


Understanding voting behavior in the Portuguese general elections from 2002 to 2019: a machine learning approach

Laura dos Reis Gonçalves



Dissertation

Master in Modelling, Data Analysis and Decision Support Systems



Supervised by
Prof. Doutor Patrício Costa
Prof. Doutor João Gama



2021

Acknowledgments

I would like to thank specially to Prof. Doutor Patrício Costa that came up with the subject of this dissertation and was kind enough to share it with the students. Also, express my gratefulness to him and Prof. Doutor João Gama that were my supervisors. They were always available to guide and help me throughout this work and to revise the document.

I would also like to thank the institution, FEP, for providing the master and the conditions to teach it, even during an atypical time. A particular acknowledgment to the MADSAD teachers that taught me a lot and made me realize that this master was the correct choice for me.

On a personal level, I would like to express enormous gratitude to my parents, Isabel and José, that supported me all over this adventure. Last but not least, a special acknowledgment to my boyfriend, Luís, who supported me emotionally and kept me on my toes in hard times.

Overall, big thanks to all who supported me and made this journey achievable!

Resumo

Esta dissertação tem como objetivo perceber o comportamento eleitoral dos portugueses nas eleições legislativas, através do uso de *Machine Learning*, nas suas duas vertentes: Turnout (refere-se à decisão do acto de votar) e Decision (refere-se à decisão “em quem votar”). Os dados utilizados neste trabalho provêm de seis inquéritos pós-eleitorais realizados pelo Instituto de Ciências Sociais da Universidade de Lisboa, juntamente com outros parceiros, entre 2002 e 2019. Foram aplicados seis algoritmos de Redes Bayesianas (*Bayesian Search*, *Naïve Bayes*, *Greedy Thick Thinning*, *Prototypical Constraint-based*, *Augmented Naïve Bayes* e *Tree Augmented Naïve Bayes*). Os modelos de previsão foram utilizados para prever os dados de 2015 e 2019, utilizando sempre dados do passado para prever o futuro. Estes modelos foram posteriormente comparados através de medidas de avaliação de performance (*Accuracy*, *Precision*, *Recall*, *Sensistivity*, *Specififcity*, *F-score* e *Area Under the ROC curve*), com a ajuda de testes estatísticos (*Friedman* e *Nemenyi*) e do cálculo da *Critical Difference* e criação do respetivo diagrama. O melhor predictor para cada vertente foi selecionado e explorado. No caso do Turnout o model selecionado foi o criado utilizando o algoritmo *Augmented Naïve Bayes*. Foi observado que as variáveis que se mostravam como maiores predictoras da variável alvo eram *Party proximity*, *Political interest*, *Frequency of attending to religious services* e *Frequency of consulting news through papers*. No caso da *Decision*, o model escolhido foi o *Greedy Thick Thinning*. As variáveis que se revelaram com maior efeito predictor eram *Party proximity* e *Syndicate member*.

Palavras-chave: *Turnout*, *Decision*, *Machine Learning*, *Comportamento eleitoral*, *Redes Bayesianas*

Abstract

This dissertation aims to understand the Portuguese voting behavior in general elections, through Machine Learning, from both perspectives: Turnout (refers to the decision to participate in the election) and Decision (refers to the decision on whom to vote for). The data used in this work come from six post-electoral surveys performed by *Instituto de Ciências Sociais da Universidade de Lisboa*, alongside different partners, between 2002 and 2019. Six Bayesian Networks algorithms were applied (Bayesian Search, Naïve Bayes, Greedy Thick Thinning, PC, Augmented Naïve Bayes e Tree Augmented Naïve Bayes) to create the predictive models that were used to predict 2015 and 2019 election outcomes, using past data to predict the future. These models were then compared through performance evaluation measures (Accuracy, Precision, Recall, Sensitivity, Specificity, F-score e Area Under the ROC curve), alongside with statistical tests (Friedman and Nemenyi) and the calculation of the Critical Difference between them and the creation of the respective diagram. The best predictor for each perspective was selected and explored. In the Turnout case, the chosen model was the one created using Augmented Naïve Bayes. We observed that the variables that most influenced the outcome of the target variable were Party proximity, Political interest, Frequency of attending religious services, and Frequency of consulting news through papers. In the Decision case, the best predictor was the model using Greedy Thick Thinning. The variables that showed the most effect on the target variable were Party proximity and Syndicate membership.

Keywords: *Turnout, Decision, Machine Learning, Voting behavior, Bayesian Networks*

Index

1	Introduction	1
1.1	Problem Definition.....	2
2	Literature Review.....	3
2.1	Turnout	3
2.2	Decision	6
2.3	Literature review conclusions	9
3	Data and Methodology	11
3.1	Data.....	11
3.1.1	Original Data.....	11
3.1.2	Data Treatment	15
3.2	Methodology.....	21
3.3	Software	25
4	Results and discussion	26
4.1	Turnout	26
4.1.1	Validation of Turnout models.....	26
4.1.2	Turnout model exploration.....	28
4.1.3	Diagnosis for Turnout states	30
4.1.4	Turnout models conclusions	32
4.2	Decision	32
4.2.1	Validation of Decision models.....	32
4.2.2	Decision model exploration.....	35
4.2.3	Diagnosis for Decision states	39
4.2.4	Decision models conclusions	40
5	Conclusions.....	41
	References	43
	Appendixes	47
	Appendix 1 - Original Dataset variables description	47
	Appendix 2 - Final Dataset variables description.....	51

List of Tables

Table 1 - Overview of some papers and their main conclusions	10
Table 2 - Operationalization of turnout.....	12
Table 3 - Operationalization of decision.....	13
Table 4 - Characterization of the sample for the turnout question – original data	14
Table 5 - Characterization of the sample for the decision question – original data	15
Table 6 - Age distribution of the sample - original data	15
Table 7 - Missing values count per variable.....	16
Table 8 - Operationalization of variables.....	17
Table 9 - Count of missing values per individuals.....	18
Table 10 – Count of missing values per variable and decision taken to replace them.....	19
Table 11 - Characterization of the sample for the turnout question – treated data	20
Table 12 - Characterization of the inquired for the decision question – treated data.....	21
Table 13 - Age distribution of the sample - treated data	21
Table 14 - Evaluation measures from Turnout predictive models	26
Table 15 - Weight of influence of variables in Turnout	29
Table 16 - Voter's profiles regarding Turnout.....	32
Table 17 - Evaluation measures from Decision predictive models.....	33
Table 18 - Evaluation measures from Decision predictive models for the year of 2019 using data from 2002 to 2011.....	33
Table 19 - Weight of influence of variables in Decision.....	36
Table 20 - Voter's profiles regarding Decision.....	40

List of Figures

Figure 1 - Work pipeline.....	11
Figure 2 - Prediction model phases.....	23
Figure 3 - CDD for 2015 prediction turnout models	27
Figure 4 - CDD for 2019 prediction turnout models	27
Figure 6 - Turnout predictive model	28
Figure 7 - Turnout predictive model with strength of influence using Euclidean distance and normalized arcs	29
Figure 8 - Sensitivity analysis for turnout model chosen.....	30

Figure 9 - Turnout diagnosis for 'Did_not_voted' state	31
Figure 10 - Turnout diagnosis for 'Voted' state	31
Figure 11 - CDD for 2015 prediction decision models	34
Figure 12 - CDD for 2019 prediction decision models	34
Figure 14 - Decision predictive model	35
Figure 15 - Decision predictive model with strength of influence using Euclidean distance and normalized arcs.....	36
Figure 16 - Sensitivity analysis for decision model chosen.....	37
Figure 17 - Decision predictive model without Party Proximity variable	38
Figure 18 - Sensitivity analysis for decision model without Party Proximity.....	38
Figure 19 - Decision diagnosis for 'Right' state.....	39
Figure 20 - Decision diagnosis for 'Right' state.....	40

1 Introduction

Portuguese democratic system was installed after the so-known Bloodless Carnation Revolution on April 25th of 1974. Portuguese first elections were held precisely one year later, obtaining a 92% turnout, and another year later, the first parliamentary elections also took place. The Portuguese constitution was then built. Nowadays, there are four types of elections: presidential elections every five years, government and parliament elections, European elections, and county elections, which occur every four years.

In democratic countries, it became essential to study voting behavior to understand what motivates turnout, determine the factors that influence the party decision, and understand if there are predominant characteristics that explain voters' choices. Voting behavior can be analyzed from two different perspectives: Turnout and Decision. While Turnout refers to the act of voting itself and can therefore be considered a binary variable as the option are turn out to vote or not to turn out to vote; Decision refers to the party choice, to whom or which party/coalition an individual will vote for, that has multiple outcomes.

This dissertation aims to understand the Portuguese voting behavior in the general elections. Analyzing the literature regarding both perspectives (Turnout and Decision) and crossing the information found regarding the main influencers with the data available from Portuguese elections, the goal is to uncover the main drivers for the voting decisions.

Data from six post-electoral surveys in Portugal between 2002 and 2019 from the Portuguese general (government and parliament) elections generate the databases. These are used to analyze electoral behavior. It is important to note that as voting behavior is analyzed from two different perspectives, there is the need to work on the data and adapt the datasets for each case.

This work contains different phases: firstly, an exploratory data analysis is performed to study the data and see the evolution in voting behavior over the years. In this step, data is treated to build a dataset for the next steps. It also allowed to find a tendency or a year where the behavior changed and try to find possible motives, framing into the Portuguese economic and political context; secondly, a model was built to predict voting. Different prediction algorithms were applied using Bayesian Networks. These models were then compared

through evaluation measures such as Accuracy, Precision, Recall, Specificity, F-score, and the Area under the curve (AUC) derived from the Receiver Operating Characteristic curves (ROC curve), Friedman's test, Nemenyi's test and Critical Difference. Bayesian Networks were also used to make iterations to draw voter's profiles using diagnosis.

The primary motivation for this work is to use several methods and software learned during the course and apply them in a real context; also, being elections a close topic to everyone's life and an important decision. It becomes crucial to be studied and discover what influences an individual on its voting behavior, not only to understand the reason behind the action but also to create and define some strategies that could increase turnout.

1.1 Problem Definition

Voting behavior is a sensitive topic that has been studied over the years to find voters' intrinsic and extrinsic motivations. The factors that influence the Turnout and the Decision can be economic, political, social, ideology, among others. The individual characteristics that influence people to their decision can be demographic, economic, or even work-related.

There has been some research done on this topic in Portugal. However, the majority use statistical methods, researchers find relations and influences, but there is a lack of usage of this information in more practical works.

That way, it seemed essential to fill in that gap with a work that uses statistical methods and the literature available to perform an analysis using machine learning methods that can be used and applied in future elections.

2 Literature Review

This section includes the literature review regarding voting behavior for both perspectives that are being analyzed. Firstly, the literature is analyzed for Turnout, and secondly, it is analyzed for the Decision.

2.1 Turnout

Turnout refers to the probability of an individual participating in the elections: turn out to vote or not. This topic has been analyzed in the latest years, mainly to uncover the main factors influencing turnout. The determinants of voter turnout in national versus subnational elections have been analyzed by Cancela and Geys (2016) using the 83 studies used previously by Geys (2006) and adding more 102 studies published between 2002 and 2015 aggregate-level data. The variables were split into three groups: socio-economic, political and institutional. The conclusion withdrawn from this study was that some political factors like campaign expenditures and election closeness, and one institutional determinant (registration requirements) have more explanatory power in national elections. In contrast, socio-economic factors (population size, concentration, stability, income and ethnic homogeneity, proportion of minorities and past turnout) and some institutional variables such as concurrent election and the electoral system, play a more critical role in explaining turnout in subnational elections.

Moreover, some works analyze some specific aspects and their impact on turnout, such as citizens' motivation to vote. It was proposed by Gerber, Green and Larimer (2008) that for predicting turnout, citizens' utility from performing their civic duty should be considered. To analyze this question, they divide the motivation into intrinsic satisfaction and extrinsic incentives to comply and answer it. They experimented with the August 2006 primary election in Michigan. The evidence from this experiment showed that a substantially higher turnout was observed among those who received communications promising to make public their turnout to their household or neighbors, which means that the extrinsic incentives to comply showed higher weight on the turnout decision. In Portugal's case, it was concluded by Barros (2017), using data from a survey, that the Portuguese give high importance to the duty of voting. The researcher revealed that the sense of responsibility given by voting increases the probability of Portuguese voters' turnout rather than the will to affect the electoral results.

Besides the utility, a different view on turnout was considered by Grönlund and Setälä (2007). They studied how citizens' evaluations of the political system and its actors affect their propensity to vote using empirical evidence from the first round of the European Social Survey (ESS) collected simultaneously in 22 European countries between 2002 and 2003 where the trust was measured on a scale from 0 to 10 where 0 meant "no trust at all" and 10 implied "complete trust", as expected, it was concluded that trust in parliament positively impacts turnout and satisfaction in democracy.

Despite this, satisfaction with democracy is not enough to explain turnout. Individuals who are satisfied with democracy tend not to vote if they do not trust the electoral system (Birch 2010). Furthermore, the perception of fairness in the elections is studied based on aggregated-level data from elections in 31 countries between 1996 and 2002, including established, new and partial democracies.

Other important factors that can influence turnout are economic factors, as these factors can have either a withdrawal or a motivation effect. Using data from the 2002 elections, it was demonstrated that the unemployment rate and real wages have a non-linear effect on turnout by Martins and Veiga (2013). The unemployment rate can have both a withdrawal and motivation effect depending on its value. It was demonstrated that if the unemployment rate is higher than 7.5%, it has a motivation effect, while if it is below 7.5%, the opposite happens.

Besides analyzing the more generic issues of democracy, and their impact on turnout, the effect of party leaders' evaluation needs to be considered as well. Ferreira da Silva and Costa (2019) developed a study with data from seven European countries using post-electoral national elections surveys. They divided the leaders' characteristics into two groups: competence and warmth, where the first one includes traits as being assertive and strong, capable of governing the country well and making decisions, while the second group contains characteristics such as being a good communicator, having charisma and close to the inquired ideas. It is concluded that the warmth dimension has a significant impact on turnout for all individuals. In contrast, the competence dimension impacts mainly on those who were abstentionists in the previous election.

The literature previously mentioned has in common that it analyzes the impact on a voter turnout of very subjective variables that are factors of opinion and not factors that

characterize the individuals, as demographic factors. Age, age squared, education, residential mobility, religion, media exposure, mobilization, vote in the previous election, party identification, political interest, and political knowledge present a consistent turnout effect (Smets and van Ham 2013). To reach this conclusion, the authors reviewed 90 studies related to voter turnout from the first decade of 2000. They aggregated what was shown to be the more commonly used variables to explain voter turnout, and the variables mentioned above were used in 10% or more of the studies analyzed.

