U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Determinants of political participation: A machine learning approach

Rita Allen Valente Guedes de Pinho

Master Dissertation
Master in Modelling, Data Analysis, and Decision Support Systems

Supervised by:
Prof. Patrício Costa
Professor João Gama

2023

Acknowledgements

I would like to express my gratitude to my supervisor, Professor Patrício Costa, and also to my co-supervisor, Professor João Gama, for their support and guidance throughout this journey. I would also like to thank to my MADSAD teachers for all their teachings.

On a personal level, I am grateful to my family, especially to my parents, for all the love, support, and motivation, and for always believing in me. To my boyfriend a special thanks for all the patience during these last months.

Abstract

This work aims to explore and develop prediction models for political participation. The political participation referred to throughout this work includes both conventional and non-conventional participation and vote. Actions related to political parties are considered conventional participation, and other activities that aim to influence political decisions or pressure politicians are characterised as non-conventional. Non-conventional participation includes loads of actions, which are divided into categories. In this work, the literature review focused on the determinants of political participation and machine learning. Data from the "The Political Participation of Youth in Portugal, 2020" survey and the European Social Survey are used for this project. Ten machine learning algorithms were used to determine the profile of the users of each category of political participation: Support Vector Machine, Random Forest, Cat Boost, Logistic Regression, Naïve Bayes, Decision Tree, K-Nearest Neighbour, XGBoost, Gradient Boosting, and Neural Networks. The area under the curve (AUC), accuracy, precision, recall, specificity, and F-score were computed to evaluate the performance of the models. Also, the Friedman test and the Nemenyi post-hoc were used to address the statistical differences between the models. The best model for each target variable was selected and analysed. Relatively to vote, the best model was Naïve Bayes and the variable with more impact was "portuguesecitizen". In a second analysis, without this variable, "incomefeel" and "politicalinterest" stood out as the ones with more impact. Gradient Boosting was the model with the best performance for conventional and non-conventional participation. The two variables with more impact on conventional and non-conventional participation were the same, "politicalinterest" and "incomecovid19". Considering participation, the model chosen was Logistic Regression, and the variables with more importance were "portuguesecitizen" and "politicalinterest".

Keywords: Political Participation, Machine Learning, Conventional Participation, Non-Conventional Participation

Resumo

Este trabalho tem como objetivo explorar e desenvolver modelos preditivos para a participação política. A participação política referida ao longo deste trabalho inclui participação convencional, participação não convencional e voto. Ações relacionadas com partidos políticos são consideradas participação convencional e outras ações que tenham como objetivo influenciar decisões políticas ou exercer pressão em políticos são caracterizadas como não convencionais. A participação não convencional inclui diversas ações que são divididas por diferentes categorias. Este trabalho apresenta revisão de literatura relativa às determinantes da participação política e *Machine Learning*. Dados do inquérito "A participação política dos jovens em Portugal, 2020", e do Inquérito Social Europeu foram usados neste trabalho. Dez algoritmos de *Machine Learning* foram usados para determinar o perfil daqueles que realizam ações de uma determinada categoria de participação política: *Support Vector Machine, Random Forest, Cat Boost, Logistic Regression, Naïve Bayes, Decision Tree, K-Nearest Neighbour, XGBoost, Gradient Boosting* and *Neural Networks*. Para avaliar a performance dos modelos *area under the curve (AUC), accuracy, precision, recall, specificity* e *F-score* foram calculados. O teste de *Friedman* e o *post-hoc de Nemenyi* foram também utilizados para avaliar as diferenças estatísticas entre os modelos. O melhor modelo para cada uma das quatro variáveis foi selecionado e analisado. Relativamente ao voto, o melhor modelo foi o *Naïve Bayes* e a variável com mais impacto foi "portuguesecitizen", numa segunda análise, sem esta variável, "incomefeel" e "politicalinterest" destacaram-se como as que tinham mais impacto. *Gradient Boosting* foi o modelo com a melhor performance para a participação convencional e não convencional. As duas variáveis com pais impacto para a participação convencional e não convencional foram as mesmas, "politicalinterest" e "incomecovid19". Considerando a participação, o modelo escolhido foi *Logistic Regression* e as variáveis com mais importância foram "portuguesecitizen" e "politicalinterest".

Palavras-chave: Political Participation, Machine Learning, Conventional Participation, Non-Conventional Participation

Table of Contents

List of Tables

List of Figures

1. Introduction

The first democratic elections in Portugal were on the 25th of April 1975, and the abstention rate was 8%. More recently, the turnout rate has been decreasing in every election. Only 51.4% of the eligible population voted in the last general elections.

In the last years, several ways of political participation have appeared. Vote is not the only way of participating in politics. Political participation is now classified into conventional and non-conventional (Ekman & Amnå, 2012). Conventional participation regards all the actions directly related to a political party. Non-conventional participation includes all the activities a person can take to influence political decisions or put pressure on politicians (Van Deth, 2001).

The determinants of the ones who practice each kind of political action differ, so with this dissertation, we aim to study the determinants of political participation with machine learning algorithms.

The primary motivation for this dissertation is to answer the research question: the determinants of political participation using machine learning. To achieve the goal, all the knowledge acquired during the master course was used: the use of several machine learning algorithms to solve the problem, as well as some measures to compare them, and also the use of two different software.

1.1 Problem definition

Political participation is essential in society, it's the only way that the majority of the population can interfere in the decisions of their country to make a change in their community. As it is so important, it's essential to measure it. With this work, we aim to develop an algorithm that predicts political participation in Portugal, whether conventional or non-conventional and also voting.

Conventional participation includes all the forms of participating in politics related to a political party, such as wearing a badge, participating in a political rally, or even donating money to a party. Non-conventional participation concerns the forms of participation that are not related to a party but aim to pressure politicians or influence political decisions. There are

many ways of participating non-conventionally, such as manifesting about labour rights, boycotting a product for ethical reasons, or even signing a petition.

Some works concern political participation, but a machine-learning approach to this subject can be more explored.

1.2 Dissertation framework

This dissertation is organised into five main chapters. The first one comprises the introduction and problem definition. Chapter 2 presents the literature review, focused on the determinants of political participation in the world, the Portuguese case, and the machine learning approach. Chapter 3 introduces the data used as well as the methodological approach. In Chapter 4, the results of the application of the models to the data are presented. The final chapter concerns the main conclusions of this work.

2. Literature Review

In this section, we present the literature review regarding political behaviour. There are four sections, the first one comprises the definition of political participation and also the different forms of participating, the second section focus on the determinants of political Participation, the third one presents the Portuguese case, and the last section contains the machine learning approaches for similar cases. The literature presented considers political participation aspects, such as the forms of political participation, the characteristics of those who participate in politics, and the machine learning algorithms that have been used to predict political participation.

2.1 Political Participation Definition

Political participation can be widely defined as citizens' actions to influence politics (Van Deth J. , 2016). The most common act is voting. However, participation can be divided into contact, party, protest, and consumer participation (Teorell, Torcal, & Montero, 2007). More recently, Teocharis and Van Deth (2018) have classified voting and the new forms of political participation, based on a survey of a sample of the German population, in six different modes: electoral participation (voting), digitally networked participation, institutionalised participation, protest participation, civic participation, and consumerist participation. Digitally networked participation involves acts on social media related to politics, for example, exposing on social media about a political/social issue. Institutionalised participation is related to the forms that deal with parties, such as attending political meetings. Protest participation is the set of protesting, like signing a petition. Civic participation consists of volunteering, and consumerist participation corresponds to the actions taken considering a brand or products for political or ethical reasons, for example, boycotting products. The more recent study includes a digital way of participation and civic participation. (Theocharis & Van Deth, 2018)

Reichert (2016) presented other classification of the different forms of political participation: voting, conventional participation, unconventional participation, and non-normative participation. Voting concerns the participation in the elections. Conventional participation regards the political actions related to a political party, such as supporting an election campaign. Unconventional participation refers to less institutionalized activities, outside political

parties, for example attend a non-violent political protest march. And non-normative participation is related to illegal forms of political demonstrations. This is the classification that is going to be used in this project, except the non-normative participation.

2.2 Determinants of Political Participation

Understand the characteristics of who votes and who doesn't it's important to democracy. Turnout is the probability of an individual participate in the elections (Kim, Alvarez, & Ramirez, 2020). Several studies concern the characteristics of those who vote in the elections. Kaat Smets and Carolien Van Hann (2013) reviewed 90 articles on individual-level turnout. The variables that had a more significant presence on turnout were age and age squared, education, residential mobility, region, media exposure, mobilisation (partisan and non-partisan), vote in the previous election, party identification, political interest, and political knowledge. The article suggests that future works include the following variables as control variables: income, marital status, and religious attendance. The variables that do not affect turnout are gender, race, occupational status and type, citizenship, union membership, trust in institutions, and the closeness of elections.

Nowadays, governments cannot be concerned only with the national population, immigration is an increasingly frequent phenomenon, and this population must also be considered. Bass and Casper (2001) studied the voting behaviour of naturalised Americans. The more established ones in society are most likely to register and vote in the elections. Beyond that, some socioeconomic and demographic variables explain naturalised citizens' registration and voting behaviour. Variables considered positively correlated with the turnout are education, income, professional occupation, age, marital status, length of time at current residence, and length of time in the US, which affect voting behaviour.

Voter behaviour is not only influenced by economic, social, and demographical variables. Also, the identification, or not, with a party influences the individuals in the decision of voting. Individuals without party identification are more likely to be influenced by considering leaders' traits (Da Silva & Costa, 2019).

Kim et al. (2020) used Fuzzy Forests to predict the United States of America 2016 presidential election turnout. Recursive variable elimination was used to choose the key variables. The

variables found as the most important were the fact that a person is registered and if he voted in the election of 2012, which means, in both cases, that he is more likely to turn out. The only demographic and social variables selected were age and retirement.

2.3 The Portuguese Case

Relatively to the Portuguese situation, there aren't many studies in this area. Political participation is more than voting in elections and could be considered multidimensional, with the modes of political participation beyond the vote divided into civic, collective conventional, online, and conventional individual involvement (Costa, 2022).

