correspond to answers) — also, a set of questions and their answers for each instance.

WVS systematically translated each survey question according to its specific language to satisfy the needs of all participants involved in this study (if applicable).The WVS survey teams have instilled specific guidelines and a code of ethics to 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, and ages, among other aspects. It is important to note that the WVS database is public with free access and available for researchers to carry out studies on this base.

The surveys started in 1981, and six waves have already finished, as shown in the table below. The 7th wave [1] of the World Values Survey is in progress.

In this work, we have chosen Wave 5,(2005-2009), Wave 6 (2010-2014), and Wave 7 (2018-2021) for our studies, because these are consistent and similar among them. For each Wave, we selected questions that belong to the central core for forming the models

[1] WVS WAVE 7 http://www.worldvaluessurvey.org/WVSContents.jsp

Table 4.1: Description of period of each Wave.

Wave	Period
Wave 1	1981-1984
Wave 2	1990-1994
Wave 3	1995-1998
Wave 4	1999-2004
Wave 5	2005-2009
Wave 6	2010-2014
Wave 7	2017-2021

and issues that do not allow bias according to the task. The purpose of this work is understanding human behavior in regards to political activism.

For this study, we analyze an extensive survey with more than two hundred questions related to individual opinions on different topics, and it was applied in several countries. The importance of friends, politics, and child obedience, or issues related to the fitness of politicians who do not believe in God to serve in public office are examples of the questions applied in the survey.

Finally, we decided to select a limited number of questions that do not correspond to demographic items, based on the expectancy of finding some relationship between values and human behavior. On the website of WVS [1], we can see the questionnaire and the list in Table 4.2 of information and numbers of questions over each wave.

Table 4.2: Description of data and questions of each Wave.

Wave	Number of Participant Countries	Number of Questions	Questions Selected
Wave 5	58	259	4 - 239
Wave 6	60	258	4 - 239
Wave 7	35	290	1 - 259

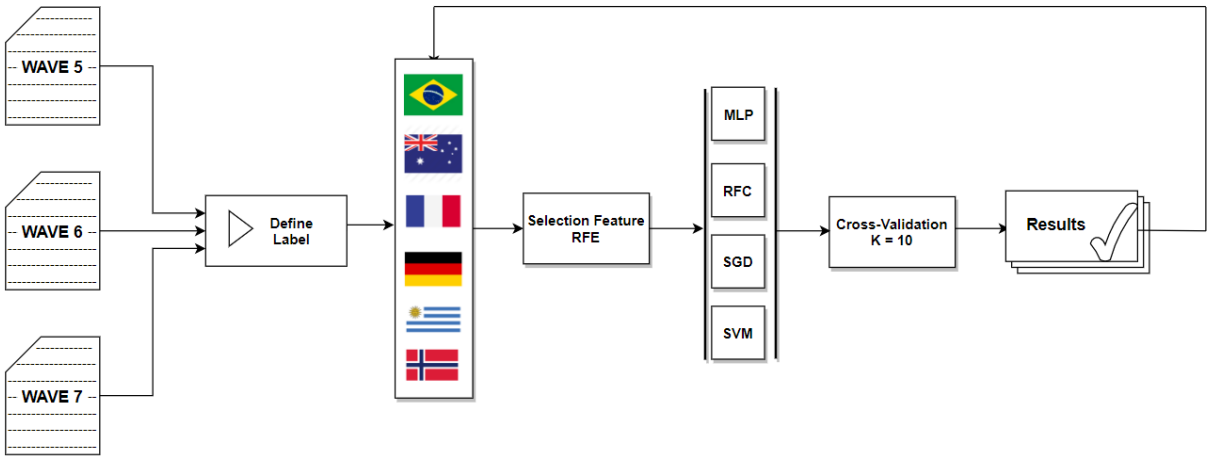
4.2 Overview of the Methodology

We defined an approach in this work, based on the idea that different nationalities may introduce distinct values and behaviors regarding the culture, politics, and process of socialization. In this way, we decided to analyze separately each country contained in each of the waves.

The process of analyzing and building the models followed these steps shown in Figure 4.1. First, for each wave, it was defined as the label of the class; in other words, a set of questions corresponding to the task of classifying and analysis a concept. Then,

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

Figure 4.1: Overview of the methodology for analyzing the data



we applied for each country the process of selection of features, reducing dimensionality. Next, using these selected features, we applied this dataset in four methods of Machine Learning, they are, Support Vector Machine (SVM)(SMOLA; SCHöLKOPF, 2004), Random Forest Classifier (RFC) (BREIMAN, 2001), linear models with Stochastic Gradient Descent (SGD) (ZHANG, 2004), and a neural network Multi-layer Perceptron (MLP)(RUMELHART; HINTON; WILLIAMS, 1986). Finally, we generated the results with the mean of k-fold cross-validation with K=10. We discuss and detail in the following sections these steps.

4.3 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 (corresponding to social activist or non-social activist). We have chosen issues related to our problem of social activism and we built two tasks for these that we will discuss in the next chapter.

For feature selection and dimensionality reduction, 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 (GUYON et al., 2002). 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, concerning

the culture, politics, and economy, which we will evaluate.

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, forty variables, were selected. Furthermore, to prevent data leakage (KAUFMAN et al., 2012), the process of data normalization during cross-validation was deployed. Above all, it seems pertinent to remember that this methodology was replicated to all the countries studied at the Waves.

For this task we considered nine questions from WVS, regarding the active participation of the interviewee in voluntary organizations from a list (see Figure 4.2).The belonging to the class was considered true when at least one of the questions about active membership was answered positively, and false, when none of the questions was answered positively.

Figure 4.2: Questions of Waves correspondents the variables for defining the class, with the list of voluntary organizations (ASSOCIATION et al., 2005).

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

In the Tables 4.3, 4.4, 4.5 we can observe the results of this criterion used to obtain the classes for the task in each Wave evaluated. The characteristics available for each country show us that there are countries that have a balanced data set of a positive and negative target, 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 evaluation for each Wave. Also, some countries do not was possible to build the model, because no occurrence of answers in both classes, and it is why there are not all the countries of each Wave shown here.

Table 4.3: Wave 5: Distribution of the obtained classes by Country

Country	Positive	Negative	Total	Country	Positive	Negative	Total
Andorra	108	895	1003	Moldova	135	911	1046
Argentina	171	831	1002	Morocco	17	1183	1200
Australia	226	1195	1421	Netherlands	147	903	1050
Brazil	772	728	1500	New Zealand	155	799	954
Bulgaria	18	983	1001	Norway	85	940	1025
Canada	615	1549	2164	Peru	378	1122	1500
Chile	225	775	1000	Poland	125	875	1000
China	55	1936	1991	Romania	95	1681	1776
Taiwan	91	1136	1227	Russia	46	1987	2033
Colombia	741	2284	3025	Rwanda	795	712	1507
Cyprus	66	984	1050	Serbia	46	1174	1220
Ethiopia	447	1053	1500	Vietnam	89	1406	1495
Finland	180	834	1014	Slovenia	129	908	1037
France	45	956	1001	South Africa	1561	1427	2988
Georgia	47	1453	1500	Spain	109	1091	1200
Germany	268	1796	2064	Sweden	67	936	1003
Ghana	1105	429	1534	Switzerland	243	998	1241
Hungary	72	935	1007	Thailand	297	1237	1534
India	438	1563	2001	Trinidad and Tobago	432	570	1002
Indonesia	758	1257	2015	Turkey	19	1327	1346
Iran	532	2135	2667	Ukraine	53	947	1000
Italy	92	920	1012	Egypit	24	3027	3051
Japan	47	1049	1096	United Kingdom	189	852	1041
Jordan	37	1163	1200	United States	466	783	1249
South Korea	228	972	1200	Burkina Faso	365	1169	1534
Malaysia	187	1014	1201	Uruguay	146	854	1000
Mali	561	973	1534	Zambia	932	568	1500
Mexico	639	921	1560				

Table 4.4: Wave 6: Distribution of the obtained classes by Country

Country	Positive	Negative	Total	Country	Positive	Negative	Total
Algeria	37	1163	1200	Malaysia	188	1112	1300
Azerbaijan	19	983	1002	Mexico	752	1248	2000
Argentina	166	864	1030	Morocco	18	1182	1200
Australia	246	1231	1477	Netherlands	207	1695	1902
Armenia	17	1083	1100	New Zealand	155	686	841
Brazil	756	730	1486	Nigeria	1397	362	1759
Belarus	76	1459	1535	Pakistan	113	1087	1200
Chile	231	769	1000	Paraguay	255	955	1210
China	24	2276	2300	Philippines	428	772	1200
Taiwan	200	1038	1238	Poland	138	828	966
Colombia	743	769	1512	Qatar	75	985	1060
Cyprus	91	909	1000	Romania	164	1339	1503
Denmark	184	1018	1202	Russia	50	2450	2500
Estonia	60	1473	1533	Rwanda	446	1081	1527
Georgia	83	1119	1202	Singapore	485	1487	1972
Palestine	55	945	1000	Slovenia	77	992	1069
Germany	270	1776	2046	South Africa	1934	1597	3531
Ghana	1078	474	1552	Zimbabwe	1043	457	1500
Haiti	191	1805	1996	Spain	77	1112	1189
Hong Kong	125	875	1000	Sweden	66	1140	1206
India	571	3507	4078	Thailand	183	1017	1200
Iraq	84	1116	1200	Trinidad and Tobago	391	608	999
Japan	91	2352	2443	Tunisia	7	1198	1205
Jamaica	45	1455	1500	Turkey	20	1585	1605
Jordan	57	1143	1200	Ukraine	68	1432	1500
South Korea	289	911	1200	Egypt	5	1518	1523
Kuwait	177	1126	1303	United States	814	1418	2232
Kyrgyzstan	123	1377	1500	Uruguay	126	874	1000
Lebanon	121	1079	1200	Uzbekistan	33	1467	1500
Libya	81	2050	2131	Yemen	36	964	1000

Table 4.5: Wave 7: Distribution of the obtained classes by Country

Country	Positive	Negative	Total
Andorra	60	944	1004
Argentina	152	851	1003
Australia	437	1376	1813
Bolivia	875	1192	2067
Brazil	773	989	1762
Chile	224	776	1000
Ecuador	241	959	1200
Germany	410	2612	3022
Greece	28	1172	1200
Indonesia	1598	1602	3200
Iraq	112	1088	1200
Jordan	56	1147	1203
South Korea	139	1106	1245
Lebanon	53	1147	1200
Malaysa	285	1028	1313
Nigeria	202	1035	1237
Pakistan	191	1809	2000
Romania	110	1147	1257
Russia	51	3584	3635
Serbia	92	1114	1206
Thailand	132	341	473

4.4 Models and Evaluation

Once the data was 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: a neural network Multi-layer Perceptron (MLP), Random Forest Classifier (RFC), linear models with Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM). Significantly, these four models were applied to each country.

We use Sklearn Library to carry out the models chosen. Table 4.6 shows the parameters applied in each model. Since the thesis purpose was to analyze human behavior, the models and the respective parameters were chosen empirically with few tests to define the most appropriate ones. The others were maintained standards.

Table 4.6: Model Parameters

Model	Parameters
Multi-layer Perceptron (HINTON, 1990)	MLPClassifier, solver=Adam, Alpha=1e-5, hidden_layer_sizes = (15,5)
Random Forest Classifier (BREIMAN, 2001)	RandomForestClassifier, criterion=entropy, n_estimators=50, max_features=15, bootstrap=True
Stochastic Gradient Descent (SAAD, 1998)	SGDClassifier, loss=squared_loss, penalty="l1"
Support Vector Machine (CHANG; LIN, 2011)	svm.SVC, Default Parameters

The validation of the model has been done using Stratified K-Folds cross-validator. In which ten different training and test sets, was separated to computing the evaluation metrics of each group to obtain the mean of the results.

We also used measures to evaluate the predictive and classification models, using the metrics F1-Score, which consists of a weighted harmonic mean of the precision and recall. To analyze the selection of features, we have built a global ranking for each wave based on the number of frequency of features showing the forty questions that have a high frequency and the forty that have low frequency in the process of RFE. With the goal of to analysis themed frequent in each wave to an understanding regarding motivations of people that choose to be a social activist.

Also, we present a table with the best results for each country, choosing the model that gave the best F1-Score. We also show the MCC and AUC in the table and the standard

Table 4.7: Matrix Confusion of Model

Predicted Class	True Class		
	Total Samples	**Social Activist**	**Non Social Activist**
	Social Activist	TP	FP
	Non Social Activist	FN	TN

deviation for each of the results.

4.4.1 Metrics

The confusion matrix has absolute variables, being TP - True Positive, TN - True Negatives, TP - False Positive, TN - False Negative, where the positive results refer to a positive case of social activism, and the negative of non social activism. Table 4.7 shows the Matrix Confusion of the Model. From the Matrix of Confusion, metrics are generated to evaluate the model's performance. The metrics used in this research will be presented below.

The Matthews Correlation Coefficient (MATTHEWS, 1975) is a measure of quality, which analyzes the (binary) classifications even when there is an imbalance between the positive and negative classes. The coefficient assumes values that vary between -1 and +1, where coefficients +1 refer to a perfect prediction, 0 random predictions, and -1 imperfect prediction (total disagreement). This measure is relevant in this work, as it makes a global analysis of predictions and indicates the quality of binary classifications in the context of the confusion matrix. The equation 4.1 shows the formula of Matthews Coefficient (MCC).

$$MCC = \frac{(TP \cdot VN - FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{4.1}$$

F1-Score in 4.2 is the harmonic mean of precision (Eq. 4.3) and recall (Eq. 4.4. Values range from zero to one.

$$F1 - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

And finally, AUC stands for "Area under the ROC Curve." The AUC measures the entire two-dimensional area below the entire ROC curve. AUC provides an aggregate measure of performance across all possible rating limits. One way to interpret AUC is to the probability that the model will rank a random positive example higher than a random negative example.

5 RESULTS AND DISCUSSIONS

In this chapter, we split the analyze for each wave, showing the performance of the models applied and thematic clusters. In the first one, we show F1-Score for each country in a heat map, differing by model, this way we can see a pattern overall.

In the second one, we developed using the conceptual cluster theory, classify each question selected by the global ranking of RFE in a thematic group. Each thematic group was defined based on the semantics of the issue.

For each wave, we create clusters based on the main subjects that appeared in the questions. Even though the clusters are not the same in all waves, we can verify a pattern. In the most frequent variables, the clusters named Family, Tolerance and Religion are present in all waves. On the other hand, the least frequency variables, the clusters that are present are those named Politics, Moral, Religion and Science.

5.1 WAVE 5

Figure 5.1 presents the results for each model applied in the countries using F1-Score. We defined a heat scale with purple that indicates relatively low F1-Score and orange that shows relatively high F1-Scores.

We can notice that the SVM (average 0.54) was the worst in most of the nations, producing low F1-Score, followed by MLP (average 0.61) and SGD (average 0.62). The most top results occurred when applied RFC, providing the max of average 0.88 in Australia (36), this model showed an excellent attain 0.65 in F1-Score on average. We can observe that Brazil (76) shows significant and similar results looking at the considered models. The difference is lower, with 0.69,0.70,0.70,0.72 to SVM, RFC, MLP, and SGD, respectively.