Also, Sobbrío and Navarra (2010) decided to study electoral participation and communicative voting in 14 European countries to analyze individuals' turnout based on their characteristics, level of information, and expressive motivations. Results show that individuals who are either independent or uninformed are less likely to turn out to vote. Furthermore, the probability of turnout is lower within individuals in an environment with a lower level of media freedom. The probability of abstention is lower for individuals whose ideology is closer to a likely winner party.

The previous works showed that there is a multitude of factors that can influence turnout. They can be either micro-level, characteristics of an individual, macro-level, characteristics of the electoral system, or even motivational, external factors that influence the voter's perspective on democracy efficiency, such as political efficacy, political interest, political trust, and satisfaction politics. The work done by Hadjar and Beck (2010) analyzes the effect of components of both levels in turnout. The authors conclude that the micro-level characteristics of the more susceptible individuals to not voting are people with a low level of education and younger voters. Regarding motivation, as expected, the lack of political interest, political efficiency, political trust, and dissatisfaction with politics increase the probability of an individual do not turn out to vote.

In the Portuguese case, age, frequency of attending religious services, social class, and income were proven to influence turnout by Cancela and Magalhães (2020). The authors analyzed turnout using post-electoral national elections surveys in Portugal between 2002 and 2019. These results are aligned with Smets and van Ham's (2013). Moreover, Costa (2020) also concluded that political interest and party identification represented the major influence on voting behavior.

Besides the factors mentioned above, other external determinants present on voter's daily basis must be considered, such as social networks. It was concluded by Magalhães (2008) that social networks have a significant influence on the decision to vote. This conclusion was taken by analyzing the effect of social networks on Portugal's electoral participation using data from two post-electoral questionnaires from 2005 and 2006 and later on supported by Falck, Gold and Heblich (2014) that found a small negative effect of the internet in turnout when comparing data from German's voting behavior before 2000 (pre-internet era) and after 2004 (post-internet era). Besides these, the research done by Jones, Bond, Bakshy, Eckles and Fowle (2017) also supported the ideal that social influence has a positive effect on turnout. Using the U.S. Presidential Election from 2012, the authors concluded that the direct recipient of the message is influenced, and the close ones.

2.2 Decision

Decision refers to whom or which party/coalition an individual will vote. It can be analyzed regarding the several parties in the election or with a left-right wings scheme, as the parties differ over the years. In the study developed by Sobbrío and Navarra (2010), electoral participation and communicative voting in 14 European countries were analyzed. Researchers concluded that left-wing extremists are more likely to vote for their most preferred party regardless of whether this party is a loser. In contrast, right-wing extremists are very strategic, which means that they are more likely to vote for a likely winner party showing that they care about the current decision making; better-educated people are more likely to vote as communicating rather than strategically, meaning that it is expected for them to vote for a sure loser party, indicating that they care more about future elections. The aim is to pass on a message. The probability of voting as communicating is higher in lower media freedom and a lower opposition party concentration.

The media's influence on the political perceptions, attitudes and voting behavior was studied by Javaid and Elahi (2014) using data from 200 Pakistan citizens from either rural and urban areas. They concluded that most people who live in rural areas vote based on personality, while voters who live in urban areas tend to vote based on performance and policy.

During electoral campaigns, media is filled with information that can swing the votes, and one of them is polls. It was concluded by Magalhães (2013) that those exposed to poll information become more likely to support candidates and parties expected to win in the elections. This conclusion was withdrawn from a study regarding the impact of polls in Portugal using data from the 2002 post-electoral study.

Besides media, the daily interactions happening might have an impact on voting behavior. According to Foladare (2019), living in a neighborhood with high concentrations of people of the same status will accentuate that status's effect as a political behavior source. The effect of social interaction in the neighborhood on people's tendencies to join specific parties and vote for specific candidates was studied by making a test of the clustering effect from a pre-election study of Buffalo's city in 1960. In recent years social networks have been a hot topic to be studied in the most diverse areas. It is known that social networks impact the user's life and its opinions, so it became important also to analyze their impact on voting behavior. According to Magalhães' (2008) study where he aborded this topic using two post-electoral questionnaires from 2005 and 2006 in Portugal, it was concluded that although social networks impact many behaviors, the voting decision is not one of them. More recently, a study from Biswas, Ingle and Roy (2014) concluded otherwise. Using data from Indian election and voters, they concluded that social media has a clear impact on younger voters. Political parties will influence more easily voter from metros cities and semi urban cities.

With the emerging of social networks and social media, one of the concerns has fallen to fake news, as they spread faster, and the fact confirmation does not happen in many cases. Following this concern, Allcott and Gentzkow (2017) developed research where they studied the effect of social media and fake news in the American 2016 election, as this was a massive polemic topic. Although fake news is more widely spread and believed in, there was no evidence that it impacted the election outcome. In the opposite direction, Biswas, Ingle and Roy (2014) concluded that social media impacts Indian voters, mainly young voters, as they usually get their information on social media. It was also shown that young voters tend to vote for digitally interactive parties and that the conversations in political forums influence the female more than male voters. It was also concluded by Cantarella, Fraccarola and Volpe (2020) when studying the effect of fake news in the 2018 Italian elections that exposure to fake news affects the voting decision. Moreover, it was found that this type of news favors

populist parties and that fake news is positively associated with prior support of these parties, which suggests a self-selection mechanism.

Besides the effect of external factors in voter's decision, the individual's specific characteristics as income and other demographic factors are also studied. It was concluded by Leigh (2011) that the individuals with characteristics such as are poor, foreign-born, younger voters, voters born since 1950, men and unmarried were shown to be more prone to vote for left-wing parties. On the opposite side, people who live in wealthier neighborhoods are more likely to vote for right-wing parties. These conclusions were based on post-election surveys of 14 000 voters in 10 Australian elections between 1966 and 2001. It was also noted that the partisan gap between men and women has closed. However, it has widened between young and old, rich and poor, and native-born and foreign-born.

Regarding the influences on voting decisions, specifically in Portugal, Costa and Ferreira da Silva (2015) studied the impact of party leaders' evaluation in voting behavior using post-electoral national surveys conducted from 2002 to 2013 in seven European countries. Given that the data used included many different countries, it could not be analyzed on a party-wise view, so the 0-10 left-right scale was used. It was concluded that Portuguese voters seem more prone to vote for right-wing parties than Germans. However, when party identification ideology is introduced, it is observed that Portugal's signal changes from negative to positive, which means that individuals who voted for right-wing parties identify themselves with left-wing parties, showing a contradiction between party identification and decision.

Regarding the characteristics, it was concluded that dimension competence was only decisive for left-wing party leaders. Party leaders or even party evaluations are typically related to the government or opposition's performances. Government performance is more easily measurable as some of their policies have a direct financial or social impact on individual life. Most of the voters' evaluations are based on these. It turns out to be essential to analyze the Portuguese government's expenditures, as it reflects the policy adopted and can even show opportunism if it differs near elections. Castro and Martins (2016), using monthly spending for the main categories and annual data, concluded that Portuguese governments tend to act opportunistically near the elections. Furthermore, general public services, social protection, and health care are more likely to be increased during the election period.

Similarly, as Turnout, the voting decision can also be determined by some demographic characteristics. Religion, social class, region, and union membership were proven to be characteristics that influence the decision (Jalali 2003). The author explored the electoral behavior of Portuguese based on previous literature data. It was also noted that over the years, the left-right self-placement has become more centrist, with fewer voters positioning themselves on the extremes. More recently, a similar study was developed with data from post-electoral studies between 2002 and 2019 in Portugal, where the characteristics to be found to have explanatory power on the decision were age, frequency of attending to religious events, social class, and schooling level (Cancela and Magalhães 2020). Although the last two studies mentioned are more than ten years apart, both show similar conclusions, having characteristics such as religion and social class prevailing.

Besides voters' demographic characteristics and evaluations, it was proven by Lewis-Beck and Lobo (2011) that ideological identification and party identification are also important factors that impact the voting decision and need to be the object of a more in-depth analysis. In this research, the authors also concluded that ideology identification has about twice the impact on voting decisions than party identification, which means voters tend to decide the party they will vote for based on the party ideology more than on the party itself. Supported on these facts and using post-electoral data from Portuguese elections in 2005 and 2009 and a survey from 2006, Freire (2013) argued that voters who tend to vote for left-wing parties are individuals who give more importance to social issues such as income equality, prefer a higher state intervention and are also more liberal in their lifestyles. Following this idea, it was also shown by Veiga and Gonçalves Veiga (2004) that, typically, left-wing parties are more penalized for increases in the unemployment rate, while right-wing parties tend to be more punished by higher inflation. Using data from monthly polls on voting intentions in Portugal's elections between 1986 and 2001, the authors also concluded that Portuguese voters tend to support their decisions on past evaluations and experiences and current economic conditions rather than on desired future economic effects.

2.3 Literature review conclusions

Researchers have analysed both Turnout and Decision, where some conclusions can be drawn as for the variables that most influence voting behaviors. Table 1 summarizes some

relevant studies and their main findings regarding the variables that most influence the voting behavior and in which sense.

PERSPECTIVE	AUTHORS	YEAR	MAIN CONCLUSIONS
TURNOUT	Cancela and Magalhães	2020	Factors that positively impact turnout: age, frequency of attending religious services and income. Ex: elderly are more prone to turn-out to vote
	Smets and van Ham	2013	Factors that positively impact turnout: age, age squared, education, home ownership, religion, media exposure, vote in the previous election, party identification, political interest, and political knowledge. Ex: individuals that are most interested in politics are more likely to turn out to vote
	Costa	2020	Political interest and party identification represented the major positive influence on voting behavior.
DECISION	Jalali	2003	Factors that positively impact decision: religion and social class Ex: Individuals who are more capitalist (belong to a higher social class) tend to vote for right-wing parties.
	Cancela and Magalhães	2020	Factors that positively impact decision: age, frequency of attending to religious events, social class, and schooling level Ex: as the age of voters increases, so does their propension to vote for right-wing parties
	Lewis-Beck and Lobo	2011	Ideological identification and party identification are also important factors that impact the voting decision. Ex: Individuals tend to vote for a party when they identify themselves with their ideology or are proximate to it.

Table 1 - Overview of some papers and their main conclusions

3 Data and Methodology

This work used data from post-electoral surveys in Portugal between 2002 and 2019. It is divided into three main parts: Exploratory data analysis, Predictive model build and Definition of voter's profile.

Figure 1 provides a scheme of the project tasks to be performed. Firstly, given that the data come from six independent post-electoral surveys, there is the need to gather and aggregate all data into one database. It is needed to analyze all questions, see which ones are common to all six surveys, and recode the answers. The next step is to treat the data: reduce variables and observations to maintain only the ones needed and replacing missing data. After having these two main tasks done, the data is ready to be used for the Data Analysis, build the Predictive model, and define the voter's profile.

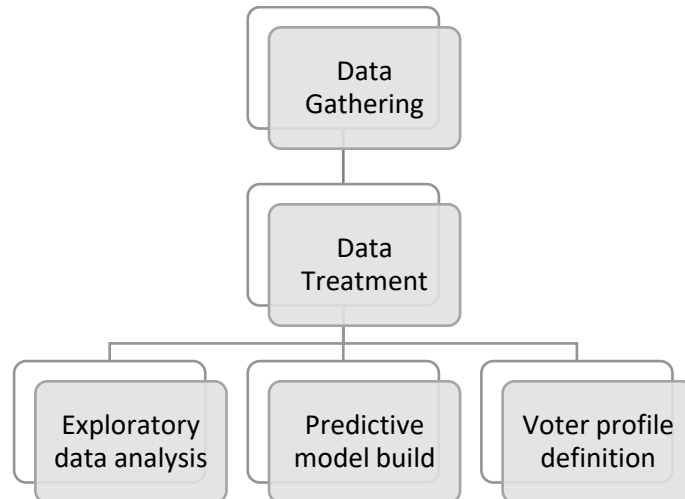


Figure 1 - Work pipeline

3.1 Data

3.1.1 Original Data

The data used in this thesis are post-electoral surveys between 2002 and 2019 (2002, 2005, 2009, 2011, 2015, and 2019) in Portugal made by researchers of the *Instituto de Ciências Sociais da Universidade de Lisboa* alongside different partners. The questions made in these surveys can be resumed into five groups: demographic, economic, evaluation, ideologic, and voting behavior. Regarding the number of participants in these studies, it was always around 1 000 participants per year, except for 2005 where the number of participants almost reached

3 000 participants. The number of participants in the last two years remained stable at 1 500, resulting in a dataset with 9 420 participants.

Electoral behavior was analyzed from two different perspectives, though the same dataset was used for both. The target variable not understudy was ignored for each perspective. This is due to the two objective variables (Turnout and Decision) being related with each other and having a causality relation. If an individual votes for a right-wing party, they turn out to vote. If an individual turns out to vote, it would vote for a specific party. Given this fact, it would be redundant having both variables in the dataset as it could also bias the analysis. The initial dataset contains 37 variables related to the party decision and/or the turnout (variable description in Appendix 1).

In the Turnout variable, there were originally four different options of answer and the purpose of this research to predict if the individual turns out to vote or not. For that reason, the original variable was recoded to have only two options as per Table 2.

Original Value	Final Value
1- Did not voted because he/she could not	0- Did not voted
2- Though of voting, but did not do it this time	
3- Usually votes, but did not do it this time	
4- Voted	1- Voted

Table 2 - Operationalization of turnout

As for the Decision, the recoding was more complex as the answer options were the different parties that have competed in the elections. Being an essential fact that the analysis was done in a time frame of eighteen years containing six elections, it is crucial to note that the parties that ran in the elections changed over the years. PS and PSD, the main parties always remained, but several small parties disappeared, and others have emerged. For that, it was decided to analyze the decision in a left-right wings scheme. So, the observations with the options of center or syncretic parties were considered ‘Other’.

In Table 3 the operationalization of this recoding and which parties were considered left, right, and other is presented. (source: [The Impact of Voter Evaluations of Leaders’ Traits on Voting Behaviour: Evidence from Seven European Countries](#). Costa and Ferreira da Silva 2015).

Party	Acronym	Spectrum	New Code
Bloco de Esquerda	B.E.	Left	1
CDS – Partido Popular	CDS-PP	Right	2
Coligação Democrática Unitária	CDU	Left	1
Partido Comunista dos Trabalhadores Portugueses	PCTP/MRPP	Left	1
Partido Social Democrata	PPD/PSD	Right	2
Partido Socialista	PS	Left	1
Pessoas-Animais-Natureza	PAN	Syncretic	998
Iniciativa Liberal	IL	Centre	998
Movimento Esperança Portugal	MEP	Centre	998
Movimento Mérito e Sociedade	MMS	Centre	998
Partido Nacional Republicano	PNR	Right	2
Partido Democrático Republicano	PDR	Centre	998
CHEGA	CH	Right	2
Aliança	A	Right	2
LIVRE	L	Left	1
Partido Popular Monárquico	PPM	Right	2

Table 3 - Operationalization of decision

Another important factor to note is that in the surveys, there were always two answer options that did not reveal any information (“Does not know” and “Not answered”), which will be considered as missing values. Given this fact, within the 9 420 initial participants, we ended up with 7 931 valid observations for Turnout and 5 307 for Decision.