Magalhães (2022) prepared a report on Youth Political Participation in Portugal, considering the years between 2002 and 2019. The data used was from the European Social Survey. The survey asks about their participation in the last elections and whether they have utilised non-conventional political participation. Relatively to conventional political participation, the Portuguese youth voted less than most young Europeans. Nonetheless, when comparing the non-conventional participation, young Portuguese people are more similar to those from Eastern Europe. The non-conventional ways of political participation used by the young Portuguese people are the boycott of products for social or political reasons, the signing of petitions, and the participation in manifestations and protests.

Costa et al. (2021) have analysed the evolution of the main determinants of voting behaviour during a general election campaign. The data is from the Portuguese 2019 general election, collected in four different moments longitudinally. The main conclusions were that the majority of the respondents maintained their opinions across four periods of time; the educational level influences participation, higher the educational level, higher the participation, and that age affects party identification and turnout.

2.4 The Use of Machine Learning

Concerning the machine learning algorithms, Hua et al. (2021) worked on predicting voter turnout in Malaysian general elections using a decision tree classifier. The models used were Classification and Regression Tree (CART), Chi-squared automatic interaction detection (CHAID), and C5.0. The model that performed better was CHAID. The authors recommended that future research consider other decision tree algorithms like Support Vector

Machine, Random Forest, and Boosting C5.0 to predict the turnout rate.

In their study, Costa et al. (2021), developed five different models to predict the outcome of the elections (abstention, left parties, and right parties). The algorithms used were Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN,) and three different models developed by experts in Political Science using Dynamic Bayesian Networks. NB and TAN were the models with the best overall performance.

3. Data and Methodology

3.1 Data

This project uses data from two sources: The Political Participation of Youth in Portugal, 2020 survey, funded by *Fundação Calouste Gulbenkian*, and the European Social Survey (ESS) conducted in 2020. To achieve the objective of this work, there was the need to merge both datasets to create a new one.

The Political Participation of Youth in Portugal survey has several questions related to attitudes concerning the economic and pandemic situation, political efficacy, populist attitudes, as well as social, civic, and political involvement. The survey gathered responses from a total of 1464 participants, consisting of 750 women and 714 men.

The European Social Survey is a biennial survey conducted across all European countries. It covers multiple areas of interest. The relevant aspects of this study include politics, gender, socio-demographics, citizen involvement, ageism, democracy, and work and well-being. Since this work focus solely on Portuguese participants, only the corresponding data from the ESS was extracted. The final sample has a total size of 1838, with 772 men and 1066 women.

As the two datasets initially contained variables that were not common to both, a specific process was applied to merge them effectively:

1. Identify the common variables across the datasets
2. Exclude variables that are not present in both datasets
3. Standardise the variables by recoding them to ensure a consistent structure across both datasets
4. Merge the datasets, combining the information from the two sources into a unified dataset

In addition to the previously elucidated merging procedure, an alternative methodology was necessitated to effectively account for the variable associated with the educational level of the participant's parents. The ESS had separate variables for the education levels of the father and the mother. At the same time, The Political Participation of Youth in Portugal survey had a single variable labelled "Highest level of education of respondent parents". To

reconcile this discrepancy, a new variable was created within the ESS dataset to keep the value of the parent that achieved the highest level of education.

The primary objective of this project is to assess political participation, including conventional and non-conventional participation, as well as voting. Three distinct variables were created. Conventional participation comprises all the actions taken by an individual related to a political party. Non-conventional participation includes actions that aim to influence political decisions or even to pressure politicians. Conventional and non-conventional participation were derived by combining multiple relevant variables, while political participation was computed as combining conventional and non-conventional participation and vote. An individual is considered to have participated in politics if he has engaged in conventional or non-conventional methods or voted.

Once the variables that are presented in both datasets were identified, the next step was to recode them to enable the merging.

Table 1 Dataset transformation

Final	ESS	FCG
Gender Gender of the respondent 1-Male (1/1) 2-Female (2/2)	**Gndr** Gender of the respondent 1-Male 2-Female	**A1** Gender of the respondent 1-Male 2-Female
Agegroup Age group of the respondent 1-15:34 (15-34/1) 2-35:64 (35-64/2) 3-65+ (65+/3)	**Agea** Age of the respondent Age number	**A2.1** Age group of the respondent 1-15:34 2-35:64 3-65+
Region Region where the respondent lives 1-Norte (PT11/1) 2-Centro (PT16/2) 3-Lisboa (PT17/3) 4-Alentejo (PT18/4) 5-Algarve (PT15/5) 6-Açores (PT20/6) 7-Madeira (PT30/7)	**Region** Region where the respondent lives PT11 Norte, PT15 Algarve, PT16 Centro, PT17 Área Metropolitana de Lisboa, PT18 Alentejo, PT20 Região Autónoma dos Açores, PT30 Região Autónoma da Madeira	**NUTSII** Region where the respondent lives 1-Norte 2-Centro 3-Lisboa 4-Alentejo 5-Algarve 6-Açores 7-Madeira
Leveleducation Level of education of the respondent 1-None (1/1) 2-1st cycle of basic education (2/2) 3-2nd cycle of basic education (3 4/3) 4-3rd cycle of basic education (5 6 7/4) 5-Secondary school (8 9/5)	**Edlvdpt** Level of education of the respondent *	**A5** Level of education of the respondent 1-None 2-1st cycle of basic education 3-2nd cycle of basic education 4-3rd cycle of basic education 5-Secondary school 6-Higher education

6-technical specialisation course (10/) 7-Higher education (11 12 13 14 15 16 17/ 6)		
Politicalinterest How much does politics interest to the respondent 1-Not at all interested (4/4) 2- Hardly interested (3/3) 3- Quite interested (2/2) 4- Very interested (1/1)	**Polintr** How much does politics interest to the respondent 1-Very interested 2-Quite interested 3-Hardly interested 4-Not at all interested	**Q9** How much does politics interest to the respondent 1-Very interested 2-Quite interested 3-Hardly interested 4-Not at all interested
Tradeunion The respondent belongs or ever belonged to a trade union 1-Yes (1 2/ 1 2 3) 2-No (3/4)	**Mbtru** The respondent belongs or ever belonged to a trade union 1-yes, currently 2-yes, previously 3-no	**Q11_2** The respondent belongs or ever belonged to a trade union 1-belongs and actively participates 2-belongs but doesn't actively participate 3-once belonged but no longer belongs 4-never belonged
Leftright Respondent political position 1-Left (0 1 2 3 4/ 0 1 2 3 4) 2-Center (5/5) 3-Right (6 7 8 9 10/ 6 7 8 9 10)	**Lrscale** Respondent political position 0-Left,1,2,3,4,5,6,7,8,9,10-right	**Q17** Respondent political position 0-Left,1,2,3,4,5,6,7,8,9,10-right
Mainactivity Respondent main activity 1-Paid work (1/2) 2-Education (2/ 1 4) 3-Unemployed looking for job (3/ 5 6) 4-Unemployed not looking for job (4/9) 5-Unable to work (5/7) 6-Retired (6/10) 7-Housework (8/8) 8-Other (7 9/ 3 98)	**Mnactic** Respondent main activity 1-Paid work 2-Education 3-Unemployed looking for job 4-Unemployed not looking for job 5-Permanently sick or disabled 6-Retired 7-Community or military service 8-Housework, looking after children, other 9-Other	**Q28** Respondent main activity 1-Student 2-Worker 3-Student-worker 4-Attending a vocational training course 5-Unemployed looking for the first job 6-Unemployed for a new job 7-Unable to work 8-Responsible for household tasks 9-Unoccupied (don't work, don't look for a job, don't study) 10-Retired
Active The respondent is part of the active population 1-Yes 2-No		
Emplrelation Respondent employment relation 1-Employee (1/4) 2-Employer (2/ 1 2 3) 3-Other situation (3/ 5 96)	**Emplrel** Respondent employment relation 1-Employee 2-Self-employed 3-Working for own family business	**Q29** Respondent employment relation 1-Employer (with employees) 2-Self-employed 3-Self-employed, on green receipts 4-Employee 5-Unpaid family worker 96-Other situation
Contract	**Wrkctra**	**Q30**

Respondent work contract type 1-Unlimited (1/1) 2-Limited (2/ 2 4) 3-No contract (3/ 3 5)	Respondent work contract type 1-Unlimited 2-Limited 3-No contract	Respondent work contract type 1-Effective 2-Term employment contract 3-Contract for services (green receipts) 4-Scholarship contract/internship 5-no contract
Incomefeel Respondent feeling about present income 1- It's very hard to live with the current income (4/4) 2- It's hard to live with the current income (3/3) 3- Current income allows to live reasonably (2/2) 4- Current income allows to live comfortably (1/1)	**Hincfel** Respondent feeling about present income 1-Current income allows to live comfortably 2-Current income allows to live reasonably 3- It's hard to live with the current income 4-It's very hard to live with the current income	**Q31** Respondent feeling about present income 1-Current income allows to live comfortably 2-Current income allows to live reasonably 3- It's hard to live with the current income 4-It's very hard to live with the current income
Incomecovid19 What happened to respondent income with covid-19 1-Reduced (1/1 2) 2-Didn't reduce (0/ 3 4 5)	**Hapirc19** What happened to respondent income with covid-19 0-No 1-Yes	**Q32** What happened to respondent income with covid-19 1-Decreased a lot 2-Decreased a little 3-Stayed the same 4-Increased a little 5-Increased a lot
Portuguesecitizen Respondent is a Portuguese citizen? 1-Yes (1/1) 2-No (2/2)	**Ctzcntr** Respondent is a Portuguese citizen? 1-Yes 2-No	**Q34** Respondent is a Portuguese citizen? 1-Yes 2-No
Maritalstatus Respondent marital status 1-Ever been in a relationship (1 2 3 4 5/ 2 3 4 5) 2-Never been in a relationship (6/1)	**Maritalb** Respondent marital status 1-Legally married 2-In a legally registered civil union 3-Legally separated 4-Legally divorced/civil union dissolved 5-Widowed/civil partner died 6-None of these (NEVER married or in legally registered civil union)	**Q36** Respondent marital status 1-Single 2-Married 3-Nonmarital partnership 4-Separated/divorced 5-Widowed
Parentseducation Highest level of education of respondent parents 1-None (1/1) 2-1st cycle of basic education (2/2) 3-2nd cycle of basic education (3 4/3) 4-3rd cycle of basic education (5 6 7/4) 5-Secondary school (8 9/5) 6-technical specialisation course (10 /) 7-Higher education (11 12 13 14 15 16 17/ 6)	**Edlvfdpt & edlvmdpt** Highest level of education of the father and mother of the respondent *	**Q38** Highest level of education of respondent parents 1-None 2-1st cycle of basic education 3-2nd cycle of basic education 4-3rd cycle of basic education 5-Secondary school 6-Higher education