Table 5.2 shows the results of each country with the best model for each country. The metrics F1-Score, AUC, and MCC present the average of the ten tests by the cross-validation.

We can notice that countries like Canada, Australia, South Korea, Netherlands, New Zealand, United Kingdom, the United States in Wave 5 presented an MCC above 0.60, which indicates that the model obtained a hit rate between the two classes positively. When MCC is closer to 1, better the model's performance means that the prediction was perfect.

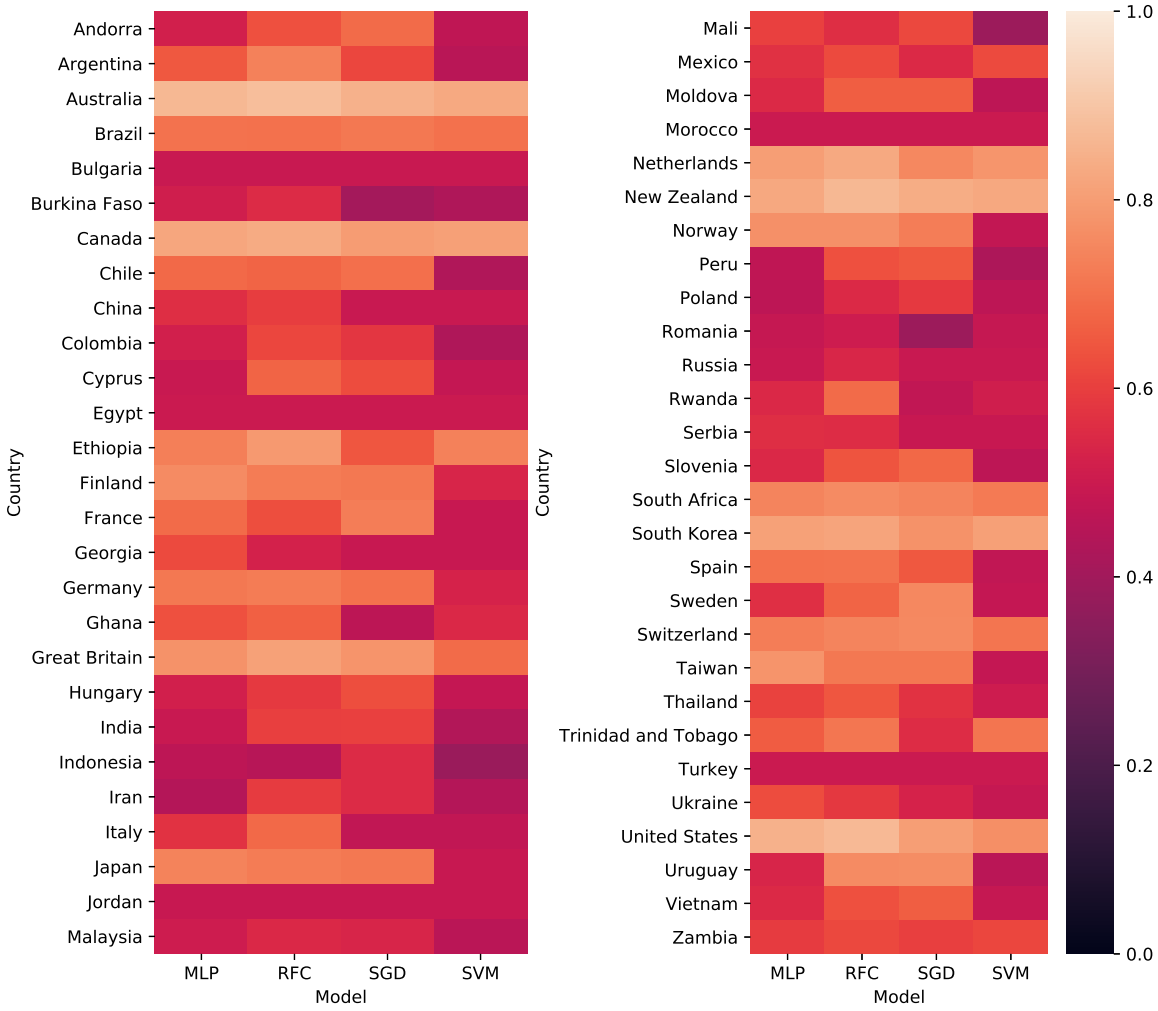
Table 5.1: Wave 5 - Best results for country

WAVE 5						
	F1-Score		AUC		MCC	
Country	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Andorra	0.69139	0.09598	0.66208	0.09276	0.45192	0.14050
Argentina	0.73698	0.07216	0.71501	0.06829	0.49312	0.13816
Australia	0.88119	0.04637	0.89404	0.04042	0.76549	0.08912
Brazil	0.71807	0.08192	0.72757	0.06284	0.48048	0.10774
Bulgaria	0.49547	0.00100	0.50000	0.00000	0.00000	0.00000
Canada	0.83205	0.03662	0.83830	0.03808	0.67288	0.06780
Chile	0.69862	0.08787	0.73281	0.08765	0.43371	0.16578
China	0.59431	0.09057	0.56793	0.06336	0.23909	0.21760
Taiwan	0.78095	0.07429	0.78850	0.09056	0.57657	0.14796
Colombia	0.61539	0.03597	0.61260	0.02739	0.26537	0.07136
Cyprus	0.67385	0.12270	0.71685	0.10941	0.42211	0.21800
Ethiopia	0.79067	0.08660	0.79415	0.08415	0.58622	0.17169
Finland	0.76017	0.06360	0.75705	0.08763	0.55243	0.09959
France	0.72916	0.10493	0.74416	0.10854	0.49396	0.21597
Georgia	0.62446	0.08964	0.69996	0.14148	0.28123	0.18759
Germany	0.71552	0.05891	0.68964	0.06101	0.45708	0.11698
Ghana	0.66417	0.05019	0.65565	0.04286	0.35904	0.10308
Hungary	0.62958	0.13482	0.64728	0.12851	0.30545	0.26239
India	0.60477	0.07990	0.63262	0.08849	0.25673	0.15620
Indonesia	0.55110	0.09195	0.59880	0.05452	0.21366	0.09663
Iran	0.59170	0.05082	0.58672	0.04423	0.20603	0.10116
Italy	0.68371	0.08669	0.65418	0.07272	0.40816	0.16327
Japan	0.74218	0.14399	0.74380	0.15939	0.49588	0.28575
Jordan	0.49217	0.00094	0.50000	0.00000	0.00000	0.00000
South Korea	0.81802	0.07963	0.82300	0.09079	0.65117	0.14373
Malaysa	0.54386	0.09038	0.54603	0.06773	0.15763	0.19516
Mali	0.62093	0.07017	0.63195	0.06314	0.27222	0.11513

Table 5.2: Wave 5 - Best results for country

WAVE 5						
	F1-Score		AUC		MCC	
Country	Mean	Std	Mean	Std	Mean	Std
Mexico	0.62226	0.06607	0.62462	0.06491	0.25230	0.12986
Moldova	0.66344	0.09214	0.64672	0.07604	0.36468	0.17164
Morocco	0.49644	0.00095	0.50000	0.00000	0.00000	0.00000
Netherlands	0.83000	0.05623	0.82310	0.06845	0.66574	0.11410
New Zealand	0.86697	0.03603	0.85248	0.05673	0.74278	0.06789
Norway	0.77197	0.10912	0.76673	0.11811	0.56732	0.20185
Peru	0.64847	0.06679	0.64961	0.06494	0.32965	0.11226
Poland	0.58671	0.08591	0.58248	0.07012	0.18727	0.17582
Romania	0.50459	0.03935	0.50903	0.02268	0.04591	0.11607
Russia	0.53947	0.07128	0.53049	0.04977	0.09850	0.16671
Rwanda	0.68874	0.04458	0.69131	0.04279	0.38964	0.08543
Serbia	0.56189	0.09413	0.54825	0.07009	0.15822	0.21488
Vietnam	0.66079	0.07279	0.69178	0.11215	0.35757	0.15551
Slovenia	0.68179	0.10161	0.69229	0.10913	0.38880	0.19713
South Africa	0.76085	0.04753	0.76221	0.04568	0.53105	0.08425
Spain	0.70409	0.07246	0.68534	0.08016	0.43640	0.13796
Sweden	0.75284	0.11371	0.77310	0.13781	0.53707	0.21495
Switzerland	0.75589	0.03292	0.76169	0.06089	0.54110	0.05821
Thailand	0.64532	0.15569	0.65949	0.11736	0.36813	0.26302
Trinidad and Tobago	0.71391	0.06850	0.71744	0.06784	0.46029	0.13698
Turkey	0.49645	0.00056	0.50000	0.00000	0.00000	0.00000
Ukraine	0.62511	0.11906	0.63294	0.11939	0.28279	0.24280
Egypt	0.49803	0.00040	0.50000	0.00000	0.00000	0.00000
United Kingdom	0.81015	0.06439	0.84266	0.06075	0.63334	0.12007
United States	0.86893	0.03203	0.87053	0.03290	0.74455	0.06167
Burkina Faso	0.55245	0.06986	0.56195	0.04727	0.18053	0.11808
Uruguay	0.76194	0.06023	0.80897	0.09311	0.54812	0.11794
Zambia	0.62091	0.07580	0.62292	0.07207	0.25234	0.15385

Figure 5.1: F1-Score for model in Wave 5



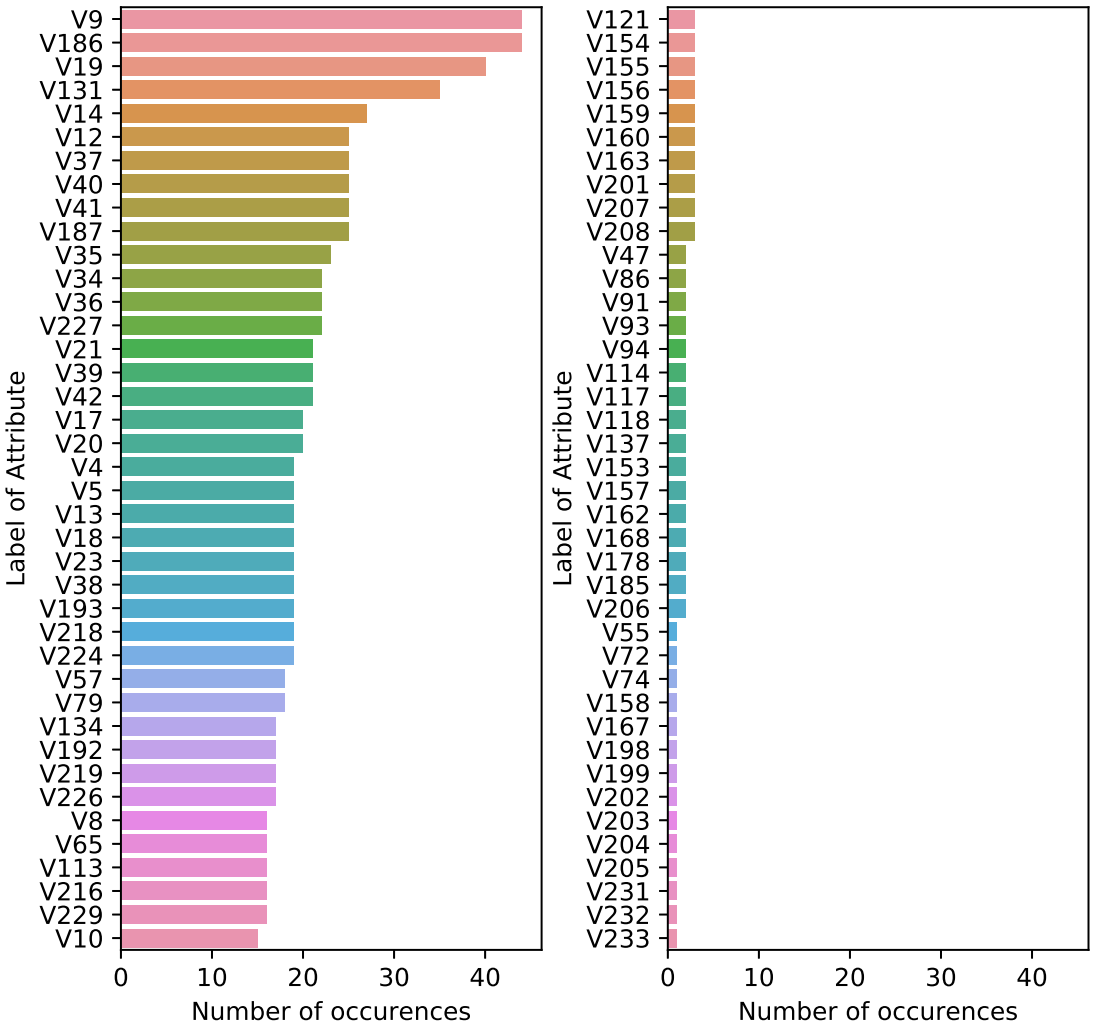
On the other hand, some countries like Bulgaria, Jordan, Morocco, Egypt, obtained a random prediction according to the MCC, which can also be seen by looking at the other table's metrics.

Wave 5 considered 55 countries. In the process of selection of features, we selected forty questions represented in Figure 5.2, by attributes.

In this ranking, we computed the number of frequency of each variable, looking if each selected feature was present in the country evaluated or not. In the left in Figure 5.2, majority of frequency shows that the questions V9, V186, V19, and V131 were more than 35 countries.

It shows these issues are representative in the task of classifying activist or non-

Figure 5.2: Ranking of frequency of questions in the countries of the Wave 5

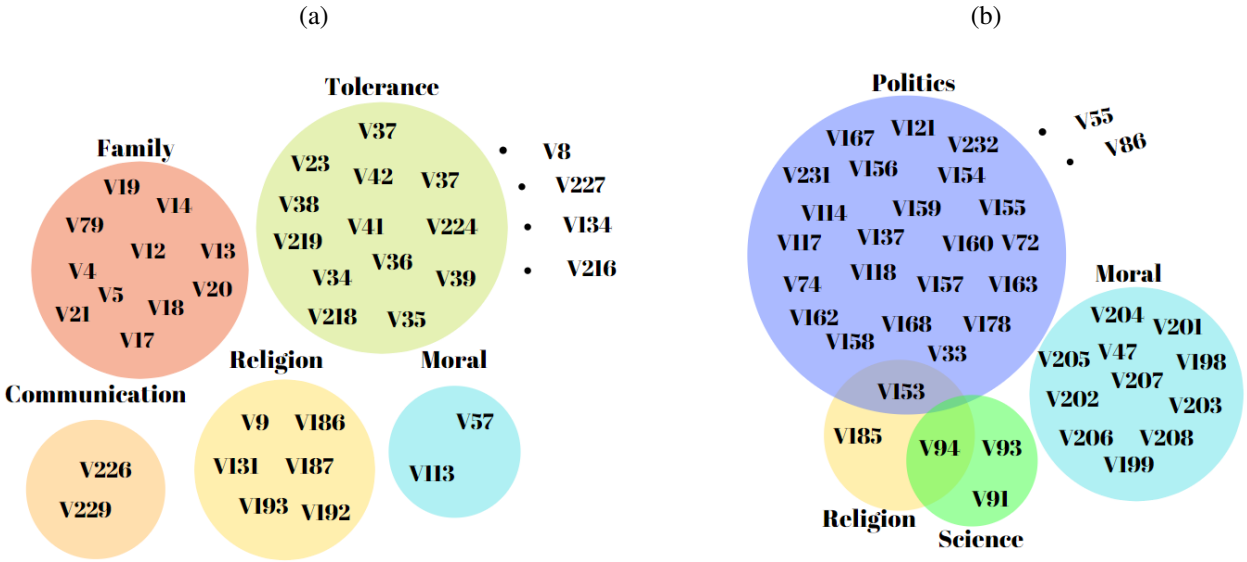


activist behaviour. From V14 with 27 frequency, the number tends to fall but in short steps, until reaching the V10 with 15 countries.