A characterization of the population representing these observations is made using a contingency table for some demographic variables in Table 4. Regarding the Turnout variable, we observed that most of the observations belong to women and that for both sexes, the turnout rate is higher than the no turnout rate. Despite this, males seem to present higher turnout levels. As for the relationship with the year, we concluded that the turnout rate does not have a clear tendency, as it is very inconsistent. Nevertheless, it reached its lowest value in the 2019 elections and its highest in the 2005 elections. Also observing the distribution of the variable year, we see that 2005 was also the year with more people inquired and on the opposite way, 2011 was the year with fewer people inquired.

Summarizing, out of the 9 420 individuals inquired, 7 931 answered the turnout question, where 5 593 responded that they did turnout to vote, corresponding to 59% of the participants.

<i>Variables</i>		<i>Turnout</i>							
		Yes		No		Missing		Total	
		n	%	n	%	n	%	N	%
SEX	Male	2604	63	925	22	636	15	4165	44
	Female	2958	57	1398	27	849	16	5205	55
	Missing	31	62	15	30	4	8	50	1
YEAR	2002	805	62	332	25	166	13	1303	14
	2005	1874	67	572	20	355	13	2801	30
	2009	810	62	312	24	195	15	1317	14
	2011	549	55	268	27	183	18	1000	11
	2015	865	58	357	24	277	18	1499	16
	2019	690	46	497	33	313	21	1500	16
Total		5593	59	2338	25	1489	16	9420	100

Table 4 - Characterization of the sample for the turnout question – original data

In Table 5, a similar analysis is done with the Decision variable. The conclusion drawn for the relationship with the variable sex is that it does not seem to have a predictor effect on the party decision. Male and female voting rates for Left and Right are balanced. As for the distribution of the Left, Right and Other votes over the years, Left parties are typically most voted by Portuguese voters, except in the 2011 elections where the percentage of voters who voted for the right-wing parties is higher. Carrying this fact to the Portuguese political history and context makes sense as the elections of 2011 were made under a crisis context and coincident with Troika's entry into Portugal's economy, being a Left party on the government prior to the elections. Making an overview of the party decision, within the initial 9 420 individuals inquired, 5 307 individuals answered the decision question. The majority (3 266) answered that they voted for Left parties, making 35% of the people inquired. Regarding the party category 'Other', it can be seen that within the individuals who answered this question, only 33 answered that they voted for a Syncretic or Centre party, which represents less than 1% of the total inquired.

Variables		Decision									
		Left		Right		Other		Missing		Total	
		n	%	n	%	n	%	n	%	N	%
SEX	Male	1522	37	938	23	12	0	1693	41	4165	44
	Female	1719	33	1064	20	21	0	2401	46	5205	55
	Missing	25	50	6	12	0	0	19	38	50	1
YEAR	2002	413	32	372	29	0	0	518	40	1303	14
	2005	1190	42	591	21	0	0	1020	36	2801	30
	2009	482	37	272	21	4	0	559	42	1317	14
	2011	206	21	295	30	0	0	499	50	1000	11
	2015	486	32	323	22	9	1	681	45	1499	16
	2019	489	33	155	10	20	1	836	56	1500	16
Total		3266	35	2008	21	33	0	4113	44	9420	100

Table 5 - Characterization of the sample for the decision question – original data

When analyzing the age distribution for each case represented in Table 6 regarding the Turnout, the average age of the group of individuals that answered that they did turn out to vote is higher than those that answered that they did not turn out to vote. As for the Decision, the average age between groups is closer. However, the group with a lower average age is the one that contains individuals that answered that they voted for center or syncretic parties. Between left and right-wing votes, the group that responded that they voted for right-wing parties has a higher average age.

Participation		Yes				No			
Decision	Mean	Std	Min	Max	Mean	Std	Min	Max	
	51	.229	18	99	45	.398	18	99	
	Left				Right				Other
	Mean	Std	Min	Max	Mean	Std	Min	Max	Mean
	51	.294	18	94	52	.391	18	99	48

Table 6 - Age distribution of the sample - original data

3.1.2 Data Treatment

After having all data aggregated, there is the need to treat the data. This consisted of observing the data available and reducing it, only having the information needed. We decided to make this data ‘cleansing’ in three main steps: variable reduction, observation reduction and missing values replacement.

3.1.2.1 Variable Reduction

The variables reduction was decided to do with two criteria: the first one to remove all variables with more than 10% of missing values. In this dataset, missing values are not

‘blank’ answers, but answers like “does not know” or “does not remember. This coding does not represent any valuable data and can interfere with the work made.

For that, it was analyzed the number of missing values per question. There are eight categories with more than 10% of missing values. These should be deleted, except for the variable ‘FREQRELIGIAO’. Although the missing values represent 10% of the observations, the literature review showed that this variable is related to voting behavior and has some explanatory effect. Hence, it is important to keep it. This criterion allowed to remove eight variables, as can be seen, highlighted in bold in Table 7.

Variable	Missing Values		Variable	Missing Values	
	n	%		n	%
SEX	50	1	SIMPATHYBE	826	9
AGE	51	1	SIMPATYCS DPP	760	8
AGEGROUP	51	1	SIMPATHYCDU	762	8
EDUCATION	61	1	SIMPATHYPPD PSD	722	8
CIVILSTATE	45	1	SIMPATHYPS	690	7
WORKINGSITUATION	212	2	POSITIONBE	1432	15
SECTORACTIVITY	2126	23	POSITIONCDSPP	1517	16
INCOME	3747	40	POSITIONCDU	1452	15
RELIGION	219	2	POSITIONPPD PSD	1373	15
NUMBEROFPEOPLE	46	0	POSITIONPS	1350	14
SINDICAT	111	1	DEMOCRACYSATISFACTION	283	3
ECONOMYEVLUTIONPT	302	3	PARTYPROXIMITY	584	6
POLITICALINTEREST	44	0	FREQPAPER	80	1
GOVERNMENTEVALUATION	602	6	FREQRADIO	124	1
WHOISINPOWER	429	5	FREQTV	76	1
WHOPEOPLEVOTE	449	5	FREQNET	4140	44
FREQRELIGIAO	976	10			

Table 7 - Missing values count per variable

The second criterion within the variable reduction was to remove the variables that weren’t proved to influence turnout or decision in literature. Observing the twenty-nine variables remaining in the dataset, it was concluded that the variables NUMBEROFPEOPLE, ECONOMYEVLUTIONPT, GOVERNMENTEVALUATION, WHOISINPOWER, WHOPEOPLEVOTE, SIMPATHYBE, SIMPATYCS DPP, SIMPATHYCDU, SIMPATHYPPD PSD and SIMPATHYPS were not shown in the literature as having any predictive effect on voting behavior. With this criterion, it was possible to remove ten more variables from the dataset.

After having the first step of data ‘cleansing’ done, the dataset remained with nineteen variables (variable description in Appendix 2). To also have a more objective dataset, there is the need to operationalize some variables to reduce the number of answer options. Below in Table 8 are represented the operationalizations done in two variables that remained in the dataset.

In variable WORKINGSITUATION nine different possible answers specified the working situation of the individual. Given that there is no need to go into much detail in this variable, it was decided to transform it into a binary variable with the options ‘Working’ and ‘Not working’.

Also, another variable that was needed to operationalize was the variable PARTYPROXIMITY. For the same reasons as we operationalized the variable Decision, it was decided to transform the several parties into a left- and right-wing scheme. It was also added the option ‘Other’ that refers to center and syncretic parties.

Variable	Original Value	Final Value
WORKINGSITUATION	1- Working full time	1- Working
	2- Working part time	
	3- Working less than part time,	
	4- Family worker unpaid,	
	5- Unemployed	2- Not working
	6- Student	
	7- Retired	
	8- Permanent disability	
	9- domestic	
PARTYPROXIMITY	0- None	0- None
	1- BE	1- Left
	2- PCP	1- Left
	3- PS	1- Left
	4- IL	3- Other
	5- CDS-PP	2- Right
	6- PSD	2- Right
	8- PAN	3- Other

Table 8 - Operationalization of variables

3.1.2.2 Observation Reduction

The next step in data treatment was the observation also consisted of two criteria. The first one was to remove all observations with more than two missing values. As stated previously, in this dataset, what is considered as missing values are the answers “does not know” or “does not remember”.

As resumed below in Table 9, most of the individuals inquired have no missing value, which is very good. Given this information, it was decided to exclude all individuals from the dataset with more than two missing values, making a total of 668 individuals deleted, representing 7% of the total of inquired.

Count of Missing Values	Number of individuals	Count of Missing Values	Number of individuals
0	6 051	7	8
1	1 480	8	1
2	1 221	9	2
3	486	10	2
4	126	11	2
5	32	16	1
6	8		

Table 9 - Count of missing values per individuals

As a second criterion for reducing the observations, it was decided to remove all participants with missing values in demographic variables. As these are sensitive variables representing the population, it does not seem appropriate to replace these values. The variables considered as demographic are SEX, AGE, AGEGROUP, EDUCATION, CIVILSTATE and RELIGION. This criterion allowed to deletion of a total of 247 observations (around 3%).

3.1.2.3 Missing Values Replacement

The last step in the Data ‘cleansing’ is the missing values replacement. There is a total of 1 501 missing values in the dataset that need to be analyzed and replaced. We decided to take different decisions according to the type of variable, as some measures are more or less sensitive to extreme values. Thus, it was decided to replace the missing values in ordinal variables with the median and replace the missing values in nominal variables with the mode.

Table 10 presents the current count of missing values per variable and the decision regarding replacing these values. For example, for the variable ‘WORKINGSITUATION’, it was decided to replace the missing values with the mode, which is ‘Working’ (option 1).

Variable	Count of missing values	Replace by
WORKINGSITUATION	148	Mode
SINDICAT	61	Mode
POLITICALINTEREST	22	Median
DEMOCRACYSATISFACTION	200	Median
PARTYPROXIMITY	229	Mode
FREQPAPER	35	Median
FREQRADIO	58	Median
FREQTV	33	Median
FREQRELGIAO	630	Median

Table 10 – Count of missing values per variable and decision taken to replace them

After having this data treatment done, the dataset remained with nineteen variables and 8 505 observations. A characterization of the population representing these observations using a contingency table for some demographic variables and observing their relation to the objective-variables Turnout and Decision was performed, similar to the previous analysis done.

As per Table 11, out of the 8 505 inquired, 7 600 answered the turnout question. Within these, 5 379 answered that they did turn out to vote, representing 63% of the population. Regarding the variable sex, the conclusions are similar to the previously taken. Most participants are female, and as for the relationship of the variable sex and Turnout, it still seems that the variable sex has some explanatory effect on turnout, as the rate of males answering that they did turn out to vote larger than in the female. The distribution of the variable year shows that 2005 was the year with more people inquired, while on the opposite side, 2011 was the year with fewer people inquired.

In the next steps of this work, the observations with Missing Values in the Turnout question will be ignored for analysis, as they do not give any relevant information and can influence the work done.

Variables		Turnout							
		Yes		No		Missing Values		Total	
		n	%	n	%	n	%	N	%
SEX	Male	2495	67	884	24	359	10	3738	44
	Female	2884	61	1337	28	546	11	4767	56
	Missing	-	-	-	-	-	-	-	-
YEAR	2002	800	66	322	26	95	8	1217	14
	2005	1748	70	525	21	229	9	2500	29
	2009	756	67	275	24	93	8	1124	13
	2011	539	61	259	29	82	9	880	10
	2015	854	61	352	25	201	14	1407	17
	2019	684	50	488	35	205	15	1377	16
Total		5379	63	2221	26	905	11	8505	100

Table 11 - Characterization of the sample for the turnout question – treated data

In Table 12, a similar analysis is done with the Decision variable. As observed, out of the 8 505 individuals inquired, 5 102 gave valid answers to the decision question. Within these individuals, taking only valid answers, the majority answered that they voted for left-wing parties (2 937 individuals, representing 35% of the population). Similar to the previous conclusions drawn, the variable sex does not have a predictor effect on the decision, as the rates are balanced for both sexes.

We also observed that out of the individuals who gave valid answers to the decision question, only 33 answered that they voted for syncretic or center parties, representing less than 1% of the population under study. For that reason, it was decided that for the next steps, not only the missing values will be ignored, as the ‘Other’ will also not be taken into consideration as their rate is very low. Given this fact, we ended up with 5 069 individuals that gave valid answers to the decision question.

Variables		Decision									
		Left		Right		Other		Missing		Total	
		N	%	N	%	N	%	N	%	N	%
SEX	Male	1363	36	991	27	13	0	1371	37	3738	44
	Female	1574	33	1141	24	20	0	2032	43	4767	56
	Missing	-	-	-	-	-	-	-	-	-	-
YEAR	2002	409	34	371	30	0	0	437	36	1217	14
	2005	1100	44	558	22	0	0	842	34	2500	29
	2009	441	39	262	23	4	0	417	37	1124	13
	2011	121	14	370	42	0	0	389	44	880	10
	2015	383	27	416	30	9	1	599	43	1407	17
	2019	483	35	155	11	20	1	719	52	1377	16
Total		2937	35	2132	25	33	0	3403	40	8505	100

Table 12 - Characterization of the inquired for the decision question – treated data

The age distribution for each perspective is represented in Table 13. We observed that regarding the Turnout, the average age is higher within individuals that answered that they did turn out to vote when comparing with individuals that answered that they did not turn out to vote. This fact shows that the data under study follows the premise that younger voters are more likely to be abstentionists. As for the Decision, and considering only left and right-wing votes, we remarked that the average age between individuals who voted for left-wing parties and right-wing parties is close. Yet, it is slightly higher among individuals who answered that they voted for right-wing parties.

	Turnout				Yes				No			
	Mean	Std	Min	Max	Mean	Std	Min	Max	Mean	Std	Min	Max
Decision	51	.233	18	99	45	.413	18	95				
	Left				Right							
	Mean	Std	Min	Max	Mean	Std	Min	Max				
	51	.300	18	94	52	.394	18	99				

Table 13 - Age distribution of the sample - treated data

3.2 Methodology

Having all data treated and a final dataset built, the core of this work is presented: to understand Portuguese voting behavior through machine learning. A prediction model was created that projects the voting behavior in the two perspectives: Turnout and Decision.

Bayesian Networks will be applied to create this predictor model. Several algorithms will need to be used to test them. Afterwards, one will be selected according to its

performance. Many algorithms such as Gaussian Naïve Bayes, Averaged One-Dependence Estimators, or Bayesian Belief Network could have been chosen. Given the software limitations, the selected algorithms to be tested are presented next, based on GeNIe Modeler User Manual from Bayes LLC (2019).

- Bayesian Search (BS) is one of the oldest and most known algorithms used. It was introduced by Cooper & Herkovitz (1992) and refined three years later by Heckerman, Geiger, and Chickering (1995). This algorithm follows a “hill climbing procedure guided by a scoring heuristic with random restarts”.