Religiousattendance	**Rlgatnd**	**Q40**
How often do the respondent attend to religious events	How often do the respondent attend to religious events	How often do the respondent attend religious events
1- Never (7/7) 2- Even less times (6/6) 3- Only on holy days (5/5) 4-At least once a month (4/4) 5- Once a week (3/3) 6- More than once a week (2/2) 7- Everyday (1/1)	1-Everyday 2-More than once a week 3-Once a week 4-At least once a month 5-Only on holy days 6-Even less times 7-Never	1-Everyday 2-More than once a week 3-Once a week 4-At least once a month 5-Only on holy days 6-Even less times 7-Never

*1-Nenhum; 2-Ensino Básico (até à 4ª classe, instrução primária (3º ou 4º ano)); 3-Ensino Básico 2 (preparatório, 5º e 6º anos/classe, 1º ciclo dos liceus/ do ensino técnico comercial ou industrial); 4-Cursos de educação e formação de tipo 1. Atribuição de 'Diploma de qualificação profissional de nível 1'; 5-Ensino Básico 3 (9º ano; 5º ano dos liceus; escola comercial/industrial; 2º ciclo dos liceus ou do ensino técnico); 6-Cursos de educação e formação de tipo 2. Atribuição de 'Diploma de qualificação profissional de nível 2'; 7-Cursos de educação e formação de tipo 3 e 4. Atribuição de 'Diploma de qualificação profissional de nível 2'; 8-Ensino Secundário – cursos científico-humanísticos (12º ano; 7ºano dos liceus; propedêutico; serviço cívico); 9-Ensino Secundário – cursos tecnológicos, artísticos especializados, ou profissionais. CEFs de tipo 5,6 e 7; 10-Cursos de especialização tecnológica. Atribuição de 'Diploma de Especialização Tecnológica'; 11-Ensino superior politécnico: bacharelato de 3 anos; antigos cursos médicos; 12-Ensino superior politécnico: licenciaturas de 3-4 anos curriculares; licenciatura complemento de formação; 13-Ensino superior universitário: licenciaturas de 3-4 anos curriculares; licenciatura bietápica de 4 anos; 14-Pós-graduação: especialização pós-licenciatura sem atribuição de grau académico, MBA; 15-Ensino superior universitário: licenciatura com mais de 4 anos curriculares; licenciatura bietápica de 5 anos; 16-Mestrado (inclui Mestrado Integrado); 17-Douturamento; 5555-Other

3.2 Target Variables

As stated before, in this work there are four target variables. They are vote, conventional participation, non-conventional participation, and participation.

The variable vote, expresses if the respondents have vote or not in the national elections of 2019. This variable was the same for both data sets and had only one transformation, which was to eliminate all the respondents of the ESS that weren't eligible to vote.

Relatively to conventional participation, both data sets had some variables that were included in conventional participation.

Table 2 Variables that composes conventional participation variable

ESS	FCG
Contplt: respondent has contacted a politician or government official last 12 months	**Q11_1**: Respondent belongs/belonged to a political party
Donprty: respondent has donated to or participated in a political party or pressure group last 12 months	**Q14_3**: Respondent attended a party or candidate rally
Badge: respondent worn or displayed campaign badge/sticker last 12 months	**Q14_4**: Respondent contacted or tried to contact, a politician or other public official to express his opinions
	Q14_5: Respondent has donated money or collected funds for a social, civic, or political activity
	Q14_10: Respondent distributed political flyers

The variables from the ESS were already code in 1-Yes and 2-No. Relatively to the variables from the FCG survey, they were coded in a different way. The recodification is in the following table.

Table 3 Recodification of the conventional variable from the FCG survey

1-Didn't and never would	2-No
2-Didn't but could have	
3-Did it in a more distant past	1-Yes
4-Did it during the last year	

The non-conventional participation variable is also a combination of variables from the ESS and the FCG survey. The process of recodification was the same for the conventional participation variable.

Table 4 Variables that compose non-conventional participation variable

ESS	FCG
Sgnptit: Respondent signed a petition last 12 months	**Q14_1**: Respondent has signed a petition
Pbldmna: Respondent taken part in public demonstration last 12 months	**Q14_2**: Respondent has boycotted certain products for political issues or to favour the environment
Bctprd: Respondent boycotted certain products last 12 months	**Q14_6**: Respondent has expressed his opinions to the media
Pstplonl: Respondent posted or shared anything about politics online last 12 months	**Q14_7**: Respondent participated in a forum or in a online political discussion group
Volunfp: Respondent volunteered for not-for-profit or charitable organisation	**Q14_8**: Respondent participated in a demonstration related to social, civic or political issues
	Q14_9: Respondent did volunteer
	Q14_11: Respondent wrote or graffiti political messages on walls
	Q14_12: Respondent published, commented or shared contents online, about political or social issues

The last variable is the participation, which is also coded as 1-Yes and 2-No. It comprises vote, conventional and non-conventional participation. If the respondent practices at least one of the forms of participation in politics, it is considered as participating politically.

3.3 Pre-processing

Before analysing the data, there was the need to analyse missing values. In this dataset, responses such as "doesn't know", "no answer", or "doesn't remember" were considered as missing values. To determine the best approach to deal with missing values, several steps were taken. The first one was to analyse the percentage of missing values for each variable.

Table 5 Percentage of missing values per variable

Variable	Missing Values	%
gender	0	0
agegroup	1	0
region	0	0
leveleducation	0	0
politicalinterest	8	0.2
tradeunion	16	0.5
leftright	484	14.7
mainactivity	17	0.5
active	17	0.5
emplrelation	754	22.8
contract	1163	35.2
incomefeel	26	0.8
incomecovid19	13	0.4
portuguesecitizen	1	0
maritalstatus	12	0.4
parentseducation	114	3.5
religiousattendance	15	0.5
vote	163	4.9
conventionalparticipation	28	0.8
nonconventionalparticipation	28	0.8
participation	55	1.7

Variables "emplrelation" and "contract" were excluded from the dataset due to their high percentage of missing values. Despite the high proportion of missing values in the "leftright" variable, given its relevance, since it is related to politics, the decision was to keep it.

The next phase aimed to determine whether the missing values are missing completely at random or not so we can decide to impute or eliminate them. The test used was Little's MCAR this analysis's results are in the appendix. Given that the majority of the missing values are not completely at random (MCAR), the decision was to eliminate all of them.

Having the data cleaned, it's possible to do some analysis of the dataset.

3.4 Participants Analysis

This section aims to do a preliminary analysis of the cleaned dataset to understand the data.

Table 6 Characteristics of the sample for vote

		Yes		No		Total	
		n	%	n	%	n	%
Gender	**Male**	987	85.5%	168	14.5%	1155	46.7%
	Female	1074	81.4%	246	18.6%	1320	53.3%
Age group	**15-34**	500	81.8%	111	18.2%	611	24.7%
	35-64	1027	83.9%	197	16.1%	1224	49.5%
	65+	534	83.4%	106	16.6%	640	25.9%
Education	**None**	48	76.2%	15	23.8%	63	2.5%
	1st cycle of basic education	360	79.5%	93	20.5%	453	18.3%
	2nd cycle of basic education	224	83.3%	45	16.7%	269	10.9%
	3rd cycle of basic education	344	80.2%	85	19.8%	429	17.3%
	Secondary school	519	82.5%	110	17.5%	629	25.4%
	Technical specialisation course	9	100%	0	0%	9	0.4%
	Higher education	557	89.4%	66	10.6%	623	25.2%
Total		2061	83.3%	414	16.7%	2475	100%

Table 5 shows the characterisation of the vote per sociodemographic variable. Relatively to gender, both males and females have turnout rates higher than 80%, with the males having the highest (85.5%). The conclusion is the same for the variable "agegroup", respondents aged between 35 and 64 have a higher turnout rate (83.9%), but every age group has a turnout rate higher than 80%. Considering the level of education, there is no clear relation with the turnout rate, but the highest rates belong to "Technical specialisation course" and "Higher education", which are the highest levels of education in this dataset. Vote is really unbalanced, 83.3% of the respondents answered that they voted in the national elections of 2019.

Table 7 Characteristics of the sample for conventional participation

		Yes		No		Total	
		n	%	n	%	n	%
Gender	**Male**	482	41.7%	673	58.3%	1155	46.7%
	Female	513	38.9%	807	61.1%	1320	53.3%
Age group	**15-34**	272	44.5%	339	55.5%	611	24.7%
	35-64	529	43.2%	695	56.8%	1224	49.5%
	65+	194	30.3%	446	69.7%	640	25.9%
Education	**None**	20	31.7%	43	68.3%	63	2.5%
	1st cycle of basic education	117	25.8%	336	74.2%	453	18.3%
	2nd cycle of basic education	108	40.1%	161	59.9%	269	10.9%
	3rd cycle of basic education	175	40.8%	254	59.2%	429	17.3%
	Secondary school	264	42%	365	58%	629	25.4%
	Technical specialisation course	3	33.3%	6	66.7%	9	0.4%
	Higher education	308	49.4%	315	50.6%	623	25.2%
Total		995	40.2%	1480	59.8%	2475	100%

The practice of conventional participation is more balanced. However, more people never practised conventional participation (59.8%) than the ones who did (40.2%). According to "gender", more men are participating than women. Relatively to the "agegroup", the younger, the highest participation. Considering the level of education, there is no pattern, but the ones with higher education have the highest rate of conventional participation (49.4%).