On the other hand, in the right in this figure, we can observe that attributes are not typical for all the countries analyzed, but indeed for a minority of them, reaching a maximum of 3 countries. Thus we can recognize two extremes, one with a pattern in most nations and other that is specific for some countries.

Seeking to understand the meanings, we developed an approach for classifying and building thematic clusters from the forty questions frequent and non-frequent in the countries of each wave based on substantive Social Science theory (using a inductive perspective). In this Wave, we identify five clusters (Tolerance, Family, Religion, Moral

Figure 5.3: Themed Cluster - Wave 5. Figure (a) - Attribute distribution according to the theme of the most frequent variables in the analyzed countries. Figure (b) - Attribute distribution according to the them of the least variables in the analyzed countries



and Communication) in the most frequent variables and four clusters (Politics, Moral, Science and Religion) in the least frequent variables.

Figure 5.3 (a) shows the distribution of questions in five different clusters. We can notice a considerable number of issues in the clusters that represent questions about Tolerance and Family. These specific questions search to measure behavior about how comfortable are the respondent regarding Social diversity, and concerning the values taught to children. Moreover, the questions V8, V227, V134, V216 are different from the clusters and do not have an explicit relationship to the found pattern, in other words, these questions do not belong to any cluster.

On the other hand, Figure 5.3 (b) referent minority questions selected have a representative grouping concerning Politics. These topics cover elections, characteristics of democracy, government, laws, and political ideology to understand positioning and concerns about the political scenario. In the Moral cluster debatable themes, and corrupt behavior are analyzed , but, in fewer attributes compared to Politics. Also, the questions V55 and V86 do not belong to any clusters.

Regarding the general issues, we can notice that in this Wave, most of the countries and attributes used in the models describe a behavior based on ethical and moral topics.

Only discerning the distribution in Figure 5.3 (a) and (b), using normative analysis, we can suppose that the people's behavior in the countries analyzed tend to be more

concerned with issues of moral-ethical conduct than politics.

The social activists, at least in the Wave 5, are more inclined to consider important subjects as Tolerance, Family and Religion, as we can see in the clusters. So, in this case, we can assume that moral-ethical issues are more important to the social activists. In this sense, Traditional Values, which means focused on the need for material survival, respect for the authorities, strong religious convictions and well-established family relationships (INGLEHART; WELZEL, 2005).

Additionally, we collect the frequencies variables of the top five of the ranking of Table 5.2 in percent among all countries. We present and analyze the distribution of the interviewers' choices in each of the essential questions in Wave 5.

Table 5.3: Valid Percentage of Top five questions

FREQUENCIES VARIABLES
WAVE 5

V186 - How often do you attend religious services	**Valid percent**
More than once a week	17,3
Once a week	18,6
Once a month	10,8
Only holy days	15,5
Once a year	5,8
Less often	10,7
Never	21,3
Total	100,0

V9 - Religion Important	**Valid percent**
Very important	48,7
Rather important	22,4
Not very important	17,4
Not at all important	11,5
Total	100,0

V131 - Confidence: Churches	**Valid percent**
A great deal	33,7
Quite a lot	32,4
Not very much	23,6
None at all	10,2
Total	100,0

V14 - Child qualities: feeling of responsibility	**Valid percent**
Mentioned	72,2
Not mentioned	27,8
Total	100,0

V19 - Child qualities: religious faith	**Valid percent**
Mentioned	40,9
Not mentioned	59,1
Total	100,0

Indeed, observing the choices of interviewers, we can notice that the presence of religion is more related to belief than a lack of religious faith. Besides, when it comes to children's education, we verify that the most part believes that their children's sense of responsibility is a value dominant. Under the congruence of the five most relevant

features of the survey to classify social activism for countries, we realized that the previous assumption the themed clusters makes sense, once we have the trend of values on ethical-moral issues and conservative standards of society.

5.2 WAVE 6

Figure 5.4 shows the F1-Score in the heat map in Wave 6 (2010-2014). The worst result was the model SVM (0.53), followed by SGD (0.58), MLP (0.61) and RFC (0.62) in average.Seeing the colors in the heat map, we can notice this behavior, where the SVM presents the colors darker, and RFC clearer, and slight variation differences between MLP and RFC.

Looking at all the performances in the models, the country that got the best result was (554) New Zealand 0.90 of F1-Score using RFC, and the worst was (332) Haiti with F1-Score below to 0.40 using SGD.

Table 5.5 shows the results of each country with the best model for each country. The metrics F1-Score, AUC, and MCC present the average of the ten tests by the cross-validation.

Following the idea presented in Wave 5, the countries that obtained an MCC above 0.60 were Australia, Chile, Estonia, Hong Kong, South Korea, Netherlands, New Zealand, and the United States. In particular, New Zealand had the best MCC, averaging 0.81, which shows that the predictions were assertive in most classifications.

Likewise, countries like Algeria, Armenia, Tunisia, Turkey, Egypt, and Uzbekistan had a random prediction.

Wave 6 was analyzed in 60 countries. In the left of Figure 5.5, we can observe that attributes V145, V9, V108, V19 despite having a few different codes compared with the four tops questions on Wave 5, both topics are around the concept of religion, and these are present in more than 30 countries.

Following the Wave 5 standard, the number of countries per ranking question decreases after these. Also, in the right of Figure 5.5, the number of frequency of the questions is still less than three, and it represents the behaviors in specific countries.

In this Wave, we identify four clusters (Tolerance, Family, Religion and Politics) in the most frequent variables and six clusters (Politics, Moral, Science, Religion, Work and Safety) in the least frequent variables. Moreover, the questions V10 and V110 do not have an explicit relationship to the pattern found, thus these questions do not belong to

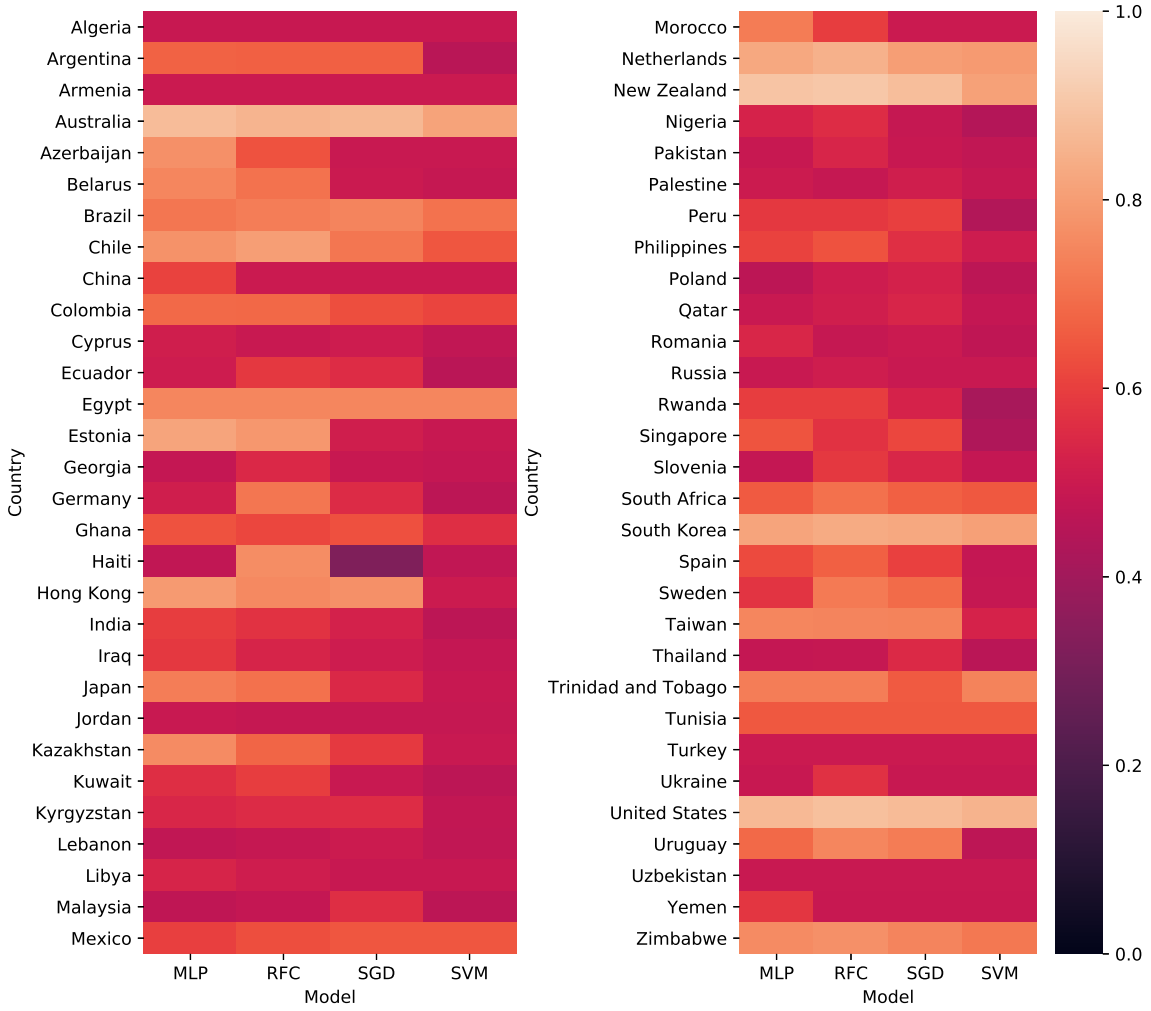
Table 5.4: Wave 6 - Best results for country

WAVE 6						
	F1-Score		AUC		MCC	
Country	Mean	Std	Mean	Std	Mean	Std
Algeria	0.49217	0.00094	0.50000	0.00000	0.00000	0.00000
Azerbaijan	0.77055	0.19072	0.77398	0.20657	0.56265	0.38393
Argentina	0.67177	0.05576	0.72108	0.07559	0.40097	0.07970
Australia	0.87834	0.04080	0.91237	0.02210	0.77154	0.07020
Armenia	0.49611	0.00104	0.50000	0.00000	0.00000	0.00000
Brazil	0.74251	0.04628	0.74794	0.04236	0.52346	0.08381
Belarus	0.74936	0.08363	0.82753	0.07813	0.53716	0.12074
Chile	0.80295	0.05352	0.79301	0.05910	0.61432	0.09794
China	0.60893	0.14196	0.60701	0.13393	0.22530	0.28915
Taiwan	0.74849	0.05436	0.73043	0.07148	0.53143	0.10022
Colombia	0.68581	0.05282	0.69770	0.04749	0.42263	0.09294
Cyprus	0.51355	0.08388	0.52668	0.06371	0.08287	0.17107
Ecuador	0.58260	0.04192	0.57651	0.03788	0.19548	0.10669
Estonia	0.81804	0.07404	0.84746	0.09932	0.66147	0.13450
Georgia	0.54367	0.07495	0.54726	0.06895	0.10932	0.16204
Palestine	0.51263	0.05371	0.51728	0.03475	0.05597	0.11213
Germany	0.71454	0.04232	0.71552	0.05526	0.43883	0.08246
Ghana	0.63932	0.06358	0.63358	0.05116	0.31010	0.13450
Haiti	0.76197	0.10197	0.75761	0.09530	0.53162	0.19621
Hong Kong	0.79463	0.06238	0.82163	0.08105	0.61187	0.12000
India	0.59712	0.05069	0.58963	0.04262	0.23540	0.09406
Iraq	0.58373	0.11813	0.61531	0.11227	0.22035	0.21544
Japan	0.72824	0.10692	0.70400	0.10385	0.47455	0.21178
Kazakhstan	0.76143	0.12072	0.76906	0.13198	0.58645	0.20407
Jordan	0.49248	0.01671	0.51904	0.06155	0.01496	0.06267
South Korea	0.83479	0.04496	0.84292	0.04878	0.68108	0.08480
Kuwait	0.59604	0.10362	0.58411	0.07501	0.27346	0.20308
Kyrgyzstan	0.55832	0.08350	0.55330	0.06837	0.18292	0.18426
Lebanon	0.50050	0.05977	0.51620	0.03688	0.06773	0.13843
Libya	0.53339	0.06849	0.54439	0.08574	0.09921	0.16071

Table 5.5: Wave 6 - Best results for country

WAVE 6						
	F1-Score		AUC		MCC	
Country	Mean	Std	Mean	Std	Mean	Std
Malaysa	0.55972	0.05441	0.55591	0.04010	0.15603	0.11128
Mexico	0.64553	0.05316	0.65851	0.05190	0.32617	0.08911
Morocco	0.72454	0.16870	0.92382	0.15485	0.53320	0.29907
Netherlands	0.84778	0.03196	0.85641	0.04067	0.69883	0.06319
New Zealand	0.90619	0.04854	0.90766	0.05468	0.81791	0.09617
Nigeria	0.55615	0.05911	0.56715	0.05368	0.24339	0.12623
Pakistan	0.53644	0.09792	0.53923	0.06386	0.13424	0.21187
Peru	0.59918	0.06455	0.62489	0.05241	0.26227	0.08870
Philippines	0.63686	0.04831	0.63651	0.04794	0.28993	0.09714
Poland	0.52518	0.06074	0.53371	0.03683	0.12454	0.14716
Qatar	0.53543	0.07699	0.57835	0.06943	0.14143	0.11864
Romania	0.54000	0.07882	0.55848	0.07263	0.12526	0.14603
Russia	0.51151	0.05037	0.50959	0.03015	0.04255	0.13373
Rwanda	0.59605	0.04307	0.59370	0.03493	0.24196	0.08649
Singapore	0.64156	0.08794	0.67409	0.08245	0.31672	0.14960
Slovenia	0.58545	0.09618	0.57018	0.06884	0.23602	0.20622
South Africa	0.70247	0.03082	0.70140	0.03044	0.40891	0.06252
Zimbabwe	0.77225	0.04263	0.75520	0.04241	0.56064	0.08200
Spain	0.66700	0.11198	0.63920	0.09339	0.35881	0.23477
Sweden	0.72015	0.07805	0.78246	0.06635	0.48221	0.13203
Thailand	0.54935	0.07661	0.56373	0.06564	0.14941	0.12444
Trinidad and Tobago	0.73954	0.06085	0.76025	0.05413	0.51341	0.10231
Tunisia	0.64855	0.23008	0.50000	0.00000	0.00000	0.00000
Turkey	0.49687	0.00001	0.50000	0.00000	0.00000	0.00000
Ukraine	0.56860	0.05766	0.54700	0.03473	0.20644	0.16810
Egypt	0.74918	0.25082	0.50000	0.00000	0.00000	0.00000
United States	0.88623	0.03562	0.89039	0.03199	0.77557	0.06575
Uruguay	0.74786	0.08410	0.71815	0.08405	0.52244	0.16330
Uzbekistan	0.49444	0.00075	0.50000	0.00000	0.00000	0.00000
Yemen	0.58111	0.10213	0.60333	0.10106	0.20259	0.20692