- PC (prototypical constraint-based) is also one of the oldest and most known learning algorithms. It was introduced by Spirtes et al. in 1993. Using independences observed in data, it infers the structure that has generated them.

- Greedy Thick Thinning (GTT) was described by Cheng et al. (1997) as being based on the Bayesian Search approach. This algorithm builds the model under two phases: the thickening phase and the thinning phase. During the first phase, the algorithm starts with an empty graph and adds connections between variables that maximally increases the marginal probability until no addition results in an increase. In the second phase, the reverse happens: the algorithm removes connections that improve the probability until no removal results in a positive increase.

- Tree Augmented Naïve Bayes (TANB) structure learning algorithm is based on the Bayesian Search method being a semi-naïve structure learning method introduced by Friedman et al. (1997). It starts with a Naïve Bayes structure and adds connections between the feature variables to account for possible dependence between them, imposing the limit of only one additional parent of every feature variable

- Augmented Naïve Bayes (ANB) structure learning algorithm is also a semi-naïve structure learning method based on Bayesian Search approach. Similar to the TANB, it starts with a Naïve Bayes structure and adds connections between the feature variables to account for possible dependence between them. The difference is that there is no limit on the number of additional connections entering each feature variable unless imposed by one of the algorithm's parameters (Friedman, Geiger, and Goldszmidt 1997).

- Naïve Bayes (NB) learning algorithm is a naïve structure learning method. A network structure is not learned but rather fixed by assumption: The class variable is the only parent of all remaining feature variables. There are no other connections between the nodes of the

network. The Naive Bayes structure assumes that the features are independent conditional on the class variable, which leads to inaccuracies when they are not independent.

To validate the models there is the need to select the validation method. There are three in the software: test only, leave-one-out (LOOCV) and K-fold cross-validation. Given that this dissertation uses past information to predict the future, the option chosen will be ‘test only’.

To test these algorithms, the dataset was split into train and test set and given that the aim is to use the information of the past to predict the future it was decided to split according to the years. To also have a higher accuracy of the models it was also decided to train and test the model in two phases. Figure 2 presents the several phases applied to all algorithms. Firstly, the model was trained with the information of four years (2002 to 2011) and tested to predict the voting behavior of 2015; afterwards the model was trained again with the information of five years (2002 to 2015) and tested to predict the results of 2019. Having double training and testing the model will theoretically improve the model's accuracy, ending with better predictions.

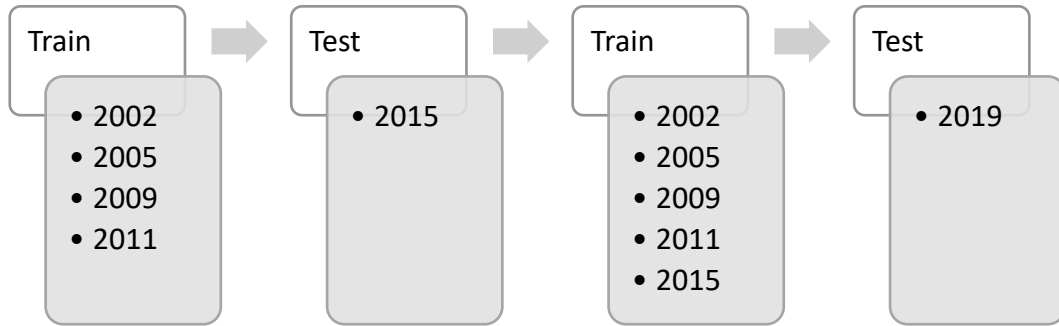


Figure 2 - Prediction model phases

After having all models trained and tested, they were compared through the evaluation of model performance measures. Accuracy, Precision, Recall and Specificity are metrics are calculated based on the confusion matrix.

$$Accuracy = \frac{TP + TN}{P + N} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{P} \quad (3)$$

$$Specificity = \frac{TN}{FP + TN} \quad (4)$$

F-score, is a measure that uses precision and recall to be calculated and results in a value between 0 and 1, where 1 means perfect precision and recall.

$$F - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

The Area under the curve (AUC) derived from the Receiver Operating Characteristic curves (ROC curve) is also used, where the ROC space is drawn, and the several methods are represented by curves and then compared.

Friedman's test is a nonparametric test used to compare models.

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1) \quad (6)$$

Nemenyi's test is a post-hoc test that ranks the models.

$$q = \frac{\bar{R}_{j1} - \bar{R}_{j2}}{\sqrt{\frac{k(k+1)}{6n}}} \quad (7)$$

The Critical Difference (CD) between models was e calculated, and Critical Difference Diagram (CDD) was drawn to compare the differences between models visually.

$$CD = q_\alpha \sqrt{\frac{A(A+1)}{6N}} \quad (8)$$

The models were then compared, and the best one was chosen and explored to analyze the relationships and influences between variables and the target variable. The model chosen was also used to diagnose both states of the target variable to determine the voter profile for each case.

To explore the influence between variables, there was the need to use some tools, namely the *strength of influence*. This instrument shows the influence between variables gives the distance method chosen. There are few options, but the ones described using information from Koiter (2006) are the ones used.

- Euclidean distance calculates the actual spatial distance between two points.
- J-Divergence is the average of the two possible values of the Kullback-Leibler distance, which is the overall difference between two distributions.
- CDF (cumulative distribution functions) is targeted towards ordinal distributions and compares the cumulative distribution functions of the two distributions to be compared.

Another tool used is the *Sensitivity Analysis*, a technique used to investigate small changes in numerical parameters on the output. The outcome is showed in a colors scheme where the variable with the biggest effect is colored in red and the variable with no effect is grey.

3.3 Software

This work used different methods and techniques within Data Analysis and Machine Learning. To perform this dissertation three software were used. Below is the description of where each software was used.

- Exploratory data analysis and missing values imputation was performed using IBM SPSS Statistics version 26 (2020).
- Predictive model was built using Bayesian Networks. Different prediction algorithms were applied, such as Bayesian Search, PC, Greedy Thick Thinning, Tree Augmented Naïve Bayes, Augmented Naïve Bayes and Naïve Bayes. The definition of the voter's profile was also performed using GeNIe version 3.0 (2020).
- RStudio version 1.4.1106 (2020) was used to calculate Friedman's and Nemenyi's statistics, calculate the Critical Difference and draw the CDD using package *scmamp* (Calvo and Santafe 2016).

4 Results and discussion

This section pretends to answer the central question of this dissertation: to understand voting behavior using Data Mining techniques, namely Machine Learning, for both perspectives: Turnout and Decision.

The models presented in this part were built using the algorithms and the methods explained in **Erro! A origem da referência não foi encontrada..** Afterward, they were compared through Evaluation Measures and Statistical Nonparametric tests presented in the same section to define the best predictive model.

Therefore, we explored the model chosen and analyzed it to define the variables that most influence the outcome and draw voters' profiles.

4.1 Turnout

4.1.1 Validation of Turnout models

As mentioned previously, two predictions were performed: data from 2002 to 2011 was used to predict 2015 and data from 2002 to 2015 was used to predict 2019. Therefore, evaluation measures will be taken for both predictions and compared.

Table 14 presents the model's performance metrics for Turnout for both predictions performed.

MODEL	2015 PREDICTION						2019 PREDICTION					
	Accuracy	Precision	Recall (sensitivity)	Specificity	F-score	AUC	Accuracy	Precision	Recall (sensitivity)	Specificity	F-score	AUC
BAYESIAN SEARCH	0.750	0.660	0.293	0.938	0.406	0.724	0.737	0.769	0.527	0.887	0.625	0.783
PC	0.687	0.450	0.321	0.838	0.375	0.642	0.557	0.474	0.582	0.539	0.522	0.617
GREEDY THICK THINNING	0.750	0.660	0.293	0.938	0.406	0.724	0.726	0.787	0.469	0.909	0.588	0.782
TREE AUGMENTED NAÏVE BAYES	0.746	0.604	0.381	0.897	0.467	0.752	0.743	0.749	0.576	0.863	0.651	0.796
AUGMENTED NAÏVE BAYES	0.748	0.613	0.369	0.904	0.461	0.746	0.751	0.761	0.586	0.868	0.662	0.798
NAÏVE BAYES	0.743	0.573	0.466	0.857	0.514	0.769	0.727	0.676	0.660	0.775	0.668	0.796

Table 14 - Evaluation measures from Turnout predictive models

The models seem to have similar evaluation measures, so it is important to perform some nonparametric tests to see statistically significant differences. All the performance measures for each prediction were used excepting F-score since it is calculated based on

precision and recall, making it redundant and repetitive, knowing that these tests consider the rank of the models and not the measures.

It was concluded that considering the performance measures of the six models for the two predictions (2015 and 2019), no significant differences were found for a 5% significance level (2015 prediction: *Friedman's* $\chi^2(5, n= 6) = 8.46$, p-value= 0.13; 2019 prediction: *Friedman's* $X^2(5, 6) = 10.6$, p-value= 0.060). To confirm this conclusion, the *Nemenyi* test was applied, and the critical difference (CD) was also calculated for a significance level of 5%.

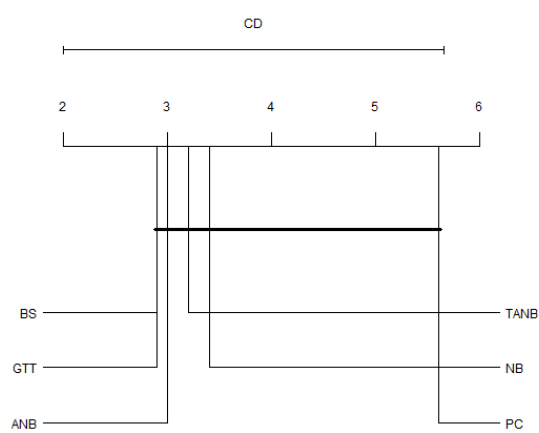


Figure 3 - CDD for 2015 prediction turnout models

In **Erro! A origem da referência não foi encontrada.**, the Critical difference diagram (CDD) is represented for the 2015 predictive models. In **Erro! A origem da referência não foi encontrada.**, the CDD for 2019 predictive models. These diagrams show the proximity between models and compare them with the CD calculated (3.66). It shows that the models are all within the CD range. This fact supports the conclusion taken from

the *Friedman* test.

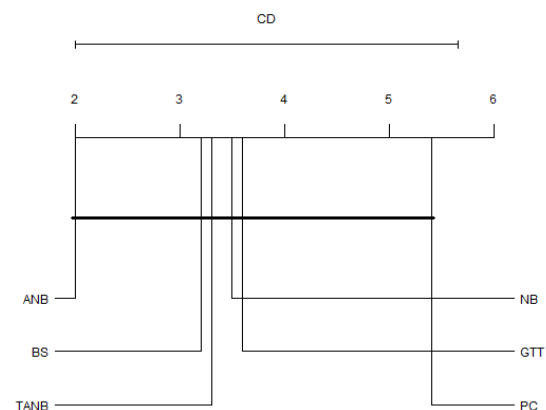


Figure 4 - CDD for 2019 prediction turnout models

Given that the conclusion of the nonparametric tests performed showed no statistically differences between the models, and observing the performance measures of the models represented in both **Erro! A origem da referência não foi encontrada.** and Figure 4. We can conclude that for the

2015 predictions the model with better performance are Bayesian Search and Greedy Thick Thinning, followed by Augmented Naïve Bayes. For 2019 predictions, the best model is the Augmented Naïve Bayes.

For that, Augmented Naïve Bayes seems to be the model with better performance on both predictions; therefore, this was the model chosen.

4.1.2 Turnout model exploration

This model was built with the dataset and the probabilities automatically calculated by the system. There is a 73% probability of individual voting in the initial model against a 27% probability of an individual not voting. We also observed in Figure 5 that in this model all variables influence the target variable.

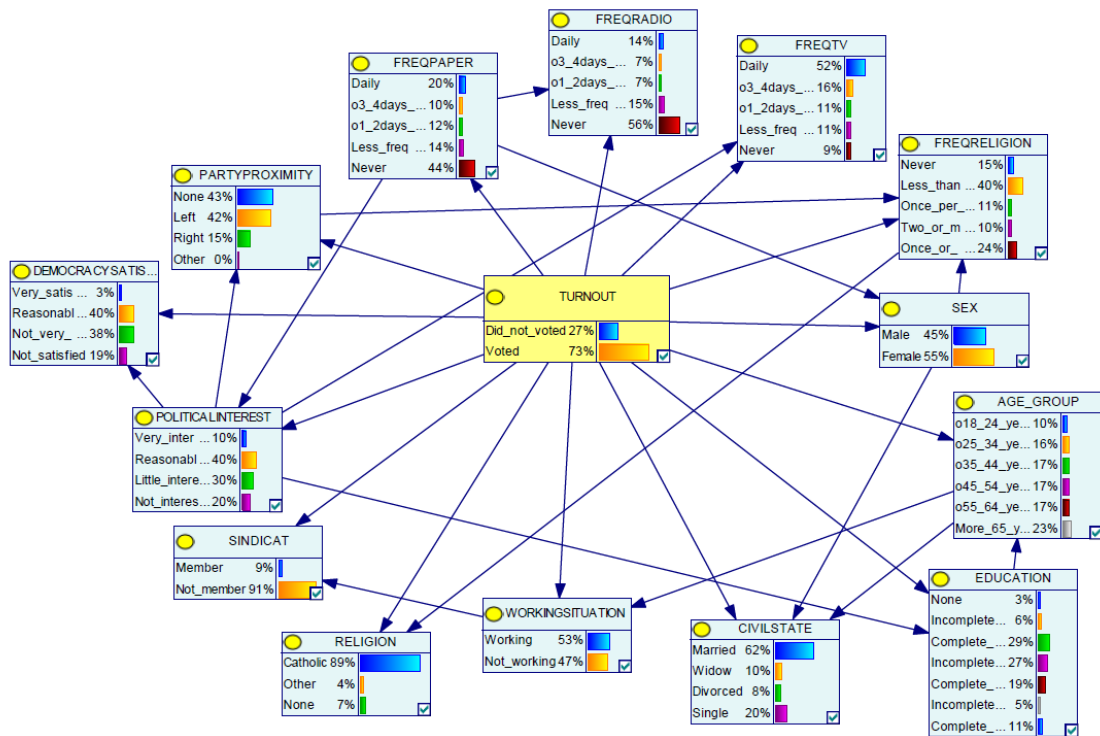


Figure 5 - Turnout predictive model

To find out which variables have higher influence on the target variable Turnout, it is needed to run the *strength of influence*. The weight of the influence of the variables in the target variable is resumed in Table 15.

Using this tool, it was selected as distance measure the Euclidean distance and average. The strongest influence is between variables Education level and Age group, followed by Age group and Civil state and Age group and Working situation. Considering the target variable (Turnout), we stated that this variable has the strongest influence from Party proximity. This is a straightforward relation to understand as individuals close to a

party are more likely to turn out to vote. Other variables that strongly influence the target variable are Political Interest and Frequency of attending religious services.