Table 8 Characteristics of the sample for non-conventional participation

		Yes		No		Total	
		n	%	n	%	n	%
Gender	**Male**	642	55.6%	513	44.4%	1155	46.7%
	Female	695	52.7%	625	47.3%	1320	53.3%
Age group	**15-34**	445	72.8%	166	27.2%	611	24.7%
	35-64	679	55.5%	545	44.5%	1224	49.5%
	65+	213	33.3%	427	66.7%	640	25.9%
Education	**None**	26	41.3%	37	58.7%	63	2.5%
	1st cycle of basic education	111	24.5%	342	75.5%	453	18.3%
	2nd cycle of basic education	132	49.1%	137	50.9%	269	10.9%
	3rd cycle of basic education	230	53.6%	199	46.4%	429	17.3%
	Secondary school	371	59%	258	41%	629	25.4%
	Technical specialisation course	7	77.8%	2	22.2%	9	0.4%
	Higher education	460	73.8%	163	26.2%	623	25.2%
Total		1337	54%	1138	46%	2475	100%

Non-conventional participation is also balanced, 54% of the participants have already participated in a non-conventional form. Males have a highest rate of participation than females (55.6% against 52.7%). As for conventional participation, the younger the participants, the higher the participation rate in a non-conventional form. Except for "1st cycle of basic education" and "Higher education", the non-conventional participation rate increases with the level of education.

Table 9 Characteristics of the sample for political participation

		Yes		No		Total	
		n	%	n	%	n	%
Gender	**Male**	1050	90.9%	105	9.1%	1155	46.7%
	Female	1173	88.9%	147	11.1%	1320	53.3%
Age group	**15-34**	560	91.7%	51	8.3%	611	24.7%
	35-64	1111	90.8%	113	9.2%	1224	49.5%
	65+	552	86.3%	88	13.8%	640	25.9%
Education	**None**	49	77.8%	14	22.2%	63	2.5%
	1st cycle of basic education	380	83.9%	73	16.1%	453	18.3%
	2nd cycle of basic education	243	90.3%	26	9.7%	269	10.9%
	3rd cycle of basic education	380	88.6%	49	11.4%	429	17.3%
	Secondary school	564	89.7%	65	10.3%	629	25.4%
	Technical specialisation course	9	100%	0	0%	9	0.4%
	Higher education	598	96%	25	4%	623	25.2%
Total		2223	89.8%	252	10.2%	2475	100%

Regarding political participation, almost every respondent has participated (89.8%). The conclusions for each variable are similar to the ones derived for vote, conventional and non-conventional participation. Males have a higher percentage than females. The younger participants are the ones with higher participation and relatively to education, the ones with no education are the ones with the lower rate, and participants with the highest levels of education are the ones with the highest rates.

3.5 Data Modelling

After treating all the data, it's time to apply the machine learning algorithms. The objective is to determine the participants' profile for each mode of political participation. The algorithms that will be used are the ones that were worked on during the master and that are the most suitable for this problem: Decision Tree, Random Forest, Cat Boost, Neural Network, Gradient Boosting, Naïve Bayes, XGBoost, Logistic Regression, Support Vector Machine, and K-Nearest Neighbor. All these algorithms will be tested in several ways. Feature selection will be applied to choose the best variables to predict the outcome. Also, the optimisation of the hyperparameters will be used.

- In a Decision Tree algorithm, instances are classified by passing through nodes from the root to a leaf node, based on their feature values. Each node signifies a decision or a test condition on an instance attribute, and each branch corresponds to a

potential value for that feature. (Dey, 2016)

- Random forest is an ensemble method that uses bagging to create several decision trees. The output of all the decision trees is combined to make the final prediction. (Alzubi, Nayyar, & Kumar, 2018)

- Cat Boost is also known as Categorical Boosting. It transforms categorical values into numbers by utilizing a range of statistics on combinations of categorical features, as well as combinations of both categorical and numerical features. (Chaplot, Pandey, Kumar, & Sisodia, 2023)

- Neural Network is inspired by the structure of the human brain. It consists of interconnected nodes organized in layers. These networks learn from data to perform tasks such as classification, regression, or pattern recognition. (Dey, 2016)

- Gradient Boosting is also an ensemble model. It combines multiple weaker models, usually decision trees, to create a stronger predictive model, by focusing on the errors of the previous models.

- Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes theorem. Bayes theorem computes the probability of an event, based on prior knowledge of circumstances that are related to the event. (Chaplot, Pandey, Kumar, & Sisodia, 2023)

- XGBoost, extreme gradient boosting, is a variation of Gradient Boosting. XGBoost has an improved performance since it's optimized. (Chaplot, Pandey, Kumar, & Sisodia, 2023)

- Logistic Regression estimates the probability of occurrence of an event, using a logistic function.

- In Support Vector Machine, every data point is represented as a point in an n-dimensional space, where n corresponds to the number of features in the dataset. Each feature's value is aligned with its respective coordinate. The algorithm categorizes data into distinct classes by identifying a line (or hyperplane) that divides the training dataset into these classes. Its effectiveness lies in maximizing the distance between

the closest data points (in each class) and the hyperplane. (Alzubi, Nayyar, & Kumar, 2018)

- In Nearest Neighbor, the algorithm identifies the k-nearest neighbors, neighbors are the cases similar to the target, to predict the outcome. (Chaplot, Pandey, Kumar, & Sisodia, 2023)

3.6 Performance Metrics

After creating the models, it was necessary to validate them. To do that, the data set was divided into two: train and test split, with a split ratio 70/30.

After testing all the models, it was necessary to compare them to choose the best one. Accuracy, precision, recall, specificity, and F-score were used to evaluate each model based on the confusion matrices.

Confusion Matrix		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FN
	Negative	FP	TN

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{n}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{F} - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Also, the area under the curve (AUC) from the Receiving Operating Characteristics (ROC curve) were computed. The ROC curve was also computed based on the confusion matrix, and each model was represented by a curve based on the true positive and false positive rates (Fawcett, 2006).

Friedman test was used to evaluate if there were statistically significant differences between the models:

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

To rank the models, the Nemenyi post-hoc test was used:

$$q = \frac{\overline{R_{J1}} - \overline{R_{J2}}}{\sqrt{\frac{k(k+1)}{6n}}}$$

3.7 Software

Data analysis and the application of several machine learning models to the data was conducted throughout this work to achieve the goal. The software that was used in developing this work is IBM SPSS Statistics version 29 for data analysis and data treatment, and Python in Jupyter Notebooks for the machine learning part.

The main packages used were:

- Autorank, this package was used in the comparison of the models, it also computed the critical difference diagram (Herbold, 2020).
- Shap, this package was used to understand and explain the outcome of the models used in this project (Welcome to the shap documentation, 2023).

4. Results

This section focuses on the central question of this dissertation, explaining the determinants of political participation, conventional participation, non-conventional participation, and voting through the utilisation of machine learning algorithms.

A uniform procedure was applied across the four dependent variables. Initially, all the algorithms were used, combining feature selection methods and data calibration techniques. The feature selection methods applied were no feature selection, extra trees, and selecting the 10 best. Relatively to calibration, no calibration, undersampling, oversampling, and SMOTE were used. Subsequently, the Friedman test, followed by the Nemenyi posthoc analysis, was conducted to identify the optimal algorithm for each case. Lastly, SHAP values were computed to determine which variables are the most influential.

4.1 Feature Selection

As previously mentioned, two feature selection techniques were applied, resulting in the following outcomes across the four target variables:

Variables	Extra trees for vote	Select 10 best for vote	Extra trees for conventional participation	Select 10 best for conventional participation	Extra trees for non-conventional participation	Select 10 best for non-conventional participation	Extra trees for participation	Select 10 best for participation
Gender		X	X		X		X	
Agegroup	X		X	X	X	X		X
Region	X	X	X	X	X		X	
Leveleducation	X	X	X	X	X	X	X	X
Politicalinterest	X	X	X	X	X	X	X	X
Tradeunion		X		X				X
Leftright	X		X		X		X	
Mainactivity	X	X	X	X	X	X	X	X
Active				X		X		X
Incomefeel	X	X	X	X	X	X	X	X
Incomecovid19	X			X		X		X
Portuguesecitizen		X						X
Maritalstatus						X	X	
Parentseducation	X	X	X	X	X	X	X	X
Religiousattendance	X	X	X		X	X	X	

As can be noticed, all variables were selected by the feature selection techniques at least once. In the cases of the vote, conventional participation, and non-conventional participation, both selection methods shared seven common variables. While in participation, only five variables are common to both approaches. The combination of feature selection techniques and data calibration methods led to twelve distinct ways of application of the algorithms.

4.2 Models' Comparison

The next tables present all models' accuracy, precision, recall, F-score, specificity, and AUC for all the target variables.

Relatively to the models, none stands out as better. Considering vote, Logistic Regression was the model with the highest accuracy, precision, F-score and AUC; Support Vector Machine was the model with the highest recall; and Naïve Bayes was the model with the highest specificity.

In the case of conventional participation, with the highest accuracy and precision, Logistic Regression, with the highest recall and specificity Naïve Bayes, with the highest F-score Cat Boost and with the highest AUC Gradient Boosting.

Considering non-conventional participation, Gradient Boosting was the model with the best accuracy and AUC, Logistic Regression with the best precision, Support Vector Machine with the best recall, and Neural Network with the best F-score and specificity.

Relatively to participation, Logistic Regression was the model with the best performance in all the performance metrics.

Regarding the calibration techniques, no calibration has the highest performance, but undersampling also performs well. Relatively to the feature selection techniques, no feature selection stands out as the best.