Figure 5.4: F1-Score for model in Wave 6

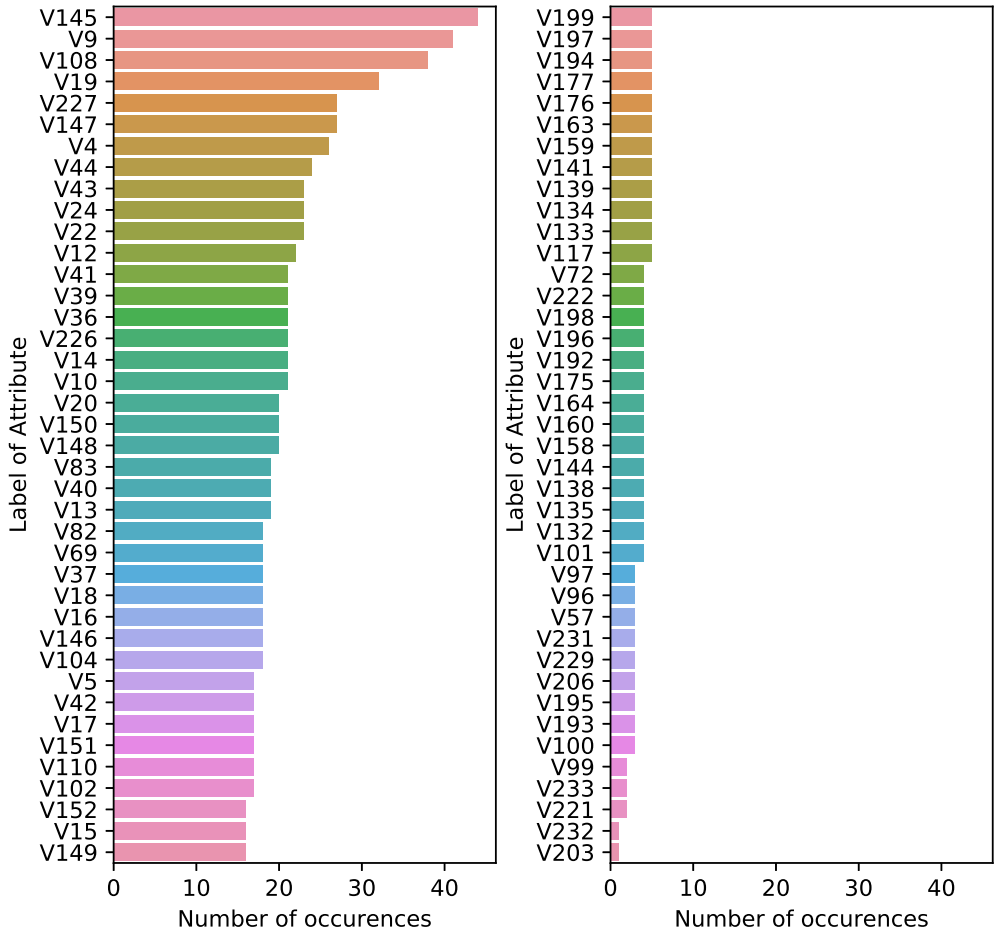


any cluster in the most frequent variables. Similarly, the questions V57, V144, V221 and V222 do not belong to any cluster in the least frequent variables.

In Figure 5.6 (a) the clusters represent issues in most of Family and Tolerance, as well as in Wave 5, and additionally Religion. Moreover, a few questions about Politics. In Figure 5.6 (b), we can notice that the number of clusters increases, including Politics, Moral, Religion, Science, Work, and Safety compared with (a). Despite this difference, the questions regarding moral-ethical arguments prevail in the majority of countries.

Following the structure of the clusters, we can notice that the concern is changing, especially in Figure 5.6 (b) that we see a movement even if in a few countries regarding other topics and interest in the period between 2010 and 2014 using this analysis.

Figure 5.5: Ranking of frequency of questions in the countries of the Wave 6



Comparing with the Wave 5, we can see now the cluster Politics inside the most frequent variables, but it is also in the least frequent variables. Based on this result, we can infer that politics is becoming more important.

In the table below we show the results of the response frequencies in the Top Five Features of 5.5 Wave 6.

Confirming the previous assumption, we assessed that religious belief and participation in elections are relevant to the task in a positive manner, as was noted in Wave 5.

Figure 5.6: Themed Cluster - Wave 6. Figure (a) - Attribute distribution according to the theme of the most frequent variables in the analyzed countries. Figure (b) - Attribute distribution according to the them of the least variables in the analyzed countries

(a) (b)

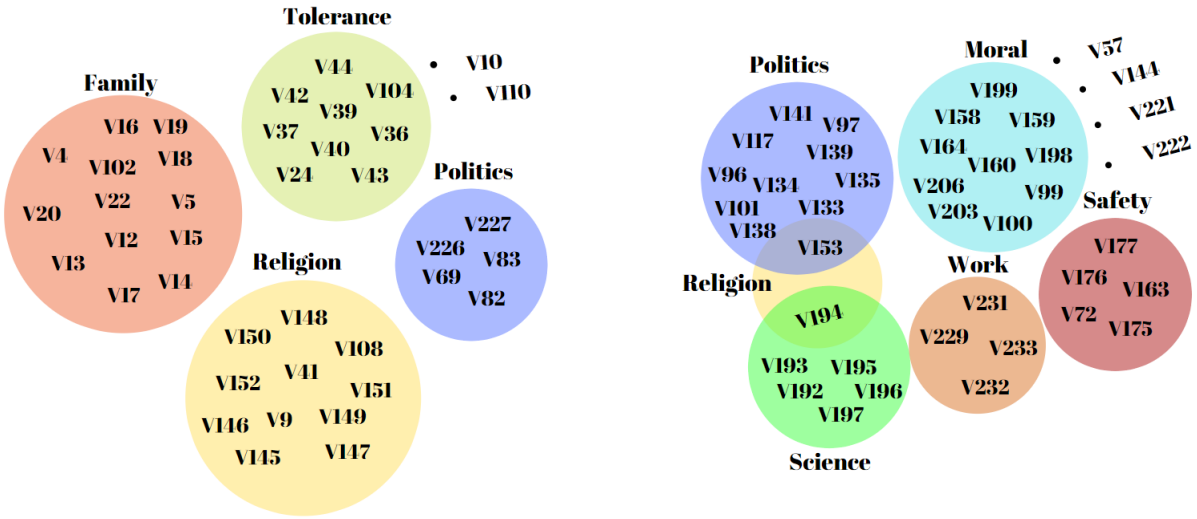


Table 5.6: Valid Percentage of Top five questions

FREQUENCIES VARIABLES

WAVE 6

V145 - How often do you attend religious services	Valid percent
More than once a week	14,3
Once a week	18,5
Once a month	9,8
Only on special holy days	17,0
Once a year	5,5
Less often	10,8
Never, practically never	24,0
Total	100,0

V9 - Important in life: Religion	Valid percent
Very important	50,3
Rather important	22,2
Not very important	16,1
Not at all important	11,4
Total	100,0

V108 - Confidence: The Churches	Valid percent
A great deal	34,2
Quite a lot	30,3
Not very much	23,0
None at all	12,6
Total	100,0

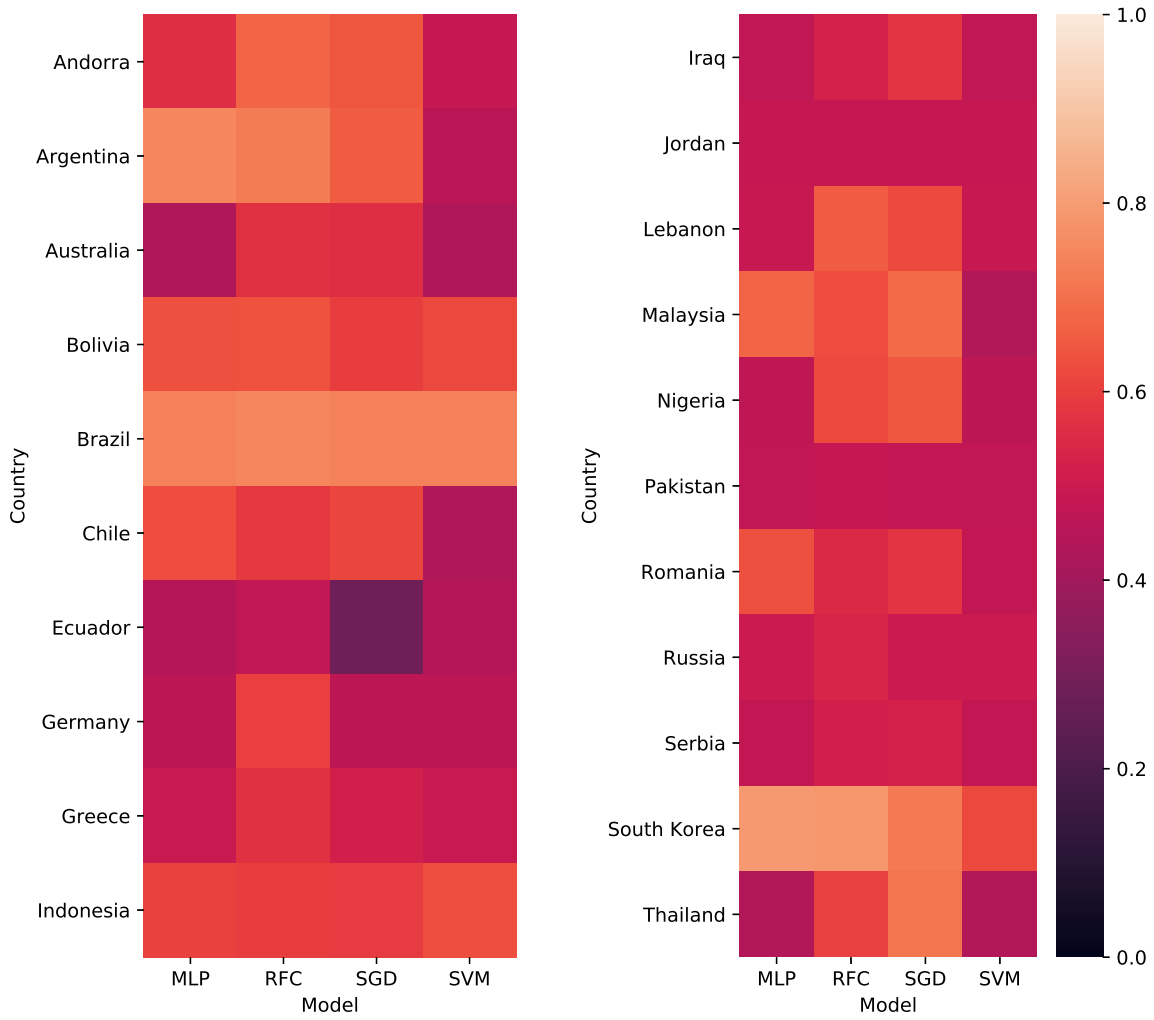
V227 - Vote in elections: National level	Valid percent
Always	58,5
Usually	23,0
Never	18,5
Total	100,0

V19 - Important child qualities: Religious faith	Valid percent
Mentioned	41,1
Not mentioned	58,9
Total	100,0

5.3 WAVE 7

The heat map in Figure 5.7 shows the results of Wave 7 (2017-2021).

Figure 5.7: F1-Score for model in Wave 7



Following the behavior of Wave 5 and Wave 6, RFC shows the best average (0.60), and SVM the worst average (0.50). The SGD and MLP the differences are lower, 0.58 and 0.56, respectively.

Wave 7 has analyzed 35 countries representing the period between 2017 and 2021. The nations collected the data during the year of 2018, applying the Survey.

Table 5.7 shows the results of each country with the best model for each country.

Table 5.7: Wave 7 - Best results for country

WAVE 7						
	F1-Score		AUC		MCC	
Country	Mean	Std	Mean	Std	Mean	Std
Andorra	0.67802	0.11810	0.68202	0.12652	0.37468	0.23587
Argentina	0.74617	0.06246	0.72597	0.08450	0.53321	0.10515
Australia	0.57025	0.04736	0.57283	0.03707	0.19954	0.07914
Bolivia	0.64035	0.04020	0.66373	0.02640	0.33665	0.04478
Brazil	0.74742	0.04491	0.75165	0.04227	0.50867	0.08487
Chile	0.62821	0.09563	0.63789	0.09525	0.32388	0.18220
Ecuador	0.47630	0.03872	0.50435	0.02137	0.00484	0.08074
Germany	0.60189	0.08013	0.58118	0.05535	0.30107	0.17805
Greece	0.56973	0.11524	0.55000	0.07638	0.17174	0.26234
Indonesia	0.63074	0.03983	0.64177	0.03381	0.30045	0.06628
Iraq	0.57656	0.07132	0.59470	0.07004	0.21685	0.14623
Jordan	0.48809	0.00098	0.50000	0.00000	0.00000	0.00000
South Korea	0.79014	0.08244	0.76543	0.11061	0.61322	0.14124
Lebanon	0.65830	0.13052	0.63108	0.12179	0.36671	0.26463
Malaysa	0.69076	0.06374	0.70502	0.06849	0.40700	0.10658
Nigeria	0.64817	0.21690	0.74744	0.14319	0.43921	0.27502
Pakistan	0.48490	0.02072	0.50471	0.01093	0.04050	0.08966
Romania	0.63332	0.09970	0.65569	0.11819	0.31270	0.18986
Russia	0.53768	0.08997	0.52819	0.06226	0.09121	0.19829
Serbia	0.52538	0.07293	0.53182	0.05153	0.08231	0.14158
Thailand	0.71093	0.11497	0.72704	0.09436	0.47702	0.18867

The metrics F1-Score, AUC, and MCC present the average of the ten tests by the cross-validation.