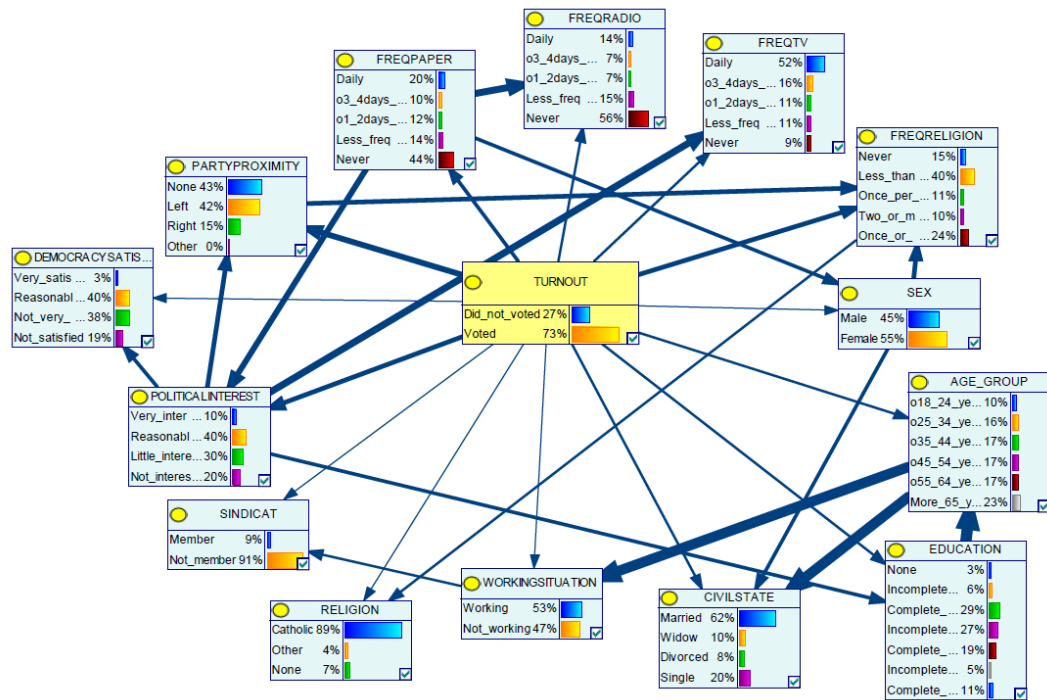


Figure 6 - Turnout predictive model with strength of influence using Euclidean distance and normalized arcs

When selecting as distance measure ‘J-Divergence’ and keeping the normalized arcs, we can observe that the influence of Party proximity in the variable ‘Turnout’ decreases. The variable Political interest appears as the strongest influencer to the target variable, followed by Frequency of attending religious services. This decrease intensifies when selecting as distance measure ‘CDF’. With this distance measure, the frequency of consulting the news through papers emerges as the variable with the most influence on turnout, followed by Political interest.

VARIABLES	EUCLIDEAN DISTANCE	J-DIVERGENCE	CDF
Party proximity	0.210	0.0628	0.107
Political interest	0.173	0.0669	0.139
Frequency of attending to religious services	0.149	0.0635	0.127
Frequency of consulting news through papers	0.131	0.0328	0.141

Table 15 - Weight of influence of variables in Turnout

Using *Sensitivity analysis*, represented in Figure 7, we can see that all variables are colored in grey. This means that these are not used in calculating the posterior probability distributions over the variable Turnout, which is the target variable.

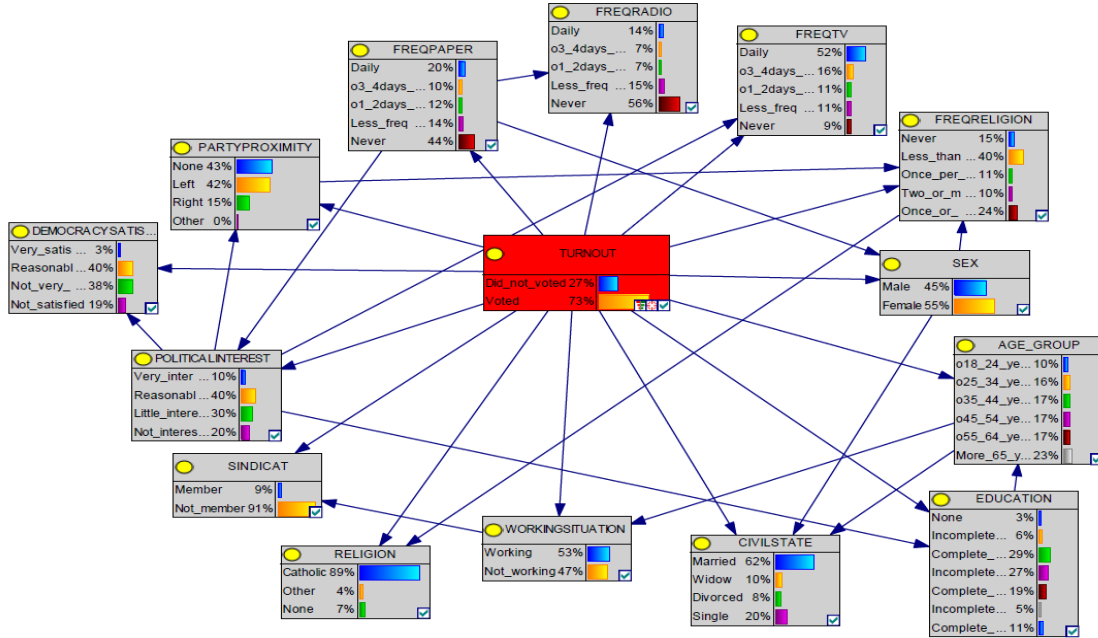


Figure 7 - Sensitivity analysis for turnout model chosen

4.1.3 Diagnosis for Turnout states

4.1.3.1 “Did not voted” State

A diagnosis can be made to the model, for example, to see the profile of an individual that chooses not to vote in the Portuguese general elections.

Selecting the state “Did_not_voted” in the Turnout node and observing only the previously shown variables to have the strongest influence in the target variable. We can observe in Figure 8 the voter’s profile. An individual who chooses to abstain in the Portuguese general elections has more probably not proximity to any party, is not politically interested, attends to religious services less than once per month and never consults the political news through papers.

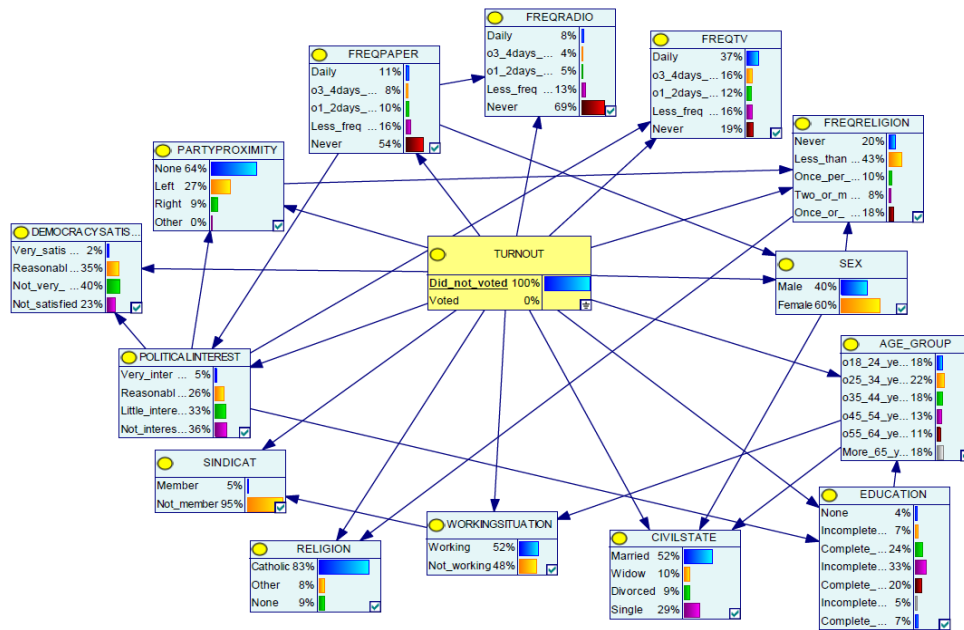


Figure 8 - Turnout diagnosis for 'Did_not_voted' state

4.1.3.2 “Voted” State

Similar diagnosis was made for the “Voted” state. When selecting that state in the Turnout node and updating, the below model was obtained. In Figure 9, we stated that an individual who chooses to vote in the Portuguese general elections is more probably proximate to a left-wing party, is reasonably politically interested, attends to religious services less than once per month and never consults the political news through papers.

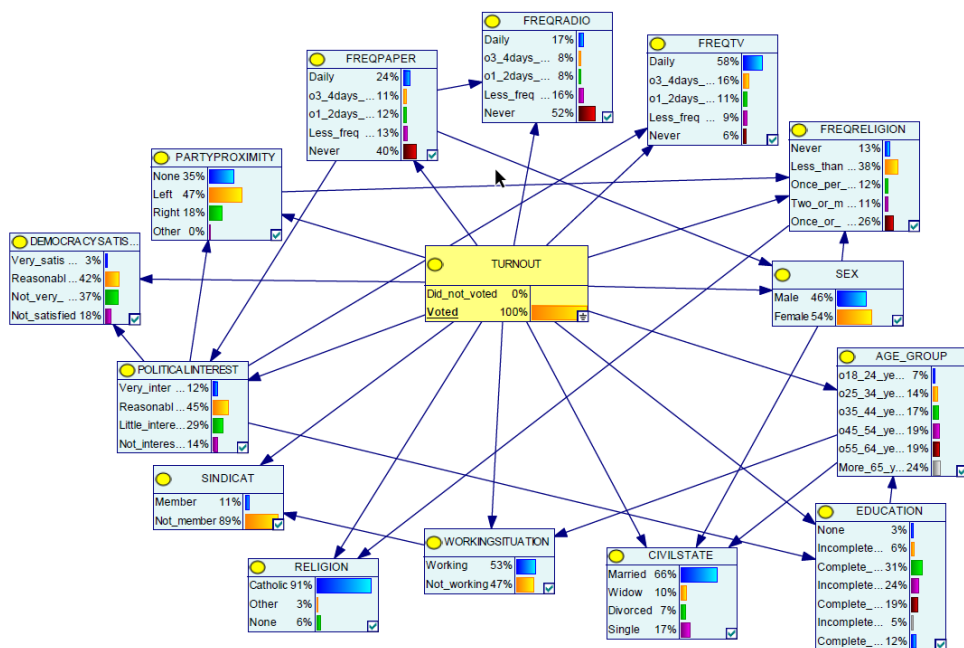


Figure 9 - Turnout diagnosis for 'Voted' state

4.1.4 Turnout models conclusions

Some conclusions can be drawn regarding Turnout and the variables that most influence the decision of voting or not voting.

The models were built and compared, and Augmented Naïve Bayes was the best predictor for the data. Using some tools and techniques, it was noticed that four variables appeared as the biggest influencers for Turnout: Party proximity, Political interest, Frequency of attending to religious services and Frequency of consulting news through papers. With this information, the voter's profile for both outcomes was drawn, ending with the characteristics shown in Table 16.

VARIABLE	State 'Did not voted'	State 'Voted'
Party proximity	No proximity to any party	Proximate to a left-wing party
Political interest	Not politically interested	Reasonably politically interested
Frequency of attending to religious services	Attends to religious services less than once per month	Attends to religious services less than once per month
Frequency of consulting news through papers	Never consults the political news through papers.	Never consults the political news through papers

Table 16 - Voter's profiles regarding Turnout

4.2 Decision

4.2.1 Validation of Decision models

As mentioned previously, two predictions will be performed: data from 2002 to 2011 will be used to predict 2015 and data from 2002 to 2015 will be used to predict 2019. Therefore, evaluation measures will be taken for both predictions and compared.

After training and testing the several models for both predictions, the below table with evaluation measures was obtained.

MODEL	2015 PREDICTION						2019 PREDICTION					
	Accuracy	Precision	Recall (sensitivity)	Specificity	F-score	AUC	Accuracy	Precision	Recall (sensitivity)	Specificity	F-score	AUC
BAYESIAN SEARCH	0.486	0.477	0.775	0.219	0.591	0.511	0.926	0.950	0.952	0.845	0.951	0.957
PC	0.496	0.482	0.708	0.300	0.573	0.509	0.850	0.923	0.874	0.774	0.898	0.866
GREEDY THICK THINNING	0.522	0.501	0.791	0.274	0.613	0.535	0.934	0.930	0.988	0.768	0.958	0.952
TREE AUGMENTED NAÏVE BAYES	0.508	0.407	0.687	0.344	0.511	0.531	0.843	0.942	0.845	0.839	0.891	0.917
AUGMENTED NAÏVE BAYES	0.508	0.491	0.684	0.346	0.572	0.537	0.865	0.950	0.867	0.858	0.907	0.924
NAÏVE BAYES	0.528	0.506	0.642	0.423	0.566	0.520	0.817	0.944	0.805	0.852	0.869	0.909

Table 17 - Evaluation measures from Decision predictive models

There is a discrepancy between the evaluation measures of the prediction of 2015 and the evaluation measures of prediction of 2019. The first predictions are very poor, contrasting with the last predictions that are very accurate. It is important to remember that 2015 was predicted with data from 2002 to 2011, while 2019 predictions were done with data from 2002 to 2015. To understand what may be causing this difference, 2019 was predicted once more using the same data used to predict 2015 (2002 to 2011), being the goal is to study the impact of 2015 data in the 2019 prediction.

Looking at Table 18 and comparing the measures with the previous evaluation measures for the 2019 prediction, we can observe that both are very similar. Where in some cases the prediction of 2019 has better evaluation measures without using 2015 data. In other, the prediction is worse. It can be concluded that 2015 is an atypical year and that the information retrieved in this year, does not contribute much to the prediction of 2019. However, it does not seem to have significant a negative impact on the predictions. For that, the data of 2015 will still be included.

MODEL	2019 PREDICTION (WITHOUT 2015 DATA)					
	Accuracy	Precision	Recall (sensitivity)	Specificity	F-score	AUC
BAYESIAN SEARCH	0.904	0.947	0.925	0.839	0.936	0.949
PC	0.887	0.944	0.905	0.832	0.924	0.902
GREEDY THICK THINNING	0.929	0.945	0.963	0.826	0.954	0.956
TREE AUGMENTED NAÏVE BAYES	0.851	0.947	0.851	0.852	0.896	0.923
AUGMENTED NAÏVE BAYES	0.851	0.949	0.849	0.858	0.896	0.924
NAÏVE BAYES	0.813	0.940	0.805	0.839	0.867	0.913

Table 18 - Evaluation measures from Decision predictive models for the year of 2019 using data from 2002 to 2011

Studying Portuguese political history and, more specifically, the 2015 elections, it was indeed an atypical year. The voting decision was balanced between right-wing and left-wing votes. In these elections, a right-wing coalition composed of two parties (PSD and CSP-PP) won without majority votes. That caused a left-wing coalition to be created composed of four left-wing parties (PS, PCP, BE and *Os Verdes*) and later known as *A Geringonça*. This was a unique event that supports the abnormality of the 2015 elections.