Table 10 Performance metrics for vote without feature selection

No FS	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.747	0.8398	0.8425	0.8425	0.8479	0.8573	0.8102	0.8573	0.8546	0.821	0.5559	0.646	0.6393	0.6581	0.638	0.8022	0.5882	0.6528	0.6568	0.6016
Precision	0.8731	0.8675	0.87	0.87	0.8686	0.8628	0.8765	0.8628	0.8546	0.8638	0.888	0.9133	0.9238	0.9262	0.9217	0.8801	0.9207	0.9414	0.9241	0.8879
Recall	0.8236	0.9591	0.9591	0.9591	0.9685	0.9906	0.9055	0.9921	1	0.9386	0.5496	0.6472	0.6299	0.652	0.6299	0.8898	0.5669	0.6331	0.652	0.611
F-score	0.8476	0.911	0.9124	0.9124	0.9159	0.9223	0.8908	0.9224	0.9216	0.8996	0.679	0.7576	0.7491	0.7652	0.7484	0.8849	0.7018	0.7571	0.7645	0.7239
Specificity	0.2222	0.3659	0.3953	0.3953	0.4286	0.5714	0.3103	0.5833	imp	0.2642	0.1829	0.2355	0.2419	0.2534	0.2394	0.3069	0.2188	0.2627	0.2508	0.1928
AUC	0.5621	0.6823	0.7195	0.6875	0.7234	0.7124	0.6686	0.7365	0.6266	0.6052	0.5711	0.698	0.7169	0.7206	0.7039	0.7224	0.6715	0.7341	0.698	0.6075
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.7564	0.8197	0.7604	0.6918	0.7012	0.7847	0.7631	0.6622	0.6999	0.6608	0.7039	0.7793	0.7402	0.7026	0.6918	0.6918	0.7376	0.6824	0.6972	0.6837
Precision	0.8661	0.8744	0.8853	0.898	0.9073	0.8874	0.8806	0.9267	0.9008	0.8748	0.8766	0.8867	0.8891	0.8935	0.8919	0.8949	0.8846	0.903	0.8927	0.8802
Recall	0.8457	0.9213	0.8268	0.7213	0.7244	0.8567	0.8362	0.6567	0.7291	0.7039	0.7606	0.8504	0.7953	0.7402	0.7276	0.7244	0.7969	0.7039	0.7339	0.7291
F-score	0.8558	0.8972	0.855	0.8	0.8056	0.8718	0.8578	0.7687	0.8059	0.7801	0.8145	0.8682	0.8396	0.8096	0.8014	0.8007	0.8384	0.7912	0.8055	0.7976
Specificity	0.2033	0.3243	0.2667	0.2403	0.2585	0.3	0.2571	0.256	0.2489	0.1897	0.2083	0.291	0.2571	0.2396	0.2311	0.2358	0.2456	0.2419	0.2353	0.2074
AUC	0.5393	0.6772	0.6824	0.6768	0.6848	0.702	0.646	0.7288	0.6916	0.5993	0.5635	0.6608	0.6628	0.6696	0.6633	0.6844	0.6494	0.6799	0.6697	0.6062

Table 11 Performance metrics for vote extra trees feature selection

Extra trees	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.7416	0.8223	0.8304	0.8466	0.8466	0.8358	0.8048	0.8506	0.8546	0.8237	0.5411	0.6164	0.6433	0.6716	0.6299	0.6433	0.576	0.6299	0.6555	0.6164
Precision	0.859	0.8619	0.859	0.8583	0.8643	0.8597	0.8746	0.8541	0.8546	0.8652	0.8769	0.8995	0.9186	0.9048	0.9147	0.9243	0.9	0.9286	0.922	0.8889
Recall	0.8346	0.9433	0.9591	0.9827	0.9732	0.9654	0.9008	0.9953	1	0.9402	0.5386	0.6205	0.6394	0.6882	0.6252	0.6346	0.5669	0.6142	0.652	0.6299
F-score	0.8466	0.9008	0.9063	0.9163	0.9156	0.9095	0.8875	0.9193	0.9216	0.9011	0.6673	0.7344	0.7539	0.7818	0.7428	0.7526	0.6957	0.7393	0.7638	0.7373
Specificity	0.1667	0.25	0.2353	0.3125	0.3929	0.2667	0.2921	0	imp	0.283	0.17	0.2098	0.2392	0.2385	0.2298	0.2443	0.1983	0.2415	0.2483	0.198
AUC	0.5242	0.6508	0.6795	0.9807	0.6983	0.6932	0.631	0.7118	0.5897	0.6078	0.5464	0.6668	0.687	0.6892	0.6805	0.6986	0.6297	0.7078	0.6871	0.6062
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.7376	0.7968	0.7564	0.7268	0.7026	0.6487	0.751	0.6514	0.7026	0.6595	0.7201	0.7671	0.7281	0.7093	0.6878	0.6528	0.712	0.6676	0.7066	0.6649
Precision	0.8667	0.8656	0.8834	0.8971	0.9043	0.9193	0.8738	0.9215	0.9043	0.8805	0.8739	0.8812	0.8887	0.9084	0.8943	0.9089	0.8766	0.9076	0.9049	0.8845
Recall	0.8189	0.9024	0.8236	0.7685	0.7291	0.6457	0.8283	0.6472	0.7291	0.6961	0.7858	0.8409	0.7795	0.7339	0.7197	0.6598	0.7717	0.6803	0.7339	0.6992
F-score	0.8421	0.8836	0.8525	0.8278	0.8073	0.7586	0.8504	0.7604	0.8073	0.7775	0.8275	0.8606	0.8305	0.8118	0.7976	0.7646	0.8208	0.7777	0.8104	0.781
Specificity	0.1958	0.2346	0.2583	0.2613	0.2554	0.2424	0.227	0.2458	0.2554	0.1992	0.2093	0.2628	0.2473	0.2652	0.2328	0.234	0.212	0.2397	0.2588	0.2075
AUC	0.5361	0.6456	0.6531	0.6651	0.6684	0.6841	0.6311	0.7012	0.681	0.6048	0.5544	0.6479	0.6692	0.6684	0.6829	0.6757	0.6269	0.6885	0.6837	0.6228

Table 12 Performance metrics for vote select 10 best feature selection

Select 10 best	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.7577	0.8197	0.8452	0.8479	0.8479	0.856	0.8143	0.8587	0.8546	0.8291	0.5262	0.6164	0.646	0.7052	0.646	0.8022	0.5707	0.6326	0.6528	0.6326
Precision	0.8748	0.8744	0.8714	0.8635	0.8707	0.8626	0.8782	0.861	0.8546	0.866	0.8714	0.9014	0.9026	0.9047	0.9115	0.8742	0.8892	0.931	0.9143	0.8884
Recall	0.8362	0.9213	0.9606	0.9764	0.9653	0.989	0.9087	0.9953	1	0.9465	0.5228	0.6189	0.6567	0.7323	0.6488	0.8976	0.5685	0.6157	0.6551	0.652
F-score	0.8551	0.8972	0.9139	0.9165	0.9156	0.9215	0.8932	0.9233	0.9216	0.9044	0.6535	0.7339	0.7603	0.8094	0.758	0.8858	0.6936	0.7412	0.7633	0.752
Specificity	0.2353	0.3243	0.4186	0.4	0.4359	0.5333	0.3256	0.6667	imp	0.3061	0.163	0.2117	0.2242	0.2576	0.2337	0.2857	0.1869	0.2446	0.2396	0.2022
AUC	0.5718	0.6171	0.6897	0.6859	0.7014	0.6974	0.6341	0.7232	0.6211	0.5744	0.5281	0.6638	0.6901	0.707	0.6722	0.7083	0.6179	0.721	0.6918	0.6046
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.7443	0.7833	0.7349	0.6904	0.6743	0.7806	0.7322	0.6528	0.6837	0.677	0.677	0.7322	0.7349	0.6635	0.6891	0.7039	0.7227	0.6756	0.6891	0.6931
Precision	0.8702	0.8774	0.8815	0.901	0.9002	0.8831	0.8746	0.9217	0.8968	0.8748	0.8705	0.8772	0.9026	0.9053	0.9073	0.8998	0.8851	0.902	0.9073	0.8832
Recall	0.8236	0.8677	0.7969	0.7165	0.6961	0.8567	0.8016	0.6488	0.7118	0.726	0.7307	0.7984	0.7732	0.6772	0.7087	0.7354	0.7764	0.6961	0.7087	0.7386
F-SCORE	0.8463	0.8725	0.8371	0.7982	0.7851	0.8697	0.8365	0.7616	0.7937	0.7935	0.7945	0.8359	0.8329	0.7748	0.7958	0.8094	0.8272	0.7858	0.7958	0.8045
Specificity	0.2113	0.2696	0.2367	0.2437	0.2341	0.2835	0.2174	0.2466	0.2343	0.1944	0.1857	0.2242	0.2764	0.2351	0.251	0.25	0.2366	0.2372	0.251	0.217
AUC	0.5495	0.6269	0.6528	0.6558	0.6738	0.6905	0.6316	0.7174	0.6876	0.5834	0.5396	0.618	0.6593	0.6561	0.6691	0.6894	0.6209	0.6836	0.6657	0.5966

Table 13 Performance metrics for conventional participation without feature selection