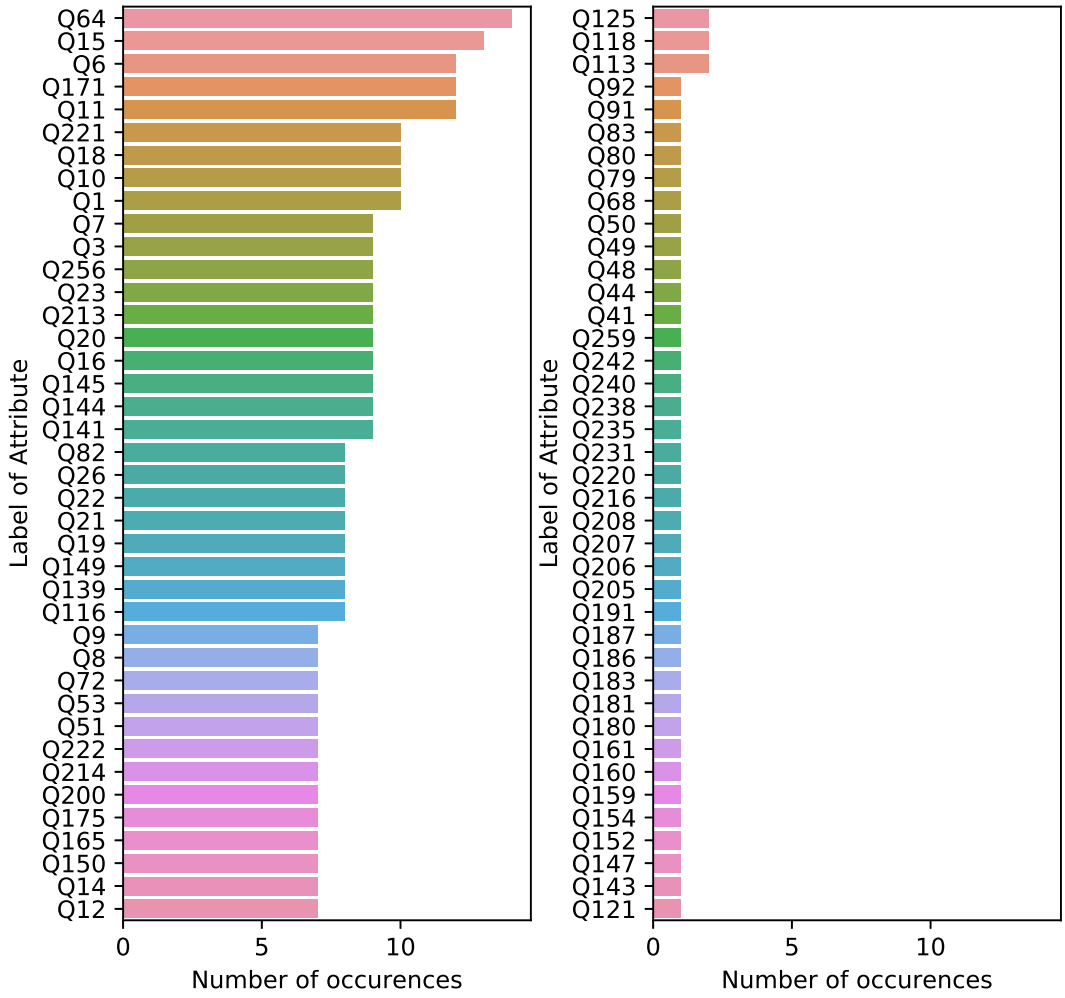
Among the countries evaluated in Wave 7, only South Korea had an MCC above 0.60. And the random prediction was in Jordan.

In the ranking in Figure 5.8, the questions on top were Q64, Q15, Q6, Q171 and Q11. These questions discuss religion, as well as in Wave 5 and Wave 6 describing the influence of this topic in most of the countries investigated.

Just as we saw in the other waves, in this one we identify six clusters (Tolerance, Family, Religion, Politics, Safety and Well-Being) in the most frequent variables and eight clusters (Politics, Moral, Science, Religion, Communication, Tolerance, Well-Being and Safety) in the least frequent variables.

Moreover, the questions Q3, Q116 and Q256 do not have an explicit relationship to the pattern found, thus these questions do not belong to any cluster in the most frequent variables. Similarly, the questions Q41 and Q259 do not belong to any cluster in the least frequent variables.

Figure 5.8: Ranking of frequency of questions in the countries of the Wave 7



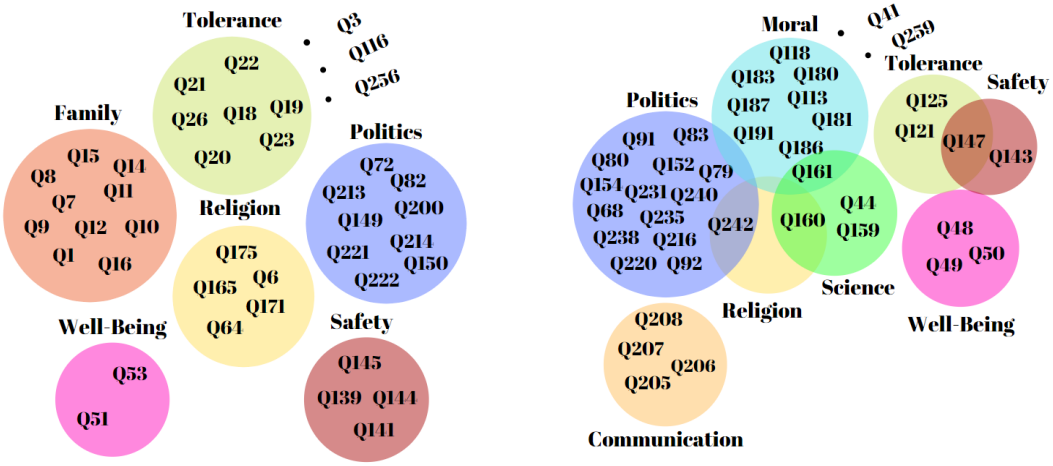
The clusters formulated in Figure 5.9 (a) show that in most of countries, differently of Wave 5 and 6, had an expressive distribution of features in theory clusters, dividing among Family, Tolerance, Religion, Politics, Safety, and Well-Being.

Differently, in Figure 5.9 (b), the distribution is more concentrated in the cluster on Politics and Moral. And, some features in other groups are present in lower quantity. Comparing with the other waves, for the first time the cluster Politics surpassed the cluster Religion in the most frequent variables. However, in the least frequent variable the cluster Politics is still important.

In the Table 5.8 issues of religion predominate, as well as the presence of questions regarding children's education, that, as other Waves present values related to con-

Figure 5.9: Themed Cluster - Wave 7. Figure (a) - Attribute distribution according to the theme of the most frequent variables in the analyzed countries. Figure (b) - Attribute distribution according to the them of the least variables in the analyzed countries

(a) (b)



servatism.

Table 5.8: Valid Percentage of Top five questions

FREQUENCIES VARIABLES
WAVE 7

Q171 - How often do you attend religious services

	Valid percent
More than once a week	10,4
Once a week	17,2
Once a month	11,7
Only on special holy days	18,2
Once a year	6,0
Less often	12,3
Never, practically never	24,2
Total	100,0

Q6 - Important in life: Religion

	Valid percent
Very important	42,1
Rather important	25,1
Not very important	19,7
Not at all important	13,1
Total	100,0

Q64 - Confidence: The Churches

	Valid percent
A great deal	27,7
Quite a lot	30,4
Not very much	27,1
None at all	14,8
Total	100,0

Q11 - Important child qualities: Imagination

	Valid percent
Mentioned	19,9
Not mentioned	80,1
Total	100,0

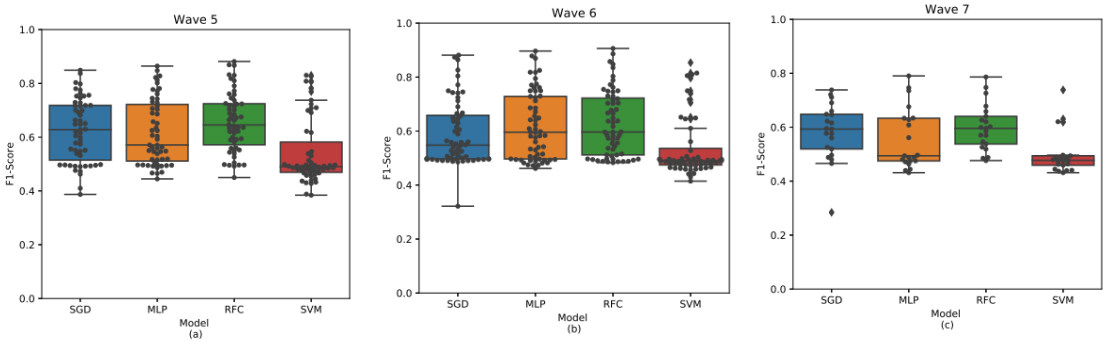
Q15 - Important child qualities: Religious faith

	Valid percent
Mentioned	32,3
Not mentioned	67,7
Total	100,0

5.4 Results Overview

In Figure 5.10, we can see an overall about the effects of F1-Score in the waves. The points on the graph represent the F1-Score of the countries.

Figure 5.10: F1-score comparison between the evaluated



The SVM model in this task had the same behavior between the waves, presenting a low variability, standard deviation, and average of F1-Score compared with other models applied. In the SGD, and MLP model had higher variability in Wave 5 and Wave 6. While in Wave 7, those models differ regarding with distribution of F1-Score.

Results of the F1-Score of the models demonstrated that the classification task related to social activism achieved satisfactory results, especially applying the RFC model.

The process of understanding the motivations to perform social activism in the RFE ranking shows that religion is an essential issue in most of the countries, knowing that the four questions in top of the rankings are around this topic. Also, other thematics as Tolerance and Family determine that is concerning the nations. It is important to note that applying a feature selection model different from the RFE, which we apply in this work, can result in different observations or even confirm this research's findings.

In Wave 6 and Wave 7, we can notice that the assumption made in Wave 5 that the world population is concerned with issues of moral-ethical rather than other aspects, such as politics, is present. However, we can perceive that in the longitudinal view, the matters of politics are starting gains in the most frequent features.

In general, the countries that obtain a more significant predictive capacity are economically well-developed, demonstrating that the study's object standards were best found in these countries. It is also clear that traditional values are more linked to social activism activities than secular-rational values in themed clusters. This relation indicates that political participation is related to moral and ethical values than concerns related to

the economy.

5.5 Social Sciences - Now and Next in Artificial Intelligence

A central objective of the social sciences is to understand political, economic and social behavior. Social Science researchers are experts in understanding and explaining human behaviour using resources and methods such as surveys, statistics and other empirical data. The use of Artificial Intelligence techniques to construct new hypotheses and explanations of political, economic and social reality is a challenge for both the Social Sciences and Computer Science.

There are several Social Science and Humanities researches that verify the moral and ethical aspects of the impact of the growing use of AI in society (FLORIDI et al., 2018) (Došilović; Brčić; Hlupić, 2018) (RAHWAN et al., 2019).

Besides, human behavior studies using Artificial Intelligence are usually related to the use of social networks, such as Twitter and Facebook, focusing on the digital world. Even knowing that people tend to reproduce behavior in the media, it is necessary to understand the practice from a socialization point-of-view.

Social scientists are beginning to use AI as a tool of analysis to understand the phenomena of society. Tiago Vier used AI to understand the aspect of Patriotism using this strategy, comparing results of traditional data analysis with the use of AI algorithms (VIER, 2020). In addition, it is worth mentioning a study that used survey data with the application of the SARIMA model to analyze and build a behavioral model of flu search (WOJCIK et al., 2020).

Our expectations are that research such as these will become more frequent, particularly at the time of the re-organization of society during and after the current pandemic. In other words, we will increasingly see the integration of the fields of Social Sciences and Artificial Intelligence.

On the one hand, the use of Social Sciences tools by AI scientists can create new Artificial Intelligence techniques and support the development of new technologies. On the other, the interest of social scientists in using AI can improve the accuracy of social studies.

Thus, interdisciplinary studies can be beneficial for the development and scientific advancement of both areas.

6 CONCLUSION

The development of this research enabled an analysis of how Artificial Intelligence can contribute to Social Sciences. We studied subjects that guided the understanding of social activism through machine learning models. In Wave 5, Wave 6, and Wave 7, we apply four models: SGD, SVM, RFC, and MLP. Also, we use a hybrid analysis approach, including studies in computer science and Social Sciences. For example, it was using a normative analysis that is commonly applied by Social Scientists.

The main finding of this thesis is that the use of AI techniques in Social Sciences structured data, as surveys databases, seems to bring new perspectives of analysis. As can be seen above, information that is not generally evidenced by the use of traditional statistical analysis techniques may appear. We found patterns in the databases using different analyses in different waves that allow both questioning and deepening existing knowledge. It is striking that Machine Learning tools have organized countries into groups according to dimensions that make sense from the perspective of the Social Sciences, but that challenge the existing knowledge.

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.

One of the factors that influence human behavior is values,as an example, which motivate actions. Understanding these values can help to improve society from measures to overcome social problems. The revolutions in our communities, for example, the abolition of slavery, the right of the feminine vote, happened with social movements from leadership who were able to identify the problems and propose solutions.

With the exponential growth of the data, stakeholders need to take ownership of these interdisciplinary scientific works to understand the effects and the motivations related to the actions, as an example, the misinformation (fake news). In this way, leaders can act with a political strategy to successfully create social change.

The findings of this thesis indicate that the research path used here is fruitful and should be deepened. In this thesis, we can infer that the causes that motivate social activism are more related to social causes than political purposes, according to the results.

However, more in-depth studies in the field of AI, as well as analyses from the perspective of the Social Sciences that could help explain the findings and propose new lines of investigation are needed. Furthermore, this research paves the way for interdisciplinary studies of AI and Social Sciences with gains for both fields.

As future work, we propose applying different strategies for selecting features to compare with the results and considerations found in this research. Also, apply the approach developed in this dissertation with other topics of interest to society for new insights.

6.1 Publications

NASCIMENTO, Francielle M; BARONE, Dante; CASTRO, Henrique Carlos.Social Activism Analysis: An Application of Machine Learning in the World Values Survey. Paper Published in MLDM - International Conference on Machine Learning and Data Mining, July 20-25, 2019, New York, USA.

VIER, Thiago. NASCIMENTO, Francielle M.The World Values Survey and the New Summer of Artificial Intelligence. Is There Anything New to be Uncovered?.IV ISA Forum of Sociology (February 23-27, 2020)."Cross-National Research: Ex-Post Survey Data Harmonization and Comparability".

REFERENCES

ASSOCIATION, W. V. S. et al. World values survey. **Wave**, v. 5, p. 2005–2008, 2005.

BHATTACHERJEE, A. Social science research: Principles, methods, and practices. Global Text Project, 2012.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Oct 2001.

CASTRO, H. C. d. O. d.; RANINCHESKI, S.; CAPISTRANO, D. O conteúdo da globalização para os latino-americanos: uma análise a partir da pesquisa mundial de valores–wvs. UNR Editora, 2015.

CASTRO, H. C. d. O. d.; REIS, F. T. Participação política no brasil no século xxi: mudanças e continuidades. 2012.

CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. **ACM transactions on intelligent systems and technology (TIST)**, Acm New York, NY, USA, v. 2, n. 3, p. 1–27, 2011.

DAVIS, J. B. Normative and positive economics. Taylor & Francis (Routledge), 1998.

Došilović, F. K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In: **2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)**. [S.l.: s.n.], 2018. p. 0210–0215.