When coming back to Table 17, we can see that the models for both predictions (2015 and 2019) have similar evaluation measures. It is then important to perform nonparametric tests to confirm this suspicion. All the evaluation measures for each prediction were used excepting F-score, as it is calculated based on precision and recall that makes it redundant and repetitive. Knowing also that these tests consider the rank of the models and not the measures themselves.

Friedman's test was performed to see if there are statistically significant differences

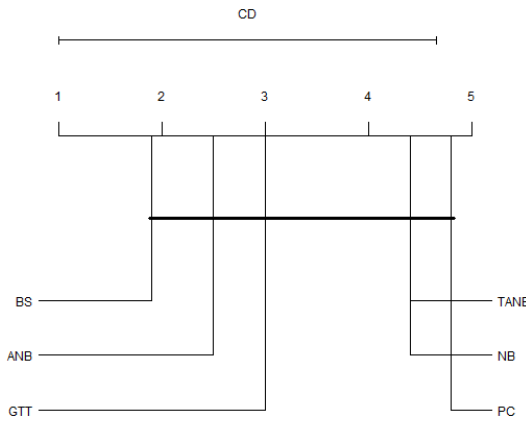


Figure 10 - CDD for 2015 prediction decision models

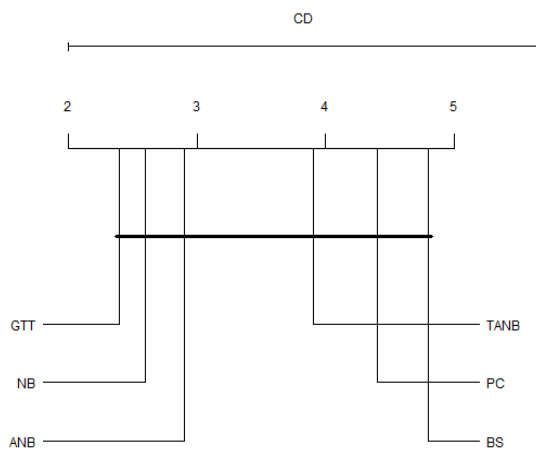


Figure 11 - CDD for 2019 prediction decision models

between the models. It was concluded that considering a significance level of 5% and the accuracy of the six models for the two predictions (2015 and 2019), no significant differences were found (2015 prediction: *Friedman's* $\chi^2(5, 6) = 3.76$, p-value= 0.58; 2019 prediction: *Friedman's* $\chi^2(5, 6) = 11.0$, p-value= 0.051). To confirm this conclusion, the *Nemenyi* test was applied, and the critical difference (CD) was calculated for both predictions, considering again a significance level of 0.05.

Figure 11 and Figure 12 show the Critical difference diagram (CDD) for the 2015 and 2019 predictive models, respectively. It is shown the proximity between models, and as it can be observed in both cases. They are all within the CD range, which supports the conclusion taken from the *Friedman* test.

Since the conclusion of the nonparametric tests performed showed no statistical differences between the models. And observing the performance measures of the models represented in both figures. We can conclude that for the 2015 predictions the model with better performance is Greedy Thick thinning, followed by Naïve Bayes and Augmented Naïve Bayes. While for 2019 predictions the best model is Bayesian Search, followed by Augmented Naïve Bayes and Greedy Thick Thinning

Given this, Greedy Thick Thinning will be the model chosen as it seems to be the model with the best performance on both predictions.

4.2.2 Decision model exploration

This model was built with the dataset and the probabilities automatically calculated by the system. The initial model shows a 55% probability of an individual voting for a left-wing party against a 45% probability of voting for a right-wing party. From Figure 12 it can also be stated that there are only two variables with an effect on Decision: Party proximity and Syndicate.

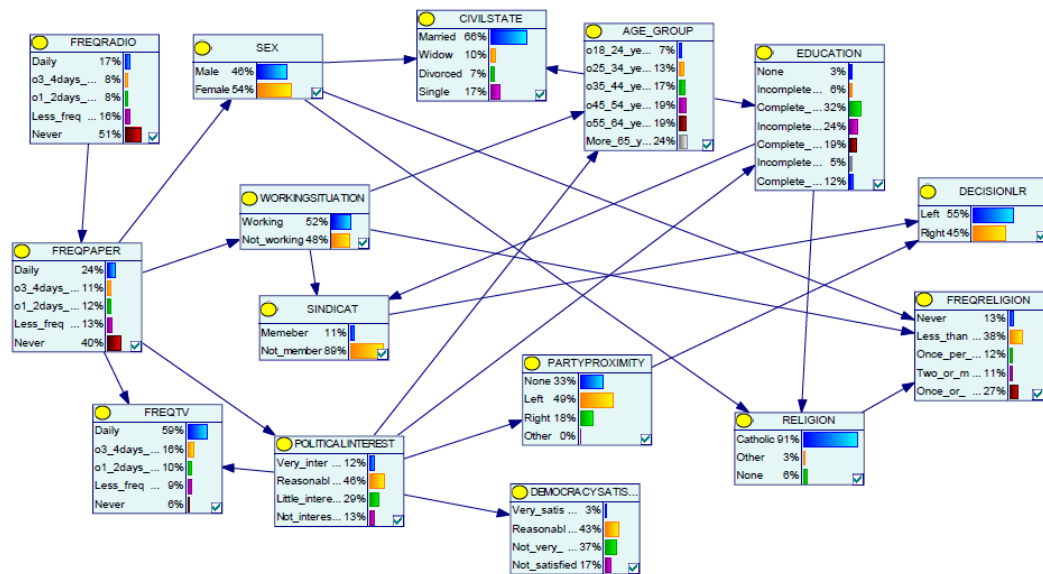


Figure 12 - Decision predictive model

To determine which variables have a bigger influence on the target variable Decision, it is needed to run the *strength of influence*. The weight of the influences of the variables in the target variable is resumed in Table 19.

Using the *strength of influence* tool and selecting as distance measure the Euclidean distance and average, we also selected the normalized widths to see the highest influence between variables. The strongest influence exists between the variables Working situation and Age group, followed by Age group and Civil State and Religion and Frequency of attending to religious services. As for the target variable (Decision), we observed that this variable has the strongest influence from Party Proximity and Syndicate membership. This is an easy relation to understand as individuals close to a party are more likely to vote for that party.

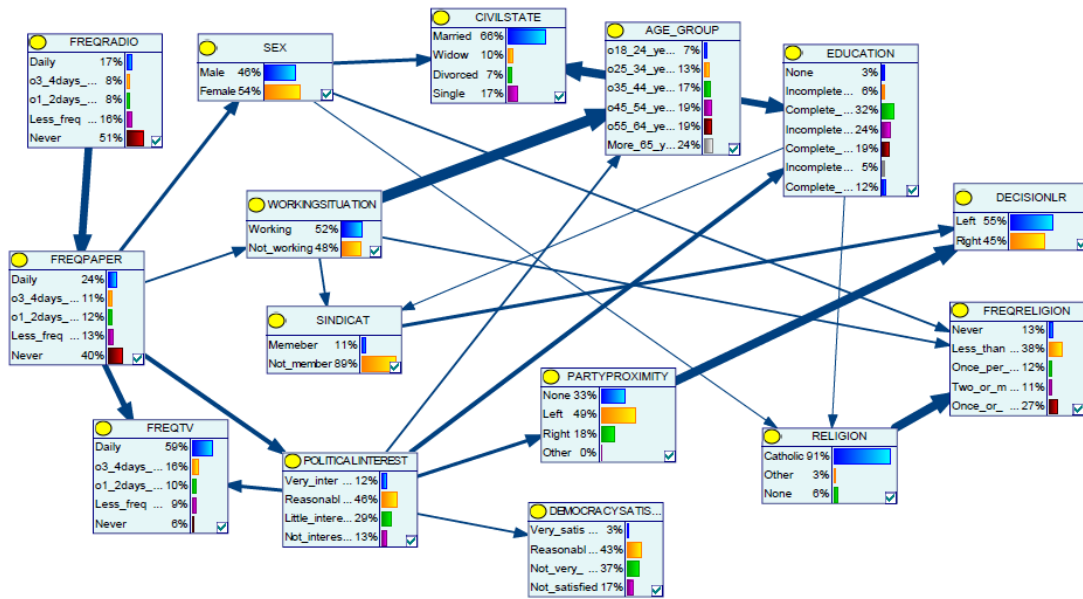


Figure 13 - Decision predictive model with strength of influence using Euclidean distance and normalized arcs

When selecting as distance measure 'J-Divergence' and keeping the normalized arcs, we see that the influence of Party Proximity in the variable Decision maintains, although it decreases its weight. This fact changes when selecting as distance measure 'CDF', where with this distance measure, the influence between the variables Party proximity and Decision emerge as the biggest influence.

VARIABLES	EUCLIDEAN DISTANCE	J-DIVERGENCE	CDF
Party proximity	0.332	0.155	0.332
Syndicate Member	0.151	0.0324	0.151

Table 19 - Weight of influence of variables in Decision

Sensitivity analysis was also done in this model, which shows us the effect of small changes in numerical parameters on the output. It gives the information on which variables

are used to calculate the posterior probability distributions over the target variable. As per Figure 14, we notice that the variable that most contributes to the probability calculation of Decision is Party proximity, followed by Syndicate. This is a similar conclusion as the previous one taken from the *strength of influence*.

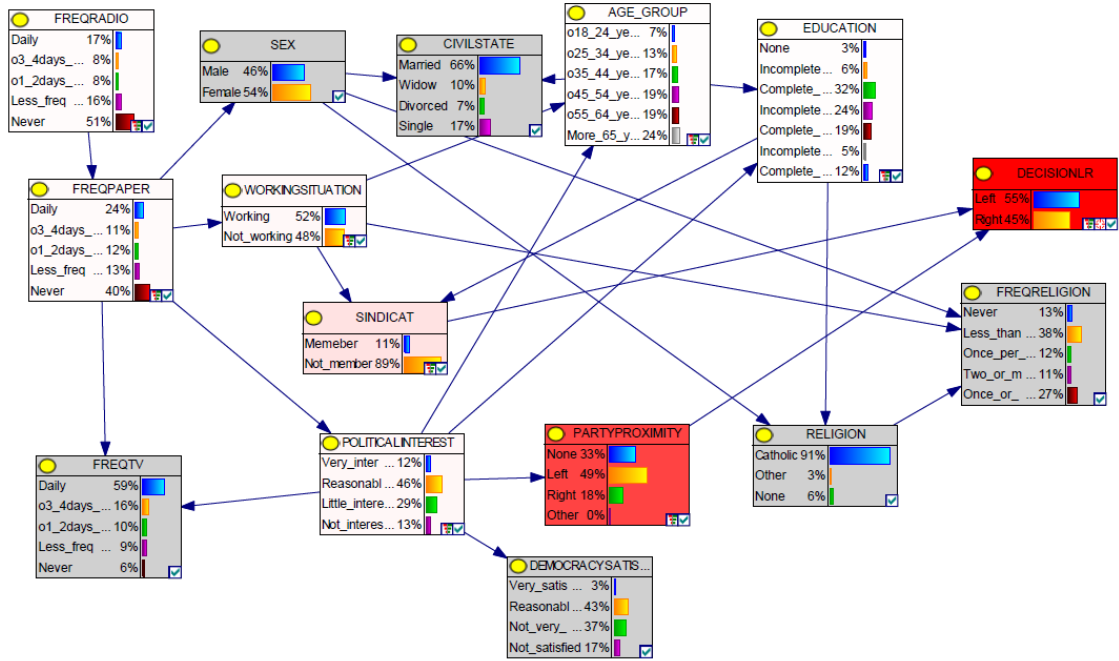


Figure 14 - Sensitivity analysis for decision model chosen

As shown previously, Party Proximity strongly influences voting decision, as mentioned on several research analyzed in the Literature Review. Given this, it seemed relevant to remove the variable from the dataset and create a new model to study its impact and see if new relationships with the target variable would emerge.

Figure 15 represents the model without the Party Proximity variable, and for simplification purposes, it was created using the same algorithm as before (Greedy Thick Thinning). We noticed that by removing a variable that strongly influenced the target variable (Decision), a new relationship has emerged that is composed by Frequency of attending to religious services and Decision.

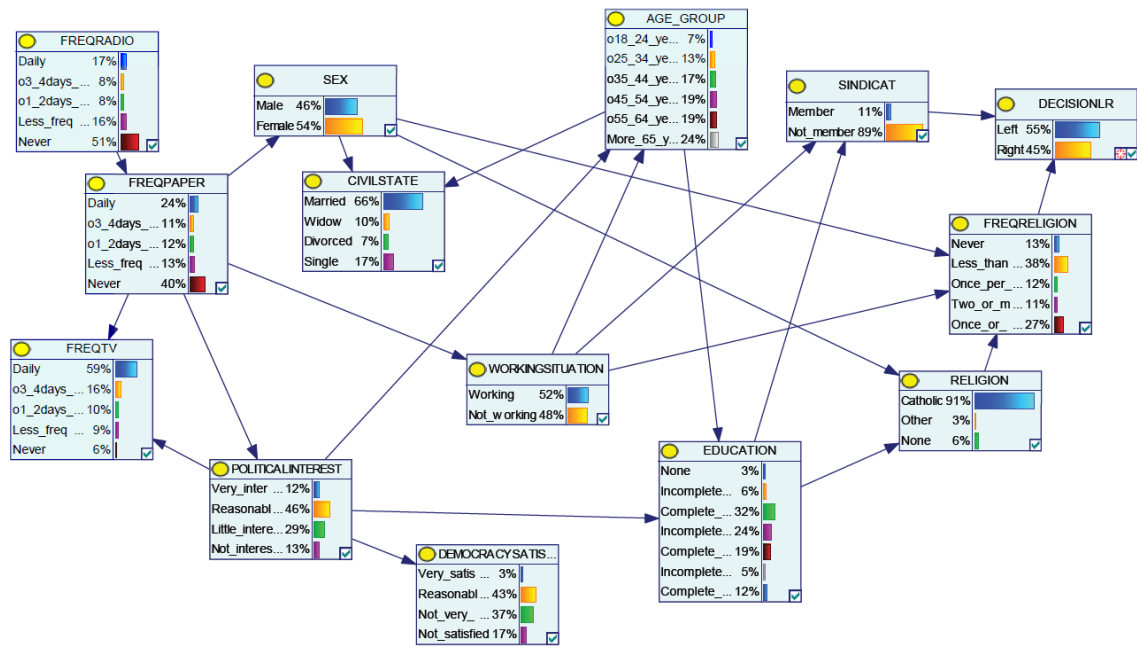


Figure 15 - Decision predictive model without Party Proximity variable

To study the strength of the relationship between these two, *Sensitivity Analysis* was applied. As per Figure 16 below, we can see that the variables Syndicate and Frequency of attending to religious services have similar weight on the probability calculation of the target variable.