No FS	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.5585	0.6245	0.6312	0.6366	0.6541	0.5747	0.5801	0.6635	0.6231	0.5882	0.5505	0.6137	0.6258	0.6285	0.6151	0.5532	0.576	0.6191	0.607	0.5666
Precision	0.4277	0.5018	0.5111	0.5203	0.5437	0.4595	0.4444	0.5625	0.5	0.4496	0.4266	0.4911	0.5025	0.5061	0.926	0.4492	0.4531	0.4961	0.4858	0.5563
Recall	0.5071	0.4929	0.4929	0.4571	0.5107	0.7286	0.4571	0.4821	0.425	0.4143	0.5607	0.6893	0.7214	0.5893	0.7143	0.8214	0.6036	0.6857	0.7357	0.6107
F-score	0.4541	0.4973	0.5018	0.4867	0.5267	0.5635	0.4507	0.5192	0.4595	0.4312	0.4846	0.5736	0.5924	0.5446	0.5831	0.5808	0.5176	0.5757	0.5852	0.5151
Specificity	0.6642	0.6966	0.6998	0.6942	0.7146	0.7458	0.6659	0.7117	0.6812	0.6619	0.672	0.7514	0.7713	0.7242	0.7626	0.7835	0.7	0.7528	0.768	0.6964
AUC	0.55	0.6581	0.6682	0.657	0.6867	0.6441	0.616	0.6861	0.6621	0.5939	0.5531	0.6525	0.6764	0.6648	0.6847	0.6507	0.6265	0.6844	0.6689	0.5975
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.5599	0.5989	0.6137	0.5868	0.6205	0.5585	0.5801	0.6191	0.5949	0.537	0.5505	0.6178	0.6218	0.6003	0.6285	0.5384	0.5868	0.6258	0.6043	0.5747
Precision	0.4259	0.4717	0.4903	0.467	0.4974	0.4518	0.4497	0.4961	0.4749	0.4024	0.4201	0.494	0.4986	0.4767	0.5054	0.4395	0.4579	0.5027	0.4822	0.4441
Recall	0.4821	0.5357	0.6321	0.6821	0.6857	0.8036	0.5107	0.6857	0.7107	0.4714	0.5071	0.5929	0.625	0.6214	0.6643	0.8179	0.525	0.6536	0.6786	0.5107
F-score	0.4523	0.5017	0.5523	0.5544	0.5766	0.5784	0.4783	0.5757	0.5694	0.4342	0.4595	0.539	0.5547	0.5395	0.5741	0.5718	0.4892	0.5683	0.5638	0.4751
Specificity	0.6596	0.6941	0.7304	0.7335	0.7535	0.7755	0.6776	0.7528	0.75	0.6434	0.6593	0.7199	0.7321	0.7196	0.7493	0.7703	0.6848	0.7441	0.7421	0.6746
AUC	0.5456	0.6417	0.663	0.6429	0.6876	0.6465	0.6276	0.6837	0.6574	0.5593	0.5433	0.6624	0.6622	0.6481	0.6824	0.6473	0.6301	0.6765	0.6591	0.5982

Table 14 Performance metrics for conventional participation with extra trees feature selection

Extra trees	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.5693	0.5976	0.5922	0.603	0.6353	0.5841	0.576	0.6272	0.6097	0.5572	0.5424	0.568	0.5639	0.5855	0.6003	0.576	0.5639	0.5949	0.5814	0.5451
Precision	0.4379	0.4649	0.4572	0.4735	0.5181	0.4498	0.4377	0.5069	0.4777	0.4089	0.4219	0.447	0.4439	0.4617	0.4784	0.4515	0.4399	0.4703	0.4621	0.4293
Recall	0.5036	0.45	0.4393	0.4786	0.4607	0.4643	0.4393	0.3929	0.3821	0.3929	0.5786	0.6179	0.6214	0.6036	0.6714	0.5821	0.575	0.5929	0.675	0.6286
F-score	0.4684	0.4574	0.4481	0.476	0.4877	0.4569	0.4385	0.4427	0.4246	0.4007	0.488	0.5187	0.5179	0.5232	0.5587	0.5086	0.4985	0.5245	0.5486	0.5101
Specificity	0.6698	0.6737	0.6688	0.6826	0.6943	0.6696	0.6602	0.6768	0.6667	0.6414	0.6713	0.6994	0.698	0.7056	0.7371	0.6937	0.6844	0.7077	0.7275	0.6877
AUC	0.5605	0.6149	0.625	0.6343	0.6528	0.6136	0.5899	0.6481	0.6283	0.5661	0.5485	0.6099	0.6189	0.6297	0.6463	0.6167	0.5952	0.6452	0.6247	0.5838
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.5612	0.576	0.5666	0.5626	0.6003	0.5693	0.5949	0.607	0.576	0.5384	0.5653	0.572	0.5653	0.5585	0.5747	0.5559	0.5518	0.6003	0.572	0.5518
Precision	0.4263	0.4448	0.4426	0.4503	0.4783	0.4459	0.4671	0.4831	0.4578	0.406	0.4313	0.4428	0.4404	0.4455	0.4545	0.4352	0.4209	0.477	0.452	0.4199
Recall	0.475	0.5036	0.5786	0.7286	0.6679	0.5893	0.5321	0.6143	0.6786	0.4857	0.4821	0.525	0.5679	0.7	0.6429	0.6	0.5036	0.6286	0.6393	0.464
F-score	0.4493	0.4724	0.5015	0.5566	0.5574	0.5077	0.4975	0.5409	0.5468	0.4423	0.4553	0.4804	0.4961	0.5444	0.5325	0.5045	0.4585	0.5424	0.5296	0.455
Specificity	0.6589	0.6737	0.687	0.7379	0.7358	0.6917	0.691	0.7209	0.7256	0.6471	0.6628	0.6764	0.6832	0.7228	0.7118	0.6863	0.6593	0.7219	0.7089	0.6578
AUC	0.5468	0.607	0.6136	0.6264	0.6576	0.614	0.5927	0.6453	0.6151	0.5619	0.5492	0.6027	0.6083	0.6218	0.634	0.6035	0.5777	0.6336	0.6154	0.5741

Table 15 Performance metrics for conventional participation with select 10 best feature selection

Select 10 best	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.5814	0.6205	0.6393	0.642	0.642	0.6043	0.6151	0.6581	0.638	0.5787	0.5518	0.5868	0.6218	0.6016	0.6137	0.5962	0.5935	0.6285	0.6043	0.541
Precision	0.4511	0.4962	0.5222	0.5267	0.5273	0.4798	0.4896	0.5551	0.5238	0.4363	0.429	0.4648	0.4987	0.4805	0.4911	0.4754	0.4696	0.5053	0.4828	0.4321
Recall	0.5107	0.4714	0.5036	0.4929	0.4821	0.5929	0.5036	0.4679	0.4321	0.4036	0.5714	0.6357	0.6964	0.7036	0.6929	0.6893	0.6071	0.685	0.7036	0.6929
F-score	0.4791	0.4835	0.5127	0.5092	0.5037	0.5304	0.4965	0.5078	0.4736	0.4193	0.49	0.537	0.5812	0.571	0.5748	0.5627	0.5296	0.5818	0.5727	0.5322
Specificity	0.6784	0.6897	0.7061	0.7048	0.7023	0.7128	0.6945	0.7061	0.6895	0.655	0.6757	0.7167	0.7585	0.7508	0.7529	0.7418	0.7113	0.7576	0.7522	0.7075
AUC	0.582	0.6436	0.6735	0.6689	0.6814	0.6369	0.6404	0.6807	0.6699	0.5892	0.5621	0.6308	0.6787	0.6797	0.6794	0.6423	0.6348	0.6801	0.6738	0.5943
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.5653	0.607	0.6083	0.5828	0.6285	0.5922	0.5908	0.6218	0.603	0.5626	0.5599	0.611	0.6285	0.5814	0.6178	0.5895	0.5976	0.6272	0.6124	0.5774
Precision	0.4304	0.4803	0.4851	0.4668	0.5051	0.4722	0.4645	0.4987	0.4817	0.4262	0.4277	0.486	0.5056	0.4652	0.4947	0.4693	0.472	0.504	0.4902	0.4452
Recall	0.475	0.5214	0.6393	0.7536	0.7036	0.6964	0.5607	0.6821	0.7036	0.4643	0.4964	0.5571	0.6464	0.7393	0.6643	0.6821	0.5714	0.6679	0.7179	0.4929
F-score	0.4516	0.5	0.5516	0.5765	0.5881	0.5628	0.5081	0.5762	0.5718	0.4444	0.4595	0.5191	0.5674	0.571	0.5671	0.556	0.517	0.5745	0.5826	0.4678
Specificity	0.6613	0.6948	0.7299	0.7629	0.7649	0.7424	0.6963	0.7528	0.7515	0.6575	0.6627	0.7062	0.7429	0.755	0.7439	0.7351	0.703	0.75	0.7628	0.6721
AUC	0.559	0.6348	0.6688	0.6587	0.6878	0.6383	0.6292	0.6796	0.6656	0.5912	0.5658	0.6367	0.6653	0.6555	0.6699	0.638	0.6292	0.675	0.6641	0.5872

Table 16 Performance metrics for non-conventional participation without feature selection

No FS	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.6151	0.7106	0.7093	0.7174	0.7147	0.6918	0.6797	0.7133	0.7039	0.6245	0.6097	0.7039	0.7106	0.712	0.7174	0.6918	0.6649	0.7079	0.7106	0.6285
Precision	0.6605	0.7049	0.7069	0.7071	0.7094	0.6838	0.7033	0.7171	0.6926	0.6585	0.649	0.726	0.7281	0.7288	0.7281	0.6868	0.7044	0.7335	0.726	0.6618
Recall	0.6152	0.8137	0.8039	0.8284	0.8137	0.8162	0.7206	0.7892	0.8584	0.6567	0.6299	0.7402	0.7549	0.7574	0.7745	0.8064	0.6716	0.7353	0.7598	0.6618
F-score	0.6371	0.7554	0.7523	0.763	0.758	0.7441	0.7119	0.7515	0.7545	0.6577	0.6393	0.733	0.7413	0.7428	0.7506	0.7418	0.6876	0.7344	0.7425	0.6618
Specificity	0.5675	0.7206	0.7133	0.7358	0.7236	0.707	0.6492	0.7075	0.7255	0.5833	0.5648	0.6758	0.6875	0.6897	0.7023	0.7008	0.6215	0.6766	0.6899	0.5881
AUC	0.6151	0.7564	0.7836	0.7766	0.7908	0.7502	0.7438	0.7783	0.7626	0.6821	0.6075	0.7556	0.7793	0.7755	0.7866	0.749	0.7384	0.7786	0.7612	0.6681
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.6245	0.712	0.7147	0.7012	0.7241	0.6945	0.681	0.7133	0.7026	0.6258	0.5976	0.7079	0.7227	0.7012	0.7079	0.6904	0.6662	0.7147	0.6972	0.6258
Precision	0.6667	0.7109	0.7197	0.7095	0.7241	0.6889	0.706	0.7327	0.7073	0.6633	0.638	0.7166	0.7285	0.7173	0.7175	0.6886	0.686	0.7379	0.7002	0.6641
Recall	0.6324	0.8015	0.7868	0.7721	0.8039	0.8088	0.7181	0.7525	0.7819	0.6471	0.6176	0.7745	0.7892	0.7525	0.7721	0.7966	0.723	0.7451	0.7843	0.6446
F-score	0.6491	0.7535	0.7518	0.7394	0.7619	0.7441	0.712	0.7424	0.7427	0.6551	0.6276	0.7444	0.7576	0.7344	0.7438	0.7386	0.7041	0.7415	0.7399	0.6542
Specificity	0.5787	0.7138	0.7071	0.689	0.7241	0.7045	0.6494	0.6883	0.6952	0.5826	0.5517	0.6954	0.7143	0.6794	0.6941	0.6937	0.639	0.6858	0.6923	0.5821
AUC	0.626	0.7551	0.7761	0.7714	0.7863	0.7498	0.7454	0.7775	0.7601	0.6756	0.5962	0.7578	0.7812	0.7759	0.7779	0.7484	0.7333	0.7767	0.7593	0.6808