FLORIDI, L. et al. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. **Minds and Machines**, Springer, v. 28, n. 4, p. 689–707, 2018.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. **Machine learning**, Springer, v. 46, n. 1-3, p. 389–422, 2002.

HARDT, M.; RECHT, B.; SINGER, Y. Train faster, generalize better: Stability of stochastic gradient descent. In: BALCAN, M. F.; WEINBERGER, K. Q. (Ed.). New York, New York, USA: PMLR, 2016. (Proceedings of Machine Learning Research, v. 48), p. 1225–1234. Available from Internet: <http://proceedings.mlr.press/v48/hardt16.html>.

HAYKIN, S. A comprehensive foundation. **Neural networks**, v. 2, n. 2004, p. 41, 2004.

HINTON, G. E. Connectionist learning procedures. In: **Machine learning**. [S.l.]: Elsevier, 1990. p. 555–610.

INGLEHART, R. et al. **World Values Survey: All Rounds - Country-Pooled Datafile Version: http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp**. Madrid, Espanha: JD Systems Institute, 2014.

INGLEHART, R. et al. **Rising tide: Gender equality and cultural change around the world**. [S.l.]: Cambridge University Press, 2003.

INGLEHART, R.; WELZEL, C. **Modernization, Cultural Change, and Democracy: The Human Development Sequence**. [S.l.]: Cambridge University Press, 2005.

JONES, E. **Developments in European Politics 2**. [S.l.]: Macmillan International Higher Education, 2011.

KANG, N. **Introducing Deep Learning and Neural Networks — Deep Learning for Rookies (1) | by Nahua Kang | Towards Data Science**. 2017. <https://towardsdatascience.com/introducing-deep-learning-and-neural-networks-deep-learning-for-rookies-1-bd68f9cf5883>. (Accessed on 08/23/2020).

KARTHIKA, S.; KIRUTHIGA, A.; BOSE, S. Behavioral profile generation for 9/11 terrorist network using efficient selection strategies. In: **Advances in Computer Science, Engineering & Applications**. [S.l.]: Springer, 2012. p. 333–344.

KAUFMAN, S. et al. Leakage in data mining: Formulation, detection, and avoidance. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM, v. 6, n. 4, p. 15, 2012.

KENDE, A. Separating social science research on activism from social science as activism. **Journal of Social Issues**, Wiley Online Library, v. 72, n. 2, p. 399–412, 2016.

KULESZA, T. et al. Why-oriented end-user debugging of naive bayes text classification. **ACM Trans. Interact. Intell. Syst.**, ACM, New York, NY, USA, v. 1, n. 1, p. 2:1–2:31, oct. 2011. ISSN 2160-6455. Available from Internet: <http://doi.acm.org/10.1145/2030365.2030367>.

LI, Q. et al. A quantitative analysis of global environmental protection values based on the world values survey data from 1994 to 2014. **Environmental Monitoring and Assessment**, v. 190, n. 10, p. 593, Sep 2018.

LIPSEY, R. G. **An introduction to positive economics**. [S.l.]: Weidenfeld and Nicolson, 1968.

LOHUIZEN, J. van; SAMOHYL, R. W. Method effects and robo-polls. **Survey Practice**, v. 4, n. 1, p. 3057, 2011.

MATTHEWS, B. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. **Biochimica et Biophysica Acta (BBA) - Protein Structure**, v. 405, n. 2, p. 442 – 451, 1975. ISSN 0005-2795. Available from Internet: <http://www.sciencedirect.com/science/article/pii/0005279575901099>.

MCCORMICK, T. H. et al. Using twitter for demographic and social science research: Tools for data collection and processing. **Sociological methods & research**, SAGE Publications Sage CA: Los Angeles, CA, v. 46, n. 3, p. 390–421, 2017.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. **Artif. Intell.**, v. 267, p. 1–38, 2017.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2.

NELSON, J. B.; KENNEDY, W. G.; KRUEGER, F. Exploratory models of trust with empirically-inferred decision trees. In: SPRINGER. **International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation**. [S.l.], 2016. p. 42–50.

NIELSEN, M. A. **Neural networks and deep learning**. [S.l.]: Determination press San Francisco, CA, 2015.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PORTA, D. D.; KEATING, M. **Approaches and Methodologies in the Social Sciences: A Pluralist Perspective**. [S.l.]: Cambridge University Press, 2008.

RAHWAN, I. et al. Machine behaviour. **Nature**, Nature Publishing Group, v. 568, n. 7753, p. 477–486, 2019.

REALE, E. et al. A review of literature on evaluating the scientific, social and political impact of social sciences and humanities research. **Research Evaluation**, v. 27, n. 4, p. 298–308, 07 2017. ISSN 0958-2029. Available from Internet: <https://doi.org/10.1093/reseval/rvx025>.

ROBILA, M.; ROBILA, S. A. Applications of artificial intelligence methodologies to behavioral and social sciences. **Journal of Child and Family Studies**, Springer, p. 1–13, 2019.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. In: RUMELHART, D. E.; MCCLELLAND, J. L.; GROUP, C. P. R. (Ed.). Cambridge, MA, USA: MIT Press, 1986. chp. Learning Internal Representations by Error Propagation, p. 318–362. ISBN 0-262-68053-X. Available from Internet: <http://dl.acm.org/citation.cfm?id=104279.104293>.

SAAD, D. Online algorithms and stochastic approximations. **Online Learning**, Cambridge Univ. Press, v. 5, p. 6–3, 1998.

SATO, A.-H. **Applied data-centric social sciences**. [S.l.]: Springer, 2014.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding machine learning: From theory to algorithms**. [S.l.]: Cambridge university press, 2014.

SMOLA, A. J.; SCHöLKOPF, B. **A tutorial on support vector regression**. 2004.

VAPNIK, V. The support vector method of function estimation. In: ____. **Nonlinear Modeling: Advanced Black-Box Techniques**. Boston, MA: Springer US, 1998. p. 55–85.

VIER, T. **O USO DA INTELIGÊNCIA ARTIFICIAL NAS CIÊNCIAS SOCIAIS: O caso do patriotismo dos brasileiros.** Thesis (PhD) — Universidade Federal do Rio Grande do Sul, Instituto de Filosofia e Ciências Humanas, Curso de Pós-Graduação em Ciência Política, Porto Alegre, 2020.

WELZEL, C.; INGLEHART, R.; KLIGEMANN, H.-D. The theory of human development: A cross-cultural analysis. **European Journal of Political Research**, Wiley Online Library, v. 42, n. 3, p. 341–379, 2003.

WOJCIK, S. et al. Survey data and human computation for improved flu tracking. **arXiv preprint arXiv:2003.13822**, 2020.

ZHANG, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: **ICML 2004: Proceedings of the Twenty-First International Conference on Machine Learning. Omnipress.** [S.l.: s.n.], 2004. p. 919–926.

Appendices

AppendixA RESUMO EXPANDIDO EM PORTUGUÊS

Contribuições de Aprendizado de Máquina para a aquisição de conhecimento na área de Ciências Sociais

A pesquisa em Ciências Sociais é fundamental para o estudo do comportamento humano. Crenças e motivações desempenham um papel importante nas decisões e escolhas das pessoas. Essa relação é relevante para explicar o comportamento de uma população e, portanto, permite delinear ações sociais para a melhoria da comunidade. Sabendo disso, propusemos uma forma de descobrir padrões a partir de um banco de dados de estudos sociais usando técnicas de Inteligência Artificial e Ciências Sociais. Nesse contexto, selecionamos o Ativismo Social para realizar a classificação por meio do banco de dados Word Values Survey (WVS). Os algoritmos aplicados na tarefa foram Random Forest, Multilayer Perceptron, Stochastic Gradient Descent e Support Vector Machine.

Além disso, usamos Recursive Feature Elimination para reduzir a dimensionalidade e analisar as features selecionadas. O dataset utilizado contém uma pesquisa aplicada em diversos países, organizada em Ondas realizadas a cada cinco anos. As ondas tratadas neste estudo foram Onda 5 (2005-2009), Onda 6 (2010-2014) e Onda 7 (2018-2022).

Especificamente, propomos o uso de técnicas de Aprendizado de Máquina para complementar os estudos de ciências sociais na base de dados WVS, com foco no entendimento das relações entre as features selecionadas no processo de construção do modelo. Tendo o Ativismo Social como alvo, desenvolvemos este trabalho a partir das relações entre o Ativismo Social e os valores humanos em torno deste tema.

O objetivo geral deste trabalho é aplicar técnicas de aprendizado de máquina em um banco de dados estruturado em pesquisas em Ciências Sociais para compreender o ativismo social. Desta forma, nesta pesquisa foram feito um levantamento de trabalhos com técnicas de Inteligência Artificial em uma base não estruturada para o estudo da sociedade, de pesquisas realizadas com a base do WVS e Ciências Sociais, e WVS e Inteligência Artificial. Além de desenvolver uma metodologia hibrida, com conhecimentos e técnicas de Inteligência Artificial e Ciências Sociais.

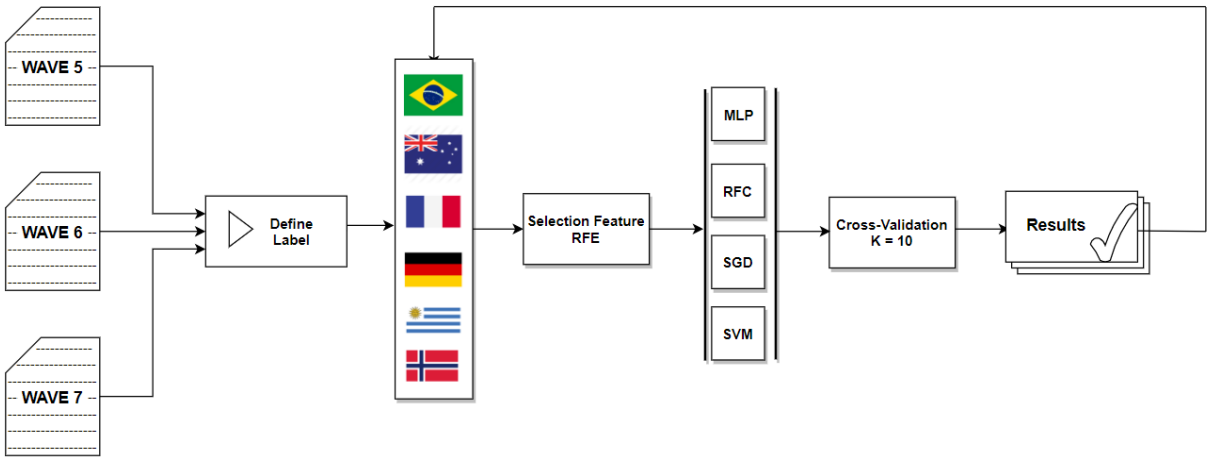
A principal contribuição desta dissertação é o uso de técnicas de IA em dados estruturados em Ciências Sociais, como bancos de dados de pesquisas, parece trazer novas perspectivas de análise. Observamos que informações que geralmente não são evidenciadas pelo uso de técnicas tradicionais de análise estatística podem aparecer. Encontramos padrões nas bases de dados utilizando diferentes análises em diferentes ondas que per-

mitem tanto questionar quanto aprofundar o conhecimento existente.

Definimos uma abordagem neste trabalho, partindo da ideia de que diferentes nacionalidades podem introduzir valores e comportamentos distintos em relação à cultura, política e processo de socialização. Desta forma, decidimos analisar separadamente cada país contido em cada uma das ondas.

O processo de análise e construção dos modelos seguiu as etapas mostradas na Figura A.1. Primeiro, para cada onda, foi definida o label da classe; em outras palavras, um conjunto de questões correspondentes à tarefa de classificar e analisar um conceito. Em seguida, aplicamos para cada país o processo de seleção de features, reduzindo assim a dimensionalidade. A seguir, utilizando esses recursos selecionados, aplicamos este conjunto de dados em quatro métodos de Aprendizado de Máquina, são eles, Support Vector Machine (SVM) (SMOLA; SCHöLKOPF, 2004), Random Forest Classifier (RFC) (BREIMAN, 2001), modelos lineares com Gradiente Estocástico Descent (SGD) (ZHANG, 2004), e uma rede neural Multi-layer Perceptron (MLP) (RUMELHART; HINTON; WILLIAMS, 1986). Por fim, geramos os resultados com a média da validação cruzada k-fold com K = 10.

Figure A.1: Visão geral da metodologia de análise dos dados



Além disso, desenvolvemos, a partir da teoria conceitual dos Clusters, uma classificação de cada questão selecionada pelo ranking global da RFE. Cada grupo temático foi definido com base na semântica da questão.

Para cada onda, criamos clusters com base nos principais assuntos que apareceram nas perguntas. Mesmo que os clusters não sejam iguais em todas as ondas, podemos verificar um padrão. Nas variáveis mais frequentes, os clusters denominados Família, Tolerância e Religião estão presentes em todas as ondas. Por outro lado, nas variáveis

de menor frequência, os clusters que estão presentes são os denominados Política, Moral, Religião e Ciência.

A pesquisa WVS foi construída com base nas teorias dos comportamentos políticos e sociais existentes nas chamadas democracias avançadas (Europa e EUA). Assim, as variáveis do conjunto de dados WVS aqui analisadas parecem ajustar-se melhor aos valores existentes nessas sociedades (como os resultados sugeriram). Em sociedades com diferentes construtos sociais e histórias, os dados da pesquisa WVS podem não refletir, pelo menos em relação à dimensão analisada (participação política), o comportamento dos indivíduos.

Um dos fatores que influenciam o comportamento humano são os valores, como exemplo, que motivam as ações. A compreensão desses valores pode ajudar a melhorar a sociedade a partir de medidas para superar os problemas sociais. As revoluções em nossas comunidades, por exemplo, a abolição da escravidão, o direito ao voto feminino, aconteceram com movimentos sociais de lideranças que conseguiram identificar os problemas e propor soluções.

Com o crescimento exponencial dos dados, os stakeholders precisam se apropriar desses trabalhos científicos interdisciplinares para entender os efeitos e as motivações relacionadas às ações, como por exemplo, a desinformação (notícias falsas). Dessa forma, os líderes podem atuar com uma estratégia política para criar mudanças sociais com sucesso.

Os resultados desta dissertação indicam que o caminho de pesquisa aqui utilizado é frutífero e deve ser aprofundado. Neste trabalho, podemos inferir que as causas que motivam o ativismo social estão mais relacionadas a causas sociais do que a propósitos políticos, de acordo com os resultados. No entanto, são necessários estudos mais aprofundados na área de IA, bem como análises na perspectiva das Ciências Sociais que possam ajudar a explicar os resultados e propor novas linhas de investigação. Além disso, esta pesquisa abre caminho para estudos interdisciplinares de IA e Ciências Sociais com ganhos para ambas as áreas.

Como trabalho futuro, propomos a aplicação de diferentes estratégias de seleção de features para comparar com os resultados e considerações encontrados nesta pesquisa. Além disso, aplicar a abordagem desenvolvida nesta dissertação com outros tópicos de interesse da sociedade para novos insights.