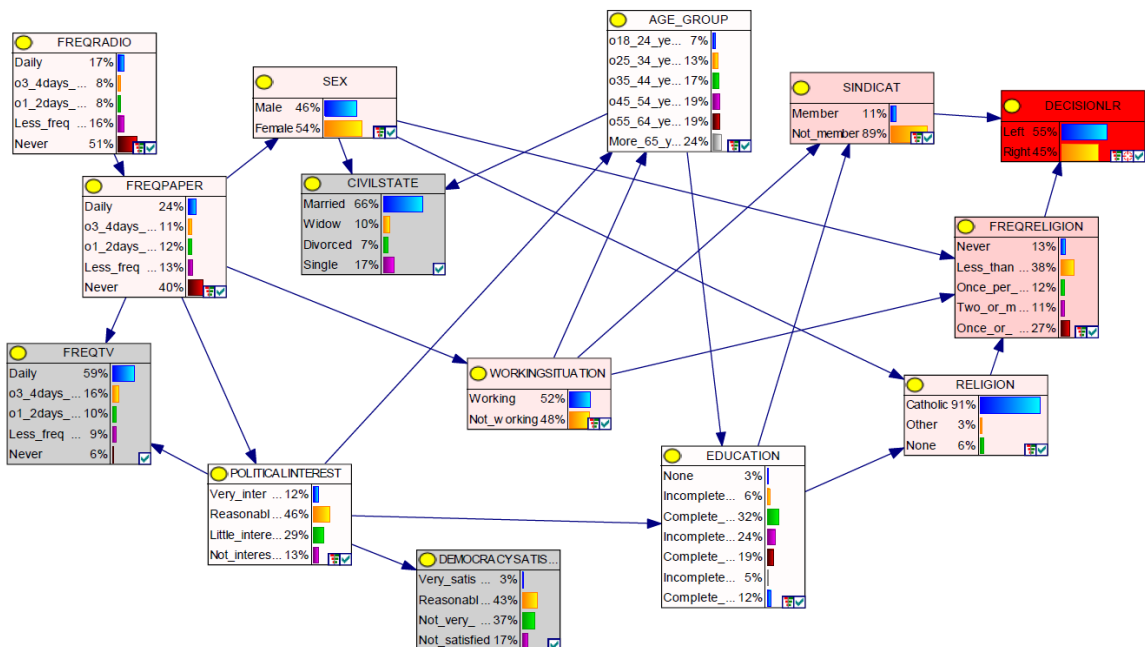


Figure 16 - Sensitivity analysis for decision model without Party Proximity

Nevertheless, by removing the variable that most influenced the variable under study, it is also important to note that the model's accuracy has dropped significantly from 0.934 to 0.726.

4.2.3 Diagnosis for Decision states

4.2.3.1 'Right' State

A diagnosis can be made to the model to see the profile of an individual that chooses to vote for a right-wing party in the Portuguese general elections. Selecting the state 'Right' stands for voting for a right-wing party in the Decision node and taking only into consideration the two variables previously shown to influence the target variable. We can observe that an individual who chooses to vote for a right-wing party is more probably not a member of a syndicate. Although Party proximity was previously shown as the variable with the strongest influence, it seems that this individual is not close to any particular party. In the below model, we also noticed that there is no obvious association between Decision and Working Situation, for example, as the probability in this last variable has a similar distribution in all states.

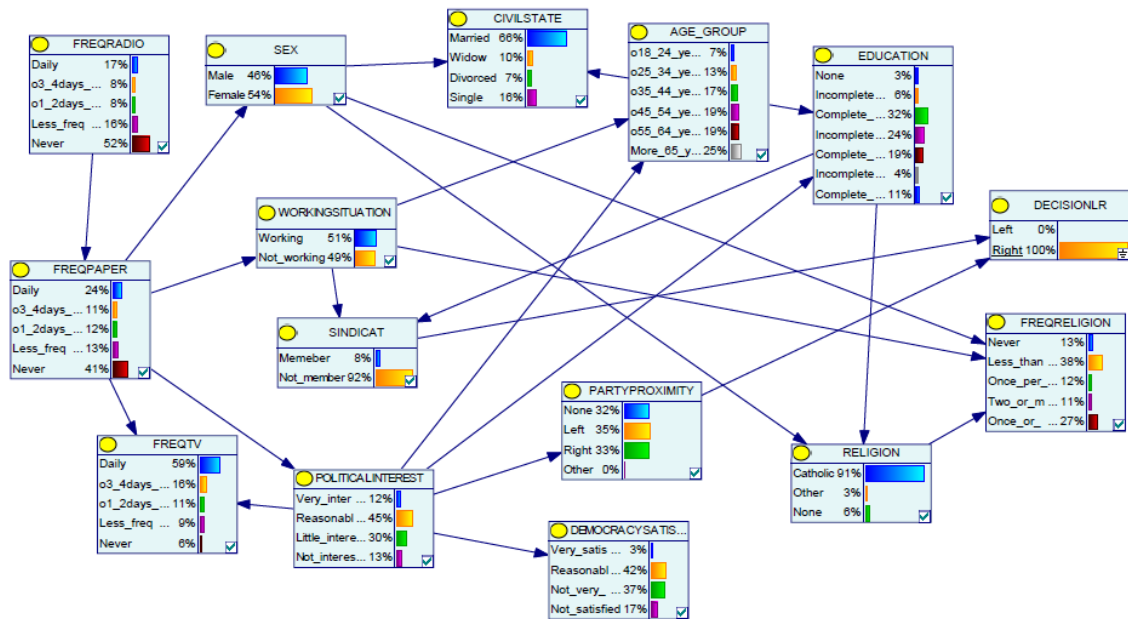


Figure 17 - Decision diagnosis for 'Right' state

4.2.3.2 'Left' state

We performed a similar diagnosis to the one shown previously, but in this case to see the typical profile of an individual that chooses to vote for a Left-wing party. Observing

Figure 18, we stated that an individual who decides to vote for a left-wing party is more probably proximate to a left-wing party and not a member of a syndicate.

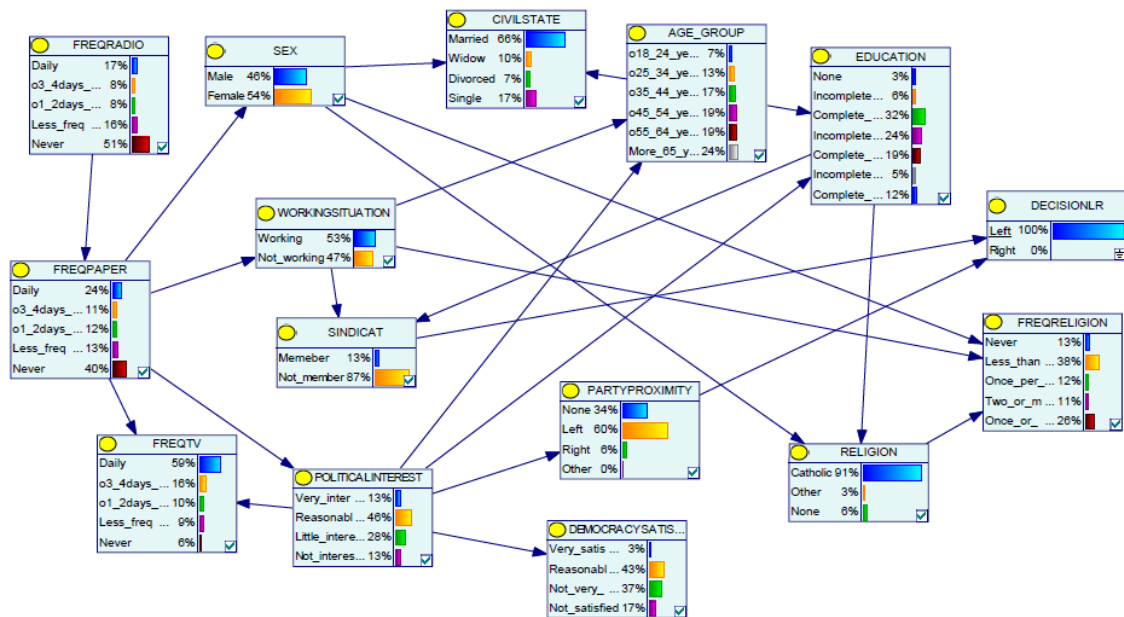


Figure 18 - Decision diagnosis for 'Right' state

4.2.4 Decision models conclusions

Some conclusions can be drawn regarding the variable Decision and the variables that most influence voting for a right or a left-wing party.

The models were built and compared, and Greedy Thick Thinning was proven to be the best predictor for the data. Two variables were the most relevant influencers for Decision: Party proximity and Syndicate member using tools and techniques. With this information, the voter's profile for both outcomes was drawn, ending with the characteristics shown in Table 20.

VARIABLE	State 'Right'	State 'Left'
Party proximity	No proximity to any party	Proximate to a left-wing party
Syndicate member	Not syndicate member	Not syndicate member

Table 20 - Voter's profiles regarding Decision

5 Conclusions

This dissertation aimed to understand Portuguese voting behavior in the general elections. To uncover the main drivers for both choices taken on elections day: Turnout (to vote) and Decision (in whom).

To study this question, data from six independent post-electoral surveys between 2002 and 2019 were aggregated and transformed to build one coherent dataset. There was the need to analyze each question of all surveys to keep only the ones common to all, recode the answers and treat the data. From the initial 9 420 observations, the dataset ended up with 8 505 and nineteen variables. Out of the individuals inquired, 7 600 answered the Turnout question, and within these, 71% (5 379) answered that they did turnout to vote. As for the Decision question, 5 102 answered it, where 58% answered that they voted for a left-wing party. This shows that there is a tendency for Portuguese voters to vote for left parties.

Six algorithms were applied to build the models (BS, PC, GTT, NB, ANB, TANB), with all probabilities calculated by the system. The models were built to make two predictions: 2015 using data from 2002 to 2011 and 2019 using data from 2002 to 2015. Always using data from the past to predict the future. They were then compared through performance measures and nonparametric tests to define the best predictor for each perspective.

We concluded that in the Turnout case, the best predictor was the Augmented Naïve Bayes. When analyzing the influences between the several variables and the target variables, we noted that four emerged as the biggest influencers: Party proximity, Political interest, Frequency of attending to religious services and Frequency of consulting news through papers. When drawing the voter's profile for both outcomes of the Turnout, the main differences between individuals who voted and individuals who abstained are that while the first one is close to a left-wing party and reasonably politically interested, the second one is not proximate to any party and is not politically interested.

As for the Decision, the model chosen was Greedy Thick Thinning. It was also rebuilt for the 2019 prediction using data from 2002 to 2011, as the performance measures between both predictions had a discrepancy. It became important to study the impact of the data of 2015 in the prediction. It was concluded that the data collected this year is redundant, as it does not improve the prediction of 2019. Studying the model chosen and the variable

influences, two variables appeared as influencers of the target variable: Party proximity and Syndicate membership. When drawing the voter's profile for both outcomes the difference between both is that while a right-wing voter is not proximate to any party, a left-wing voter is closer to a left-wing party.

Given that Party proximity was shown as a heavy influencer of the Decision, it was decided to rebuild the model excluding this variable as it may overcome and control the model. When performing this Union membership remained a variable that influences the target, and a new variable merged as an influencer: Frequency of attending to religious services. Despite this, the model's accuracy dropped, proving that Party proximity is an important variable to predict the Decision.

This study has some limitations related to the data. Although it used a considerable volume of data (9 420 observation), it ended up with a lower number of valid observations in each perspective (7 600 for Turnout and 5 102 for Decision) due to the amount of missing values (questions not answered). Furthermore, although it had an interesting range of data in some variables, for example, Age and Education, other variables had most of their observations in one option, for example Syndicate member and Frequency of attending to religious services. This fact difficult the analysis as there is not enough data to see the differences between profiles.

Despite these limitations, to the best of our knowledge, this is the first practical study related to voting behavior in Portugal using Machine Learning, which is an important fact and enriches the study in this area. It also opens the possibility to more similar studies in the future that can use this dissertation as a base and update it with future data.

References

- Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31(2): 211–36.
- Barreto, António, André Freire, Marina Costa Lobo, and Pedro C. Magalhães. 2002. *As Eleições Legislativas de 2002 - Inquérito Pós-Eleitoral, 2002*. <https://dados.rcaap.pt/handle/10400.20/1/simple-search?filterquery=Magalhães%2C+Pedro&filtername=author&filtertype>equals>.
- Bayes Fusion LLC. 2019. "GeNIe Modeler." *GeNIe Modeler*: 1–588.
- BayesFusion, LLC. 2020. "GeNIe Modeler." <http://www.bayesfusion.com/>.
- Birch, Sarah. 2010. "Perceptions of Electoral Fairness and Voter Turnout." *Comparative Political Studies* 43(12): 1601–22.
- Biswas, Aindrila, Nikhil Ingle, and Mousumi Roy. 2014. "Influence of Social Media on Voting Behavior." *Journal of Power* 2(2): 127–55.
- Calvo, Borja, and Guzman Santafe. 2016. "Statistical Comparison of Multiple Algorithms in Multiple Problems."
- Cancela, João, and Benny Geys. 2016. "Explaining Voter Turnout: A Meta-Analysis of National and Subnational Elections." *Electoral Studies* 42(2006): 264–75.
- Cancela, João, and Pedro Magalhães. 2020. "As Bases Sociais Dos Partidos Portugueses." : 1–34.
- Cantarella, Michele, Nicolò Fraccaroli, and Roberto Geno Volpe. 2020. "Does Fake News Affect Voting Behaviour?" *SSRN Electronic Journal* 18(6).
- Castro, Vítor, and R Martins. 2016. "Political Cycles and Government Expenditures: Evidence from Portugal." *Applied Economics Letter* 23(1): 34–37. http://www.nipe.eeg.uminho.pt/Uploads/Seminários 2013/2014-01-22_Castro.pdf.
- Chen, Bin, Jiarong Hong, and Yadong Wang. 1997. "The Minimum Feature Subset Selection Problem." *J. of Comput. Sci. & Technol.*: 283.
- Cooper, Gregory F., and Edward Herskovits. 1992. "A Bayesian Method for the Induction of Probabilistic Networks from Data." *Machine Learning* 9(4): 309–47.