Table 17 Performance metrics for non-conventional participation with extra trees feature selection

Extra trees	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.6029	0.6487	0.6783	0.6608	0.6891	0.6649	0.6272	0.6756	0.6689	0.6083	0.5895	0.6501	0.6729	0.6743	0.6689	0.6555	0.6245	0.6622	0.6581	0.6151
Precision	0.6438	0.6713	0.6849	0.6757	0.6895	0.6934	0.6563	0.6956	0.6746	0.6459	0.6419	0.6864	0.7027	0.6781	0.6947	0.6939	0.6711	0.7018	0.6915	0.6517
Recall	0.6201	0.7059	0.7672	0.7353	0.7892	0.6985	0.674	0.7279	0.7672	0.6348	0.5711	0.6544	0.701	0.7745	0.7083	0.6667	0.6201	0.6691	0.6814	0.6422
F-score	0.6317	0.6882	0.7237	0.7042	0.736	0.696	0.6651	0.7114	0.7179	0.6403	0.6044	0.6742	0.7018	0.7231	0.7015	0.68	0.6446	0.6851	0.6864	0.6469
Specificity	0.5571	0.6178	0.6678	0.6388	0.6884	0.6295	0.5895	0.6487	0.6595	0.5643	0.5395	0.6062	0.6369	0.6679	0.6361	0.6125	0.5765	0.6186	0.6188	0.5718
AUC	0.6027	0.705	0.7381	0.7282	0.741	0.72	0.7039	0.7363	0.7184	0.6656	0.5931	0.7058	0.7378	0.7378	0.7423	0.7198	0.6943	0.7371	0.7169	0.6534
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.611	0.646	0.6676	0.6487	0.6581	0.6541	0.6151	0.6555	0.6689	0.6043	0.5989	0.6581	0.6824	0.6649	0.6783	0.6608	0.6487	0.6662	0.6649	0.6312
Precision	0.6522	0.6739	0.6868	0.6889	0.6791	0.6921	0.6495	0.6959	0.6929	0.6508	0.6478	0.6878	0.7019	0.7076	0.6979	0.698	0.678	0.7062	0.6925	0.6727
Recall	0.625	0.6887	0.7255	0.6569	0.7157	0.6667	0.6495	0.6618	0.7132	0.6029	0.5907	0.6912	0.7328	0.6642	0.7304	0.674	0.6863	0.6716	0.701	0.6397
F-score	0.6383	0.6812	0.7056	0.6725	0.6969	0.6792	0.6495	0.6784	0.7029	0.626	0.6179	0.6895	0.717	0.6852	0.7138	0.6858	0.6821	0.6884	0.6967	0.6558
Specificity	0.5653	0.6104	0.641	0.6045	0.6294	0.6114	0.5731	0.6113	0.6378	0.5562	0.5499	0.6216	0.6562	0.6194	0.6519	0.6189	0.6121	0.6225	0.6303	0.5859
AUC	0.6103	0.7012	0.7356	0.7221	0.7408	0.7198	0.6933	0.7357	0.7126	0.6668	0.6018	0.6999	0.7351	0.7371	0.7434	0.7217	0.7031	0.7366	0.7134	0.6776

Table 18 Performance metrics for non-conventional participation with select 10 best feature selection

Select 10 best	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.5962	0.6568	0.6904	0.6703	0.6877	0.6676	0.6501	0.6918	0.6851	0.6501	0.5882	0.6595	0.6756	0.6958	0.6824	0.6676	0.6474	0.6904	0.6877	0.638
Precision	0.6274	0.6735	0.6943	0.6936	0.6864	0.6785	0.6762	0.6993	0.679	0.6805	0.6275	0.6914	0.7022	0.6944	0.7009	0.6834	0.6881	0.716	0.7075	0.6488
Recall	0.652	0.7279	0.7794	0.7157	0.7941	0.75	0.6961	0.7696	0.8088	0.6838	0.6152	0.6863	0.7108	0.7966	0.7353	07353	0.6544	0.723	0.7353	0.7426
F-score	0.6394	0.6996	0.7344	0.7045	0.7364	0.7125	0.686	0.7328	0.7383	0.6822	0.6213	0.6888	0.7065	0.742	0.7177	0.7084	0.6709	0.7195	0.7212	0.6926
Specificity	0.5549	0.6325	0.6842	0.6398	0.69	0.6507	0.6161	0.6803	0.6965	0.6126	0.5423	0.6213	0.6424	0.6982	0.6571	0.6447	0.6028	0.6586	0.6614	0.6196
AUC	0.5865	0.6887	0.7443	0.744	0.7504	0.7285	0.7032	0.7605	0.7525	0.6822	0.5815	0.6998	0.7465	0.7577	0.7522	0.7274	0.7087	0.7609	0.7541	0.6634
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.5895	0.6595	0.681	0.6689	0.681	0.6676	0.6514	0.6931	0.6904	0.6339	0.6097	0.6649	0.6756	0.6756	0.6756	0.6622	0.6514	0.6931	0.6837	0.6608
Precision	0.6241	0.6773	0.6921	0.7005	0.693	0.6842	0.684	0.7174	0.7023	0.6726	0.6453	0.6888	0.6885	0.7072	0.6868	0.683	0.6867	0.7195	0.6961	0.699
Recall	0.6348	0.7255	0.7549	0.6936	0.7525	0.7328	0.6789	0.7279	0.7574	0.6495	0.6422	0.7108	0.7475	0.6985	0.7525	0.7181	0.6716	0.723	0.7525	0.6716
F-score	0.6924	0.7006	0.7222	0.697	0.7215	0.7077	0.6814	0.7226	0.7288	0.6608	0.6437	0.6996	0.7168	0.7028	0.7181	0.7001	0.6791	0.7213	0.7232	0.685
Specificity	0.5457	0.634	0.6644	0.6313	0.6633	0.6438	0.6124	0.6626	0.6733	0.5903	0.5667	0.6335	0.6567	0.6382	0.6588	0.6338	0.6105	0.6607	0.6656	0.6182
AUC	0.5812	0.6922	0.7419	0.7468	0.7443	0.7274	0.7005	0.7588	0.7491	0.6785	0.6004	0.6966	0.7408	0.7468	0.741	0.728	0.7092	0.7587	0.7485	0.6899

Table 19 Performance metrics for participation without feature selection

No FS	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.8345	0.9031	0.9071	0.8991	0.9098	0.8816	0.8735	0.9098	0.9071	0.8856	0.6258	0.6945	0.6985	0.6783	0.6743	0.7322	0.6339	0.7133	0.6595	0.5855
Precision	0.9219	0.9157	0.9196	0.9177	0.9186	0.9309	0.9179	0.9107	0.9071	0.913	0.9342	0.9479	0.9555	0.9541	0.9576	0.9473	0.9487	0.9657	0.9488	0.9357
Recall	0.8932	0.9837	0.9837	0.9763	0.9881	0.9392	0.9451	0.9985	1	0.966	0.632	0.7018	0.7003	0.678	0.6706	0.7463	0.6306	0.7092	0.6602	0.5831
F-score	0.9073	0.9485	0.9505	0.9461	0.9521	0.935	0.9313	0.9526	0.9513	0.9387	0.754	0.8065	0.8082	0.7927	0.7888	0.8349	0.7576	0.8178	0.7787	0.7185
Specificity	0.2	0.4211	0.5	0.3846	0.5556	0.3492	0.2449	0.75	nan	0.2333	0.1359	0.1762	0.1888	0.178	0.1808	0.1934	0.1559	0.2097	0.1642	0.13
AUC	0.5759	0.7109	0.7437	0.7391	0.7566	0.7356	0.702	0.7817	0.6488	0.6111	0.5986	0.7148	0.7536	0.7517	0.7387	0.7276	0.6956	0.7658	0.731	0.6413
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.8493	0.8856	0.8439	0.7564	0.7497	0.7604	0.8493	0.7079	0.716	0.7376	0.7645	0.8156	0.8075	0.786	0.7604	0.6985	0.8129	0.716	0.7376	0.7362
Precision	0.9132	0.9189	0.9332	0.9441	0.9485	0.9444	0.9271	0.9579	0.9548	0.9269	0.9165	0.9215	0.9303	0.927	0.9336	0.9447	0.9294	0.9327	0.9331	0.9378
Recall	0.9214	0.9585	0.8917	0.7774	0.7656	0.7819	0.905	0.7092	0.7211	0.7715	0.8145	0.8709	0.8516	0.8294	0.7923	0.7092	0.8591	0.7404	0.7656	0.7596
F-score	0.9173	0.9383	0.912	0.8527	0.8473	0.8555	0.9159	0.815	0.8216	0.8421	0.8625	0.8955	0.8892	0.8755	0.8571	0.8102	0.8928	0.8255	0.8411	0.8393
Specificity	0.1587	0.3	0.2626	0.2021	0.206	0.2054	0.2471	0.1967	0.1966	0.1538	0.1319	0.1792	0.2063	0.6863	0.1813	0.173	0.2083	0.1587	0.1684	0.1777
AUC	0.5331	0.6978	0.7191	0.717	0.747	0.7359	0.7025	0.7738	0.7317	0.6213	0.5437	0.6714	0.7032	0.6863	0.7052	0.7166	0.6698	0.6949	0.6822	0.6551