AppendixB FEATURES RANKING QUESTIONS

B.1 WAVE 5 features ranking questions

Table B.1: Ranking of frequency of questions in the countries of the Wave 5 - Most frequent

Attribute	Questions	Frequency
V9	Religion (Very important, rather important, not very important, not all important)	44
V186	Apart from weddings and funerals, about how often do you attend religious services these days? (Code one answer): 1 More than once a week 2 Once a week 3 Once a month 4 Only on special holy days 5 Once a year 6 Less often 7 Never, practically never	44
V19	Qualities that children can be encouraged to learn at home: Religious faith	40
V131	could you tell me how much confidence you have in them: The churches (great deal, quite a lot, not very much, none at all) all	35
V14	Qualities that children can be encouraged to learn at home: Feeling of responsibility	27
V12	Qualities that children can be encouraged to learn at home: Independence	25
V37	Could you please mention any that you would not like to have as Neighbors? : Immigrants/foreign workers	25
V40	Could you please mention any that you would not like to have as Neighbors? : Heavy drinkers	25
V41	Could you please mention any that you would not like to have as Neighbors? : Unmarried couples living together	25
V187	Independently of whether you attend religious services or not, would you say you are (read out and code one answer): 1 A religious person 2 Not a religious person 3 An atheist	25
V35	Could you please mention any that you would not like to have as Neighbors? : People of a different race	23
V34	Could you please mention any that you would not like to have as Neighbors? : Drug addicts	22

V36	Could you please mention any that you would not like to have as Neighbors? : People who have AIDS	22
V227	Books, (Used it last week, did not use it last week)	22
V21	Qualities that children can be encouraged to learn at home: Obedience	21
V39	Could you please mention any that you would not like to have as Neighbors? : People of a different religion	21
V42	Could you please mention any that you would not like to have as Neighbors? : People who speak a different language	21
V17	Qualities that children can be encouraged to learn at home: Thrift, saving money and things	20
V20	Qualities that children can be encouraged to learn at home: Unselfishness	20
V4	Family (Very important, rather important, not very important, not all important)	19
V5	Friends (Very important, rather important, not very important, not all important)	19
V13	Qualities that children can be encouraged to learn at home: Hard work	19
V18	Qualities that children can be encouraged to learn at home: Determination, perseverance	19
V23	Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people? (Code one answer): 1 Most people can be trusted, 2 Need to be very careful,	19
V38	Could you please mention any that you would not like to have as Neighbors? : Homosexuals	19
V193	Do you take some moments of prayer, meditation or contemplation or something like that? (Yes, no)	19
V218	Being born on my country's soil, (Very important, rather important, not important)	19
V224	News broadcasts on radio or TV, (Used it last week, did not use it last week)	19
V57	If someone says a child needs a home with both a father and a mother to grow up happily, would you tend to agree or disagree? (Code one answer): 1 Tend to agree 2 Tend to disagree	18

V79	I'm going to read out a list of various changes in our way of life that might take place in the near future:More emphasis on my family life (Good, don't mind, bad)	18
V134	could you tell me how much confidence you have in them: Television (great deal, quite a lot, not very much, none at all) all	17
V192	How important is God in your life? Please use this scale to indicate, 10 means "very important" and 1 means "not at all important,"	17
V219	Adopting the customs of my country, (Very important, rather important, not important)	17
V226	In tepth reports on radio or TV (Used it last week, did not use it last week)	17
V8	Work (Very important, rather important, not very important, not all important)	16
V65	I seek to be myself rather than to follow others, (Strongly agree, Agree, disagree, strongly disagree)	16
V113	Pollution of rivers, lakes and oceans,(Very serious, somewhat serious, not very serious, not serious at all)	16
V216	Are your mother or father immigrants to this country or not? Please, indicate separately for each of them (read out and code one answer for each): Father (Immigrant; Not an immigrant)	16
V229	Talk with friends or colleagues (Used it last week, dit not use it last week)	16
V10	Taking all things together, would you say you are, (Very happy, rather happy, not very happy, not at all happy)	15

Table B.2: Ranking of frequency of questions in the countries of the Wave 5 - Least frequent

Attribute	**Questions**	**Frequency**
V121	Now I'd like you to tell me your views on various issues, How would you place your views on this scale? 1 means you agree completely with the statement on the left; 10 means you agree completely with the statement on the right; and if your views fall somewhere in between, you can choose any number in between, (People can only get rich at the expense of others ; Wealth can grow so there's enough for everyone) - scale	3

V154	People choose their leaders in free elections, (Not an essential characteristic of democracy ; An essential characteristic of democracy) - scale	3
V155	People receive state aid for unemployment, (Not an essential characteristic of democracy ; An essential characteristic of democracy) - scale	3
V156	The army takes over when government is incompetent, (Not an essential characteristic of democracy ; An essential characteristic of democracy) - scale	3
V159	Criminals are severely punished, (Not an essential characteristic of democracy ; An essential characteristic of democracy) - scale	3
V160	People can change the laws in referendums, (Not an essential characteristic of democracy ; An essential characteristic of democracy) - scale	3
V163	And how democratically is this country being governed today? Again using a scale from 1 to 10, where 1 means that it is “not at all democratic” and 10 means that it is “completely democratic,” what position would you choose? (Not all democratic ; completely democratic) - scale	3
V201	Someone accepting a bribe in the course of their duties, (Never justifiable, always justifiable - scale)	3
V207	Suicide, (Never justifiable, always justifiable - scale)	3
V208	For a man to beat his wife, (Never justifiable, always justifiable - scale)	3
V47	Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair? Please show your response on this card, where 1 means that “people would try to take advantage of you,” and 10 means that “people would try to be fair” (People would try to take advantage of you, people would try to be fair) - scale	2
V86	Adventure and taking risks are important to this person; to have an exciting life, (very much like me, like me, somewhat like me, not like me, or not at all like me)	2
V91	Science and technology are making our lives healthier, easier, and more comfortable, (Completely disagree, completely agree) - scale	2

V93	Science and technology make our way of life change too fast, (Completely disagree, completely agree) - scale	2
V94	We depend too much on science and not enough on faith, (Completely disagree, completely agree) - scale	2
V114	In political matters, people talk of "the left" and "the right," How would you place your views on this scale, generally speaking? (Left, right) - scale	2
V117	Now I'd like you to tell me your views on various issues, How would you place your views on this scale? 1 means you agree completely with the statement on the left; 10 means you agree completely with the statement on the right; and if your views fall somewhere in between, you can choose any number in between, (Private ownership of business and industry should be increased, government ownership of business and industry should be increased) - scale	2
V118	Now I'd like you to tell me your views on various issues, How would you place your views on this scale? 1 means you agree completely with the statement on the left; 10 means you agree completely with the statement on the right; and if your views fall somewhere in between, you can choose any number in between, (The government should take more responsibility to ensure that everyone is provide for ; People should take more responsibility to provide for themselves) - scale	2
V137	could you tell me how much confidence you have in them: The courts (great deal, quite a lot, not very much, none at all) all	2
V153	Religious authorities interpret the laws, (Not an essential characteristic of democracy ; An essential characteristic of democracy) - scale	2
V157	Civil rights protect people's liberty against oppression,(Not an essential characteristic of democracy ; An essential characteristic of democracy) - scale	2
V162	How important is it for you to live in a country that is governed democratically? On this scale where 1 means it is "not at all important" and 10 means "absolutely important" what position would you choose? (Not all important; Absolutely important) - scale	2

V168	Which of these problems do you consider the most serious one in your own country? (Code one answer under “most serious for own country”): People living in poverty and need, Discrimination of girls and women, Poor sanitation and infectious diseases, Inadequate education, Environmental pollution,	2
V178	Thinking at your own country’s problems, should your country’s leaders give top priority to help reducing poverty in the world or should they give top priority to solve your own country’s problems? Use this scale where 1 means “top priority to help reducing poverty in the world” and 10 means “top priority to solve my own country’s problems,” (Code one answer): (Top priority to help reducing proverty in the world; Top priority to solve my own country’s problems)	2
V185	Do you belong to a religion or religious denomination? If yes, which one? (Code answer due to list below, Code 0, if respondent answers to have no denomination!) 0 - No: do not belong to a denomination 1 - Yes: Roman Catholic 2 - Protestant 3 - Orthodox (Russian/Greek/etc,) 4 - Jew 5 - Muslim 6 - Hindu 7 - Buddhist 8 - Other (write in):	2
V206	Euthanasia - ending of the life of the incurable sick (Never justifiable, always justifiable - scale)	2
V55	Are you currently (read out and code one answer only): 1 Married 2 Living together as married 3 Divorced 4 Separated 5 Widowed 6 Single	1
V72	If you had to choose, which one of the things on this card would you say is most important? (Code one answer only under “second choice”): Maintaining order in the nation Giving people more say in important government decisions Fighting rising prices Protecting freedom of speech	1
V74	Here is another list, In your opinion, which one of these is most important? (Code one answer only under “second choice”): A stable economy Progress toward a less impersonal and more humane society Progress toward a society in which Ideas count more than money The fight against crime	1
V158	The economy is prospering, (Not an essential characteristic of democracy ; An essential characteristic of democracy) - scale	1

V167	And which is the second most serious problem for the world as a whole? (Code one answer under “next most serious for the world”): People living in poverty and need, Discrimination against girls and women, Poor sanitation and infectious diseases, Inadequate education, Environmental pollution,	1
V198	Claiming government benefits to which you are not entitled, (Never justifiable, always justifiable - scale)	1
V199	Avoiding a fare on public transport, (Never justifiable, always justifiable - scale)	1
V202	Homosexuality, (Never justifiable, always justifiable - scale)	1
V203	Prostitution, (Never justifiable, always justifiable - scale)	1
V204	Abortion, (Never justifiable, always justifiable - scale)	1
V205	Divorce, (Never justifiable, always justifiable - scale)	1
V231	If there were a national election tomorrow, for which party on this list would you vote? Just call out the number on this card, If you are uncertain, which party appeals to you most? (Code one answer): 01 Party 1 02 Party 2 03 Party 3 04 etc,	1
V232	And which party would be your second choice? If you are uncertain, which one appeals you second most? (Code one answer): 01 Party 1 02 Party 2 03 Party 3 04 etc,	1
V233	And is there a party that you would never vote for? (Code one answer): 01 Party 1 02 Party 2 03 Party 3 04 etc,	1

B.2 WAVE 6 features ranking questions

Table B.3: Ranking of frequency of questions in the countries of the Wave 6 - Most frequent

Attribute	Questions	Frequency
V145	Apart from weddings and funerals, about how often do you attend religious services these days? 1 More than once a week, 2 Once a week, 3 Once a month, 4 Only on special holy days, 5 Once a year, 6 Less often, 7 Never, practically never	44
V9	Religion (Very important, rather important, not very important, not all important)	41
V108	The churches (A great deal, Quite a lot, Not very much, None at all)	38

V19	Qualities that children can be encouraged to learn at home: Religious faith (Mentioned, Not mentioned)	32
V227	National level (Always, Usually, Never)	27
V147	Independently of whether you attend religious services or not, would you say you are, 1 A religious person, 2 Not a religious person, 3 An atheist	27
V4	Family (Very important, rather important, not very important, not all important)	26
V44	People who speak a different language (Mentioned, Not mentioned)	24
V43	Unmarried couples living together (Mentioned, Not mentioned)	23
V24	Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people? (Most people can be trusted, Need to be very careful)	23
V22	Qualities that children can be encouraged to learn at home: Self-expression (Mentioned, Not mentioned)	23
V12	Qualities that children can be encouraged to learn at home: Independence (Mentioned, Not mentioned)	22
V41	People of a different religion (Mentioned, Not mentioned)	21
V39	Immigrants/foreign workers (Mentioned, Not mentioned)	21
V36	Drug addicts (Mentioned, Not mentioned)	21
V226	Local level (Always, Usually, Never)	21
V14	Qualities that children can be encouraged to learn at home: Feeling of responsibility (Mentioned, Not mentioned)	21
V10	Taking all things together, would you say you are, (Very happy, rather happy, not very happy, not at all happy)	21
V20	Qualities that children can be encouraged to learn at home: Unselfishness (Mentioned, Not mentioned)	20
V150	With which one of the following statements do you agree most?The basic meaning of religion is: 1 To follow religious norms and ceremonies, 2 To do good to other people	20
V148	Do you believe in God? 1 Yes, 2 No	20
V83	Participated in a demonstration for some environmental cause? 1 Yes, 2 No	19
V40	Homosexuals (Mentioned, Not mentioned)	19

V13	Qualities that children can be encouraged to learn at home: Hard work (Mentioned, Not mentioned)	19
V82	Given money to an ecological organization? 1 Yes, 2 No	18
V69	Greater respect for authority, (Good, Don't mind, Bad)	18
V37	People of a different race (Mentioned, Not mentioned)	18
V18	Qualities that children can be encouraged to learn at home: Determination, perseverance (Mentioned, Not mentioned)	18
V16	Qualities that children can be encouraged to learn at home: Tolerance and respect for other people (Mentioned, Not mentioned)	18
V146	Apart from weddings and funerals, about how often do you pray? 1 Several times a day, 2 Once a day, 3 Several times each week, 4 Only when attending religious services, 5 Only on special holy days, 6 Once a year, 7 Less often, 8 Never, practically never	18
V104	People you know personally (Trust completely, Trust somewhat, Do not trust very much , Do not trust at all)	18
V5	Friends (Very important, rather important, not very important, not all important)	17
V42	Heavy drinkers (Mentioned, Not mentioned)	17
V17	Qualities that children can be encouraged to learn at home: Thrift, saving money and things (Mentioned, Not mentioned)	17
V151	And with which of the following statements do you agree most? The basic meaning of religion is: 1 To make sense of life after death, 2 To make sense of life in this world	17
V110	The press (A great deal, Quite a lot, Not very much, None at all)	17
V102	Your family (Trust completely, Trust somewhat, Do not trust very much , Do not trust at all)	17
V152	How important is God in your life? Please use this scale to indicate, 10 means "very important" and 1 means "not at all important," (Not at all important, Very important) - scale	16
V15	Qualities that children can be encouraged to learn at home: Imagination (Mentioned, Not mentioned)	16
V149	Do you believe in hell? 1 Yes, 2 No	16

Table B.4: Ranking of frequency of questions in the countries of the Wave 6 - Least frequent

Attribute	Questions	Frequency
V199	Avoiding a fare on public transport, (Never justifiable,Always justifiable) - scale	5
V197	All things considered, would you say that the world is better off, or worse off, because of science and technology? Please tell me which comes closest to your view on this scale: 1 means that "the world is a lot worse off," and 10 means that "the world is a lot better off," (Code one number): (A lot worse off ,A lot better off) - scale	5
V194	We depend too much on science and not enough on faith (Completely disagree ,Completely agree) - Scale	5
V177	Which of the following things have you done for reasons of security? Preferred not to go out at night:1 Yes,2 No	5
V176	Which of the following things have you done for reasons of security? Didn't carry much money, 1 Yes, 2 No	5
V163	with respect? (Not at all likely to be viewed that way, Very likely to be viewed that way,(Don't know))	5
V159	people over 70? (Extremely low position in society, Extremely high position in society, (Don't know))	5
V141	And how democratically is this country being governed today? Again using a scale from 1 to 10, where 1 means that it is "not at all democratic" and 10 means that it is "completely democratic," what position would you choose? (Not at all democratic , Completely democratic) - scale	5
V139	Women have the same rights as men, (Not an essential characteristic of democracy, An essential characteristic of democracy) - scale	5
V134	People receive state aid for unemployment, (Not an essential characteristic of democracy, An essential characteristic of democracy) - scale	5
V133	People choose their leaders in free elections, (Not an essential characteristic of democracy, An essential characteristic of democracy) - scale	5
V117	Parliament (A great deal, Quite a lot, Not very much, None at all)	5