- Costa, Patrício. 2020. "Learning the Dynamics of Voting Behaviour during a General Election Campaign Using a Dynamic Bayesian Network."
- Costa, Patrício, and Frederico Ferreira da Silva. 2015. "The Impact of Voter Evaluations of Leaders' Traits on Voting Behaviour: Evidence from Seven European Countries." *West European Politics* 38(6): 1226–50.
- Falck, Oliver, Robert Gold, and Stephan Heblich. 2014. "E-Lectons: Voting Behavior and the Internet." *American Economic Review* 104(7): 2238–65.
- Ferreira Da Silva, Frederico, and Patrício Costa. 2019. "Do We Need Warm Leaders? Exploratory Study of the Role of Voter Evaluations of Leaders' Traits on Turnout in Seven European Countries." *European Journal of Political Research* 58(1): 117–40.
- Foladare, Irving S. 2019. "The Effect of Neighborhood on Voting Behavior." *Exposure to Polls, Cognitive Mobilization, and Voting Behavior: the 2002 General Elections in Portugal* 83(4): 516–29.
- Freire, André. 2013. "Cleavages, Values and the Vote in Portugal, 2005-09." *Portuguese Journal of Social Science* 12(3): 317–40.
- Friedman, Nir, Dan Geiger, and Moises Goldszmidt. 1997. "Bayesian Network Classifier." *Machine Learning* 97(1–4): 131–41.
- Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102(1): 33–48.
- Geys, Benny. 2006. "Explaining Voter Turnout: A Review of Aggregate-Level Research." *Electoral Studies* 25(4): 637–63.
- Grönlund, Kimmo, and Maija Setälä. 2007. "Political Trust, Satisfaction and Voter Turnout." *Comparative European Politics* 5(4): 400–422.
- Hadjar, Andreas, Michael Beck, Andreas Hadjar, and Michael Beck. 2010. "WHO DOES NOT PARTICIPATE IN ELECTIONS IN EUROPE AND WHY IS THIS? A Multilevel Analysis of Social Mechanisms behind Non-Voting ELECTIONS IN EUROPE AND WHY IS A Multilevel Analysis of Social Mechanisms Behind." 6696.

- Heckerman, David, Dan Geiger, and David M. Chickering. 1995. "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data." *Machine Learning* 20(3): 197–243.
- Jalali, Carlos. 2003. "A Investigação Do Comportamento Eleitoral Em Portugal: História e Perspectivas Futuras." *Análise Social* 38(167): 545–72.
- Javaid, Umbreen, and Urwa Elahi. 2014. "Patterns of Political Perceptions, Attitudes and Voting Behaviour: Influence of Media." *South Asian Studies A Research Journal of South Asian Studies* 29(2): 363–78.
- Jones, Jason J. et al. 2017. "Social Influence and Political Mobilization: Further Evidence from a Randomized Experiment in the 2012 U.S. Presidential Election." *PLoS ONE* 12(4): 1–9.
- Koiter, J R. 2006. "Visualizing Inference in Bayesian Networks." *Man-machine interaction group* Master of. <http://www.kbs.twi.tudelft.nl/Publications/MSc/2006-JRKoiter-Msc.html>.
- Leigh, Andrew. 2011. "Economic Voting and Electoral Behaviour: How Do Individual, Local and National Factors Affect the Partisan Choice?" *SSRN Electronic Journal*.
- Lewis-Beck, Michael S., and Marina Costa Lobo. 2011. "Anchoring the Portuguese Voter: Panel Dynamics in a Newer Electorate." *Political Research Quarterly* 64(2): 293–308.
- Lobo, Marina Costa et al. 2019. *Estudo Eleitoral Português, 2019*. <https://dados.rcaap.pt/handle/10400.20/1/simple-search?filterquery=Magalhães%2C+Pedro&filtername=author&filtertype>equals>.
- Lobo, Marina Costa, and Pedro C. Magalhães. 2011. *Estudo Eleitoral Português, 2011*. <https://dados.rcaap.pt/handle/10400.20/1/simple-search?filterquery=Magalhães%2C+Pedro&filtername=author&filtertype>equals>.
- Lobo, Marina Costa, Pedro C. Magalhães, and António Barreto. 2009. *Estudo Eleitoral Português, 2009*. <https://dados.rcaap.pt/handle/10400.20/1/simple-search?filterquery=Magalhães%2C+Pedro&filtername=author&filtertype>equals>.
- Lobo, Marina Costa, Pedro C. Magalhães, António Barreto, and André Freire. 2005. *Estudo Eleitoral Português, 2005*. <https://dados.rcaap.pt/handle/10400.20/1/simple-search?filterquery=Magalhães%2C+Pedro&filtername=author&filtertype>equals>.

- Lobo, Marina Costa, Pedro C. Magalhães, and João Tiago Gaspar. 2015. *Estudo Eleitoral Português, 2015*. <https://dados.rcaap.pt/handle/10400.20/1/simple-search?filterquery=Magalhães%2C+Pedro&filtername=author&filtertype>equals>.
- Magalhães, Pedro. 2008. "Redes Sociais e Participação Eleitoral." XLIII: 473–504.
- Magalhães, Pedro C. 2013. "Exposure to Polls, Cognitive Mobilization, and Voting Behavior: The 2002 General Elections in Portugal." 53(9): 1689–99.
- Martins, Rodrigo, and Francisco José Veiga. 2013. "Turnout and the Modeling of Economic Conditions: Evidence from Portuguese Elections." *SSRN Electronic Journal*: 1–33.
- NY: IBM Corp. 2020. "IBM SPSS Statistics for Windows." <https://www.ibm.com/analytics/spss-statistics-software>.
- Pita Barros, Henrique. 2017. "Warm Glow Voting? An Analysis of Turnout in Portugal." *SSRN Electronic Journal* (November): 1–20.
- Smets, Kaat, and Carolien van Ham. 2013. "The Embarrassment of Riches? A Meta-Analysis of Individual-Level Research on Voter Turnout." *Electoral Studies* 32(2): 344–59. <http://dx.doi.org/10.1016/j.electstud.2012.12.006>.
- Sobbrio, Francesco, and Pietro Navarra. 2010. "Electoral Participation and Communicative Voting in Europe." *European Journal of Political Economy* 26(2): 185–207.
- Veiga, Francisco José, and Linda Gonçalves Veiga. 2004. "The Determinants of Vote Intentions in Portugal." *Public Choice* 118(3–4): 341–64.

Appendixes

Appendix 1 - Original Dataset variables description

Variable	Meaning	Initial answer coding
QUESTYEAR	Year of questionnaire	
ID	Inquired ID	
TURNOUT	Participation of the inquired in the last elections	1- Did not voted because he/she could not; 2- Though of voting, but did not do it this time; 3- Usually votes, but did not do it this time; 4- Voted; 98- Does not know; 99- Not answered
DECISION	Party of Coalition the inquired voted for in the last elections	98- Does not know; 99- Not answered
SEX	Inquired Sex	1- Male, 2- Female, 9- Not registered
AGE	Inquired Age	999- not answered
AGEGROUP	Inquired Age Group	1- 18 to 24 years, 2- 25 to 34 years, 3- 35 to 44 years, 4- 45 to 54 years, 5- 55 to 64 years, 6- more than 65 years, 99- not answered
EDUCATION	Education level of the inquired	1- None, 2- Incomplete primary, 3- Complete primary, 4- Incomplete secondary, 5- Complete secondary, 6- Incomplete degree, 7- Complete degree, 98- does not know, 999- not answered
CIVILSTATE	Civil state of the inquired	1- Married, 2- Widow, 3- Divorced, 4- Single, 98- does not know, 99- not answered
WORKINGSITUATION	Working situation of the inquired	1- Working full time, 2- Working part time, 3- Working less than part time, 4- Family worker unpaid, 5- Unemployed, 6- Student, 7- Retired, 8- Permanent disability, 9- domestic, 10- Other, 98- does not know, 99- not answered

SECTORACTIVITY	Activity Sector were the inquired works. If student: Activity sector of main contributor at home. If retired, disable or domestic: last activity sector where they worked.	1- State worker, 2- Dependent worker in private sector, 3- Independent worker in private sector, 4- Mix company worker, 5- Nonprofit organization worker, 97- Not applicable, 98- does not know, 99- not answered
INCOME	Average monthly net income of the aggregate	1- 0 to 300€, 2- 301 to 750€, 3- 751 to 1500€, 4- 1501 to 2500€, 5- more than 2500€, 98- does not know, 99- not answered
RELIGION	Inquired religion	1- Catholic, 2- Other, 3- None, 98- does not know, 99- not answered
NUMBEROFPEOPLE	Number of people in the family aggregate	
SINDICAT	Is the inquired member of a syndicate	1- Yes, 2- No, 98- does not know, 99- Not answered
ECONOMYEVLUTIONPT	Inquired opinion on the evolution of the Portuguese economy in the previous 12 months	1- Improved a lot, 2- Improved a little, 3- Maintained, 4- Worsened a little, 5- Worsened a lot, 98- does not know, 99- not answered
POLITICALINTEREST	Inquired level of political interest	1- Very interested, 2- Reasonably interested, 3- Little interested, 4- Not interested, 98- does not know, 99- not answered
GOVERNMENTEVALUATION	Inquired evaluation to prior government performance	1- Very good, 2- Good, 3- Bad, 4- Very bad, 98- does not know, 99- not answered
WHOISINPOWER	Level of agreement with “who is in power makes a difference”	1- Does not make difference, 2, 3, 4, 5- Makes a big difference, 98- does not know, 99- not answered
WHOPEOPLEVOTE	Level of agreement with “who people voter for makes a difference”	1- Does not make difference, 2, 3, 4, 5- Can make a big difference, 98- does not know, 99- not answered
SIMPATYBE	Level of sympathy with party <i>Bloco de Esquerda</i>	0- Great antipathy, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 - Great sympathy, 97- does not know party, 98- does not know, 99- not answered
SIMPATYCSDPP	Level of sympathy with party <i>CDS-PP</i>	0- Great antipathy, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 - Great sympathy, 97- does not know party, 98- does not know, 99- not answered

SIMPATHYCDU	Level of sympathy with party <i>CDU</i>	0- Great antipathy, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 - Great sympathy, 97- does not know party, 98- does not know, 99- not answered
SIMPATHYPPDPSD	Level of sympathy with party <i>PSD</i>	0- Great antipathy, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 - Great sympathy, 97- does not know party, 98- does not know, 99- not answered
SIMPATHYPS	Level of sympathy with party <i>PS</i>	0- Great antipathy, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 - Great sympathy, 97- does not know party, 98- does not know, 99- not answered
POSITIONBE	Inquired opinion on position of party <i>Bloco de Esquerda</i>	0- Left, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – Right, 97- does not know party, 98- does not know, 99- not answered
POSITIONCDSPP	Inquired opinion on position of party <i>CDS-PP</i>	0- Left, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – Right, 97- does not know party, 98- does not know, 99- not answered
POSITIONCDU	Inquired opinion on position of party <i>CDU</i>	0- Left, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – Right, 97- does not know party, 98- does not know, 99- not answered
POSITIONPPDPSD	Inquired opinion on position of party <i>PSD</i>	0- Left, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – Right, 97- does not know party, 98- does not know, 99- not answered
POSITIONPS	Inquired opinion on position of party <i>PS</i>	0- Left, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – Right, 97- does not know party, 98- does not know, 99- not answered
DEMOCRACYSATISFACTION	Level of inquired satisfaction with democracy	1- Very satisfied, 2- Reasonably satisfied, 3- Not very satisfied, 4- Not satisfied, 98- does not know, 99- not answered
PARTYPROXIMITY	Party that is closer to the inquired	0- None, 1- BE, 2- PCP, 3- PS, 5- IL, 6- CDS-PP, 6- PSD, 8- PAN, 90- Other, 98- does not know, 99- not answered
FREQPAPER	Frequency which the inquired checked political news on journals or magazines during the campaign	1- Daily, 2- 3/4 Days per week, 3- 1/2 Days per week, 4- Less frequently, 5- Never, 98- does not know, 99- not answered
FREQRADIO	Frequency which the inquired checked political news on radio during the campaign	1- Daily, 2- 3/4 Days per week, 3- 1/2 Days per week, 4- Less frequently, 5- Never, 98- does not know, 99- not answered

FREQTV	Frequency which the inquired checked political news on television during the campaign	1- Daily, 2- 3/4 Days per week, 3- 1/2 Days per week, 4- Less frequently, 5- Never, 98- does not know, 99- not answered
FREQNET	Frequency which the inquired checked political news on blogs, websites or social media during the campaign	1- Daily, 2- 3/4 Days per week, 3- 1/2 Days per week, 4- Less frequently, 5- Never, 98- does not know, 99- not answered
FREQRELGIAO	Frequency which the inquired attended religious services	1- Never, 2- Less than once per month, 3- Once per month, 4- Two or more times per month, 5- Once or more per week, 97- Not applicable, 98- does not know, 99- not answered

Appendix 2 - Final Dataset variables description

Variable	Meaning	Final answer coding
QUESTYEAR	Year of questionnaire	
ID	Inquired ID	
TURNOUT	Participation of the inquired in the last elections	0- Did not voted; 1- Voted; 999- Missing value
DECISION	Party or Coalition the inquired voted for in the last elections	0- Left; 1- Right; 998- Other; 999- Missing value
SEX	Inquired Sex	1- Male, 2- Female
AGE	Inquired Age	
AGEGROUP	Inquired Age Group	1- 18 to 24 years, 2- 25 to 34 years, 3- 35 to 44 years, 4- 45 to 54 years, 5- 55 to 64 years, 6- more than 65 years
EDUCATION	Education level of the inquired	1- None, 2- Incomplete primary, 3- Complete primary, 4- Incomplete secondary, 5- Complete secondary, 6- Incomplete degree, 7- Complete degree
CIVILSTATE	Civil state of the inquired	1- Married, 2- Widow, 3- Divorced, 4- Single
WORKINGSITUATION	Working situation of the inquired	1- Working, 2- Not Working,
RELIGION	Inquired religion	1- Catholic, 2- Other, 3- None
SINDICAT	Is the inquired member of a syndicate	1- Yes, 2- No
POLITICALINTEREST	Inquired level of political interest	1- Very interested, 2- Reasonably interested, 3- Little interested, 4- Not interested
DEMOCRACYSATISFACTION	Level of inquired satisfaction with democracy	1- Very satisfied, 2- Reasonably satisfied, 3- Not very satisfied, 4- Not satisfied
PARTYPROXIMITY	Is the inquired close to any party	0- None, 1- Left, 2- Right, 3- Other
FREQPAPER	Frequency which the inquired checked political news on journals or magazines during the campaign	1- Daily, 2- 3/4 Days per week, 3- 1/2 Days per week, 4- Less frequently, 5- Never

FREQRADIO	Frequency which the inquired checked political news on radio during the campaign	1- Daily, 2- 3/4 Days per week, 3- 1/2 Days per week, 4- Less frequently, 5- Never
FREQTV	Frequency which the inquired checked political news on television during the campaign	1- Daily, 2- 3/4 Days per week, 3- 1/2 Days per week, 4- Less frequently, 5- Never
FREQRELGION	Frequency which the inquired attended religious services	1- Never, 2- Less than once per month, 3- Once per month, 4- Two or more times per month, 5- Once or more per week