V72	Living in secure surroundings is important to this person; to avoid anything that might be dangerous,	4
V222	Email, (Daily, Weekly, Monthly, Less than monthly, Never)	4
V198	Claiming government benefits to which you are not entitled, (Never justifiable,Always justifiable) - scale	4
V196	It is not important for me to know about science in my daily life (Completely disagree ,Completely agree) - Scale	4
V192	Science and technology are making our lives healthier, easier, and more comfortable, (Completely disagree ,Completely agree) - Scale	4
V175	How frequently do the following things occur in your neighborhood? Drug sale in streets, 1 Very frequently, 2 Quite frequently , 3 Not frequently, 4 Not at all frequently, -1 DK/NA	4
V164	Please tell me how acceptable or unacceptable you think most people in [country] would find it if a suitably qualified 70* year old was appointed as their boss? (Completely unacceptable, Completely acceptable, DK) -scale	4
V160	Please tell me how acceptable or unacceptable you think most people in [country] would find it if a suitably qualified 30 year old was appointed as their boss? (Completely unacceptable, Completely acceptable) - scale	4
V158	people in their 40's? (Extremely low position in society, Extremely high position in society, (Don't know))	4
V144	Do you belong to a religion or religious denomination? If yes, which one? (Code answer due to list below, Code 0, if respondent answers to have no denomination!) No: do not belong to a denomination 0, Yes: Roman Catholic 1, Protestant 2, Orthodox (Russian/Greek/etc,) 3 Jew, 4 Muslim 5 Hindu 6 Buddhist 7 Other (write in):8	4
V138	People obey their rulers, (Not an essential characteristic of democracy, An essential characteristic of democracy) - scale	4
V135	The army takes over when government is incompetent,(Not an essential characteristic of democracy, An essential characteristic of democracy) - scale	4

V132	Religious authorities ultimately interpret the laws, (Not an essential characteristic of democracy, An essential characteristic of democracy) - scale	4
V101	People can only get rich at the expense of others , Wealth can grow so there's enough for everyone, - scale	4
V97	Private ownership of business and industry should be increased , Government ownership of business and industry should be increased - scale	3
V96	Incomes should be made more equal, We need larger income differences as incentives for individual effort - scale	3
V57	Are you currently :1 Married, 2 Living together as married, 3 Divorced, 4 Separated, 5 Widowed, 6 Single	3
V231	Are the tasks you do at work mostly manual or mostly intellectual? If you do not work currently, characterize your major work in the past, Use this scale where 1 means "mostly manual tasks" and,10 means "mostly intellectual tasks" (code one answer): (Mostly manual tasks, Mostly intellectual tasks) - scale	3
V229	Are you employed now or not? If yes, about how many hours a week? If more than one job: only for the main job (code one answer): Yes, has paid employment: 1 Full time employee (30 hours a week or more), 2 Part time employee (less than 30 hours a week), 3 Self employed ,No, no paid employment:4 Retired/pensioned ,5 Housewife not otherwise employed, 6 Student, 7 Unemployed , 8Other (write in):	3
V206	Divorce, (Never justifiable,Always justifiable) - scale	3
V195	One of the bad effects of science is that it breaks down people's ideas of right and wrong, (Completely disagree ,Completely agree) - Scale	3
V193	Because of science and technology, there will be more opportunities for the next generation (Completely disagree ,Completely agree) - Scale	3
V100	In the long run, hard work usually brings a better life , Hard work doesn't generally bring success—it's more a matter of luck and connections , - scale	3

V99	Competition is good, It stimulates people to work hard and develop new ideas, Competition is harmful, It brings out the worst in people, - scale	2
V233	How much independence do you have in performing your tasks at work? If you do not work currently, characterize your major work in the past, Use this scale to indicate your degree of independence where 1 means “no independence at all” and 10 means “complete independence” (code one answer): (No independence at all, Complete independence) -scale	2
V221	Mobile phone, (Daily, Weekly, Monthly, Less than monthly, Never)	2
V232	Are the tasks you perform at work mostly routine tasks or mostly creative tasks? If you do not work currently, characterize your major work in the past, Use this scale where 1 means “mostly routine tasks” and 10 means “mostly creative tasks” (code one answer): (Mostly routine tasks, Mostly creative tasks) - scale	1
V203	Homosexuality, (Never justifiable,Always justifiable) - scale	1

B.3 WAVE 7 features ranking questions

Table B.5: Ranking of frequency of questions in the countries of the Wave 7- Most frequent

Attribute	Questions	Frequency
Q64	For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all? The [churches]	14
Q15	Qualities that children can be encouraged to learn at home: Religious faith (Mentioned, Not mentioned)	13
Q6	Religion (Very important, rather important, not very important, not all important)	12
Q171	Apart from weddings and funerals, about how often do you attend religious services these days? (Code one answer): 1 More than once a week, 2 Once a week, 3 Once a month, 4 Only on special holy days,5 Once a year, 6 Less often, 7 Never, practically never	12
Q11	Qualities that children can be encouraged to learn at home: Imagination (Mentioned, Not mentioned)	12

Q221	When elections take place, do you vote always, usually or never? Please tell me separately for each of the following levels. Local level (Always, Usually, Never, Not allowed to vote)	10
Q18	Could you please mention any that you would not like to have as neighbors? Drug addicts (Mentioned, Not mentioned)	10
Q10	Qualities that children can be encouraged to learn at home: Feeling of responsibility (Mentioned, Not mentioned)	10
Q1	Family (Very important, rather important, not very important, not all important)	10
Q7	Qualities that children can be encouraged to learn at home: Good manners (Mentioned, Not mentioned)	9
Q3	Leisure time (Very important, rather important, not very important, not all important)	9
Q256	People have different views about themselves and how they relate to the world. Using this card, would you tell me how close do you feel to. . . ? Your [country, region, district] (Very close, close, not very close, not close at all)	9
Q23	Could you please mention any that you would not like to have as neighbors? People of a different religion (Mentioned, Not mentioned)	9
Q213	Donating to a group or campaign (Have done, Might do, Would never do)	9
Q20	Could you please mention any that you would not like to have as neighbors? People who have AIDS (Mentioned, Not mentioned)	9
Q16	Qualities that children can be encouraged to learn at home: Not being selfish (unselfishness) (Mentioned, Not mentioned)	9
Q145	Have you been the victim of a crime during the past year? And what about your immediate family – has someone in your family been the victim of a crime during the last year? Family (Yes No)	9
Q144	Have you been the victim of a crime during the past year? And what about your immediate family – has someone in your family been the victim of a crime during the last year? Respondent (Yes No)	9
Q141	Which of the following things have you done for reasons of security? (MULTIPLE RESPONSE) Carried a knife, gun or other weapon (yes, no)	9

Q82	For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all? The [European Union]	8
Q26	Could you please mention any that you would not like to have as neighbors? People who speak a different language (Mentioned, Not mentioned)	8
Q22	Could you please mention any that you would not like to have as neighbors? Homosexuals (Mentioned, Not mentioned)	8
Q21	Could you please mention any that you would not like to have as neighbors? Immigrants/foreign workers (Mentioned, Not mentioned)	8
Q19	Could you please mention any that you would not like to have as neighbors? People of a different race (Mentioned, Not mentioned)	8
Q149	Most people consider both freedom and equality to be important, but if you had to choose between them, which one would you consider more important? 1. Freedom, 2. Equality	8
Q139	Which of the following things have you done for reasons of security? (MULTIPLE RESPONSE) Yes No, Didn't carry much money	8
Q116	Among the following groups of people, how many do you believe are involved in corruption? Tell me for each group if you believe it is none of them, few of them, most of them or all of them? Civil service providers (police, judiciary, civil, servants, doctors, teachers)	8
Q9	Qualities that children can be encouraged to learn at home: Hard work (Mentioned, Not mentioned)	7
Q8	Qualities that children can be encouraged to learn at home: Independence (Mentioned, Not mentioned)	7
Q72	For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all? Political parties	7
Q53	In the last 12 months, how often have your or your family. . . ? Gone without medicine or medical treatment that you needed (Often, sometimes, rarely, never)	7
Q51	In the last 12 months, how often have your or your family. . . ? Gone without enough food to eat (Often, sometimes, rarely, never)	7

Q222	When elections take place, do you vote always, usually or never? Please tell me separately for each of the following levels. National level (Always, Usually, Never, Not allowed to vote)	7
Q214	Contacting a governmental official (Have done, Might do, Would never do)	7
Q200	When you get together with your friends, would you say you discuss political matters frequently, occasionally or never? 1 Frequently, 2 Occasionally, 3 Never	7
Q175	And with which of the following statements do you agree most? The basic meaning of religion is: 1. To make sense of life after death 2. To make sense of life in this world	7
Q165	Which, if any, of the following do you believe in? God (yes, no)	7
Q150	Most people consider both freedom and security to be important, but if you had to choose between them, which one would you consider more important?, 1. Freedom, 2. Security	7
Q14	Qualities that children can be encouraged to learn at home: Determination, perseverance (Mentioned, Not mentioned)	7
Q12	Qualities that children can be encouraged to learn at home: Tolerance and respect for other people (Mentioned, Not mentioned)	7

Table B.6: Ranking of frequency of questions in the countries of the Wave 7 - Least frequent

Attribute	**Questions**	**Frequency**
Q125	From your point of view, what have been the effects of immigration on the development of [this country]? Gives asylum to political refugees who are persecuted elsewhere (Agree, Hard to say, Disagree)	2
Q118	want to know about your experience with local officials and service providers, like police officers, lawyers, doctors, teachers and civil servants in your community. How often do you think ordinary people like yourself or people from your neighbourhood have to pay a bribe, give a gift or do a favor to these people in order to get the services you need? Does it happen never, rarely, frequently or always?	2
Q113	Among the following groups of people, how many do you believe are involved in corruption? Tell me for each group if you believe it is none of them, few of them, most of them or all of them? State authorities	2

Q92	Where are the headquarters of the International Monetary Fund (IMF) located? A) Washington DC, B) London, C) Geneva	1
Q91	Five countries have permanent seats on the Security Council of the United Nations. Which ones of the following is not a member? A) France, B) China, C) India	1
Q83	For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all? The United Nations	1
Q80	For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all? Women's organization	1
Q79	For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all? Environmental organizations	1
Q68	For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all? Labor unions	1
Q50	How satisfied are you with the financial situation of your household? Please use this card again to help with your answer. (Completely dissatisfied, completely satisfied) - scale	1
Q49	All things considered, how satisfied are you with your life as a whole these days? Using this card on which 1 means you are "completely dissatisfied" and 10 means you are "completely satisfied" where would you put your satisfaction with your life as a whole? (Completely dissatisfied, Completely satisfied) - scale	1
Q48	Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where 1 means "no choice at all" and 10 means "a great deal of choice" to indicate how much freedom of choice and control you feel you have over the way your life turns out. (No choice at all, A great deal of choice) - scale	1
Q44	I'm going to read out a list of various changes in our way of life that might take place in the near future. More emphasis on the development of technology (Good, Don't mind, bad)	1

Q41	How would you feel about the following statements? Work should always come first, even if it means less spare time (Agree strongly, agree, neither agree nor disagree, disagree, disagree strongly)	1
Q259	People have different views about themselves and how they relate to the world. Using this card, would you tell me how close do you feel to…? Your [World] (Very close, close, not very close, not close at all)	1
Q242	Religious authorities ultimately interpret the laws. (Not an essential characteristic of democracy, An essencial characteristic of democracy) - scale	1
Q240	In political matters, people talk of "the left" and "the right." How would you place your views on this scale, generally speaking? (left, right) - scale	1
Q238	Having a democratic political system. (very good, fairly good, fairly bad, very bad)	1
Q235	Having a strong leader who does not have to bother with parliament and elections. (very good, fairly good, fairly bad, very bad)	1
Q231	In your view, how often do the following things occur in this country's elections? Voters are threatened with violence at the polls (Very often, fairly often, not often, not at all often)	1
Q220	Now I'd like you to look at this card. I'm going to read out some other forms of political action that people can take using Internet and social media tools like Facebook, Twitter etc., and I'd like you to tell me, for each one, whether you have done any of these things, whether you might do it or would never under any circumstances do it. Organizing political activities, events, protests (Have done, might do, would never do)	1
Q216	Encouraging others to vote (Have done, Might do, Would never do)	1
Q208	Talk with friends or colleagues (Daily, Weekly, Monthly, Less than monthly, Never)	1
Q207	Social Media (Facebook, twitter) (Daily, Weekly, Monthly, Less than monthly, Never)	1
Q206	Internet (Daily, Weekly, Monthly, Less than monthly, Never)	1
Q205	Email (Daily, Weekly, Monthly, Less than monthly, Never)	1

Q191	Violence against other people (Never justifiable, Always justifiable) - scale	1
Q187	Suicide (Never justifiable, Always justifiable) - scale	1
Q186	Sex before marriage (Never justifiable, Always justifiable) - scale	1
Q183	Prostitution (Never justifiable, Always justifiable) - scale	1
Q181	Someone accepting a bribe in the course of their duties (Never justifiable, Always justifiable) - scale	1
Q180	Cheating on taxes if you have a chance (Never justifiable, Always justifiable) - scale	1
Q161	one of the bad effects of science is that it breaks down people's ideas of right and wrong. Completely disagree Completely agree - scale	1
Q160	We depend too much on science and not enough on faith. Completely disagree Completely agree - scale	1
Q159	Because of science and technology, there will be more opportunities for the next generation. Completely disagree Completely agree - scale	1
Q154	If you had to choose, which one of the things on this card would you say is most important? (Code one answer only under "first choice"): Maintaining order in the nation 1,2. Giving people more say in important government decisions 2 2 3. Fighting rising prices 3,4. Protecting freedom of speech	1
Q152	Would you please say which one of these you, yourself, consider the most important? 1. A high level of economic growth 1 , 2. Making sure this country has strong defense forces 2 3. Seeing that people have more say about how things are done at their jobs and in their communities 3 3 4. Trying to make our cities and countryside more beautiful	1
Q147	To what degree are you worried about the following situations? A terrorist attack (Very much, a good deal, not much, not at all)	1
Q143	To what degree are you worried about the following situations? Not being able to give my children a good education (Very much, a good deal, not much, not at all)	1

Q121	Now we would like to know your opinion about the people from other countries who come to live in [your country] - the immigrants. How would you evaluate the impact of these people on the development of [your country]? Very good, Quite good ,Neither good,nor bad,Quite bad, Very bad	1