Table 20 Performance metrics for participation with extra trees feature selection

Extra trees	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.8034	0.8991	0.9004	0.8937	0.9031	0.8748	0.8802	0.9071	0.9071	0.8843	0.607	0.6393	0.6353	0.6581	0.638	0.6003	0.5882	0.6541	0.638	0.5801
Precision	0.9125	0.9142	0.9132	0.9081	0.9112	0.9192	0.9173	0.9071	0.9071	0.9095	0.9226	0.9394	0.9508	0.9487	0.9551	0.9415	0.9423	0.9503	0.9451	0.9269
Recall	0.8665	0.9807	0.9837	0.9822	0.9896	0.9451	0.954	1	1	0.9688	0.6187	0.6439	0.6306	0.6588	0.6306	0.5964	0.5816	0.6528	0.638	0.5831
F-score	0.8889	0.9463	0.9471	0.9437	0.9488	0.932	0.9353	0.9513	0.9513	0.9382	0.7407	0.7641	0.7583	0.7776	0.7596	0.7302	0.7193	0.774	0.7617	0.7158
Specificity	0.1262	0.35	0.3529	0.1429	0.3636	0.26	0.2619	nan	nan	0.16	0.1168	0.1459	0.1588	0.1636	0.1644	0.1392	0.1376	0.1643	0.1528	0.1191
AUC	0.5388	0.6485	0.6979	0.6978	0.6936	0.6885	0.6473	0.7204	0.5636	0.5897	0.5557	0.6825	0.708	0.7024	0.7104	0.6785	0.6505	0.7058	0.7014	0.615
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.821	0.8721	0.8197	0.712	0.7066	0.6353	0.8304	0.6635	0.6783	0.7443	0.7645	0.8237	0.8008	0.7631	0.51	0.6729	0.8022	0.7052	0.7376	0.7349
Precision	0.9105	0.9165	0.9313	0.9291	0.9419	0.9549	0.9215	0.953	0.9485	0.9261	0.925	0.9196	0.9311	0.922	0.9327	0.9354	0.9271	0.935	0.9347	0.9313
Recall	0.8902	0.9451	0.865	0.7389	0.7211	0.6276	0.8887	0.6617	0.6825	0.7804	0.8056	0.8828	0.8427	0.8071	0.7819	0.6869	0.8487	0.7255	0.7641	0.7641
F-score	0.9002	0.9306	0.8969	0.8231	0.8168	0.7574	0.9048	0.7811	0.7938	0.847	0.8612	0.9008	0.8847	0.8608	0.8507	0.7921	0.8861	0.817	0.8408	0.8394
Specificity	0.119	0.2292	0.2222	0.1498	0.1718	0.1633	0.1935	0.1709	0.1705	0.1543	0.1603	0.1771	0.203	0.1503	0.1742	0.1492	0.1905	0.1591	0.1719	0.1632
AUC	0.5182	0.6449	0.671	0.6418	0.6912	0.6919	0.6476	0.7148	0.6933	0.6022	0.5902	0.6687	0.6712	0.646	0.6715	0.6721	0.665	0.6775	0.662	0.6294

Table 21 Performance metrics for participation with select 10 best feature selection

Select best	Normal										Undersampling									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.856	0.8923	0.9004	0.9071	0.9098	0.8735	0.8869	0.9112	0.9071	0.8789	0.5841	0.6487	0.6716	0.7026	0.6662	0.7402	0.6218	0.7093	0.6824	0.6999
Precision	0.9115	0.908	0.9144	0.9093	0.9163	0.9277	0.9143	0.9119	0.9071	0.9067	0.9176	0.9329	0.9442	0.9557	0.9494	0.9413	0.9262	0.9617	0.9488	0.9279
Recall	0.9318	0.9807	0.9822	0.997	0.9911	0.9332	0.9659	0.9985	1	0.9659	0.595	0.6602	0.678	0.7047	0.6677	0.7611	0.6335	0.7077	0.6869	0.7255
F-score	0.9215	0.9429	0.9471	0.9512	0.9522	0.9305	0.9394	0.9533	0.9513	0.9353	0.7219	0.7732	0.7893	0.8113	0.784	0.8417	0.7524	0.8154	0.7969	0.8143
Specificity	0.1481	0.1333	0.3684	0.5	0.5714	0.3077	0.2581	0.8	nan	0.08	0.1078	0.1391	0.1622	0.1911	0.1673	0.1869	0.1241	0.2024	0.1725	0.1435
AUC	0.6094	0.6733	0.6999	0.7377	0.725	0.728	0.6704	0.769	0.6464	0.593	0.5385	0.6567	0.7029	0.7375	0.7146	0.723	0.6277	0.7572	0.7278	0.6469
	Oversampling										SMOTE									
	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN	DT	RF	CB	NN	GB	NB	XGB	LR	SVM	KNN
Accuracy	0.7766	0.782	0.7497	0.7106	0.7201	0.7577	0.7604	0.6985	0.6985	0.7995	0.7295	0.751	0.7429	0.7201	0.6958	0.6985	0.7483	0.6985	0.6689	0.7927
Precision	0.9247	0.9252	0.9296	0.9456	0.9533	0.9379	0.9261	0.9611	0.9592	0.916	0.9231	0.9179	0.9229	0.9413	0.9392	0.9482	0.922	0.95	0.944	0.914
Recall	0.8205	0.8264	0.7834	0.7226	0.727	0.7849	0.7997	0.6958	0.6973	0.8576	0.7656	0.7967	0.7819	0.7374	0.7107	0.7062	0.7893	0.7047	0.6751	0.8516
F-score	0.8695	0.873	0.8502	0.8192	0.8249	0.8546	0.8583	0.8072	0.8076	0.8858	0.837	0.8531	0.8466	0.827	0.8091	0.8095	0.8505	0.8092	0.7872	0.8817
Specificity	0.1655	0.1702	0.1657	0.1798	0.1965	0.1899	0.1615	0.1961	0.1937	0.1429	0.1413	0.1329	0.1453	0.1767	0.1631	0.1784	0.1446	0.1811	0.1609	0.1304
AUC	0.5935	0.6577	0.6567	0.6891	0.7181	0.7261	0.6649	0.7625	0.72	0.5819	0.5672	0.6275	0.6437	0.6768	0.6789	0.7224	0.6176	0.7198	0.6828	0.5969

As observed in the previous section, the performance metrics were similar for all the models. So, to test if there were statistically significant differences between the models, the Friedman test was applied to all the metrics except for the F-score, as it is a combination of precision and recall, so that it would be redundant.

The conclusion of the Friedman test was the same for vote, conventional participation, non-conventional participation, and participation. There were significant differences among the models at a 5% significance level. The results for vote were X^2(9, n=10) = 104.7, p < 0.001. In the case of conventional participation, the results were X^2(9, n=10) = 333.8, p < 0.001. For non-conventional participation, the results were X^2(9, n=10) = 387.6, p < 0.001. And lastly, for participation, the results were X^2(9, n=10) = 108.7, p< 0.001. Subsequently, the Nemenyi test was applied to validate these differences, alongside constructing a critical difference diagram (CDD) to illustrate it.

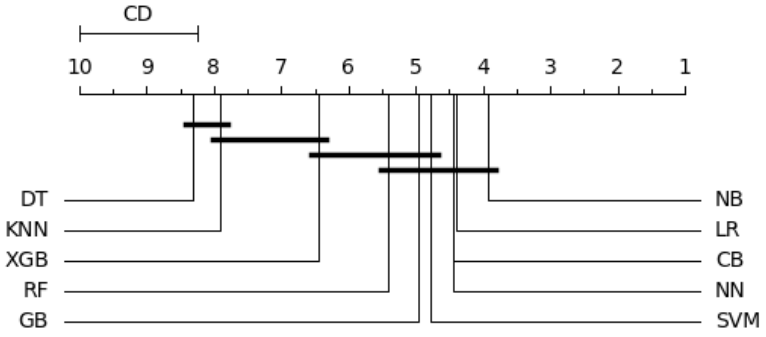

Figure 1 CDD for vote models

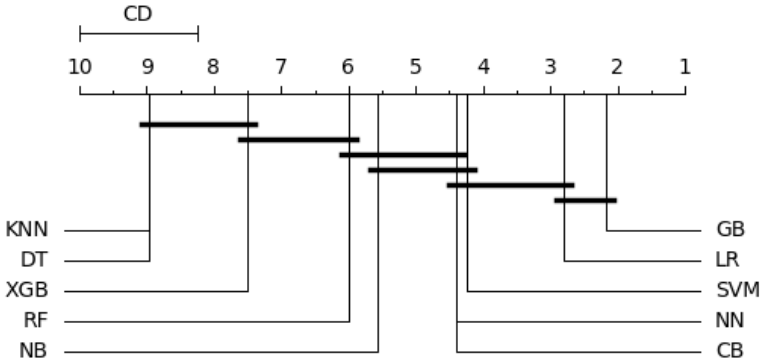

Figure 2 CDD for conventional participation models

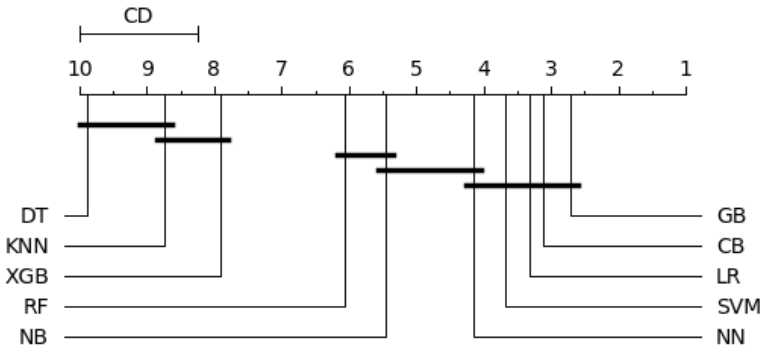

Figure 3 CDD for non-conventional participation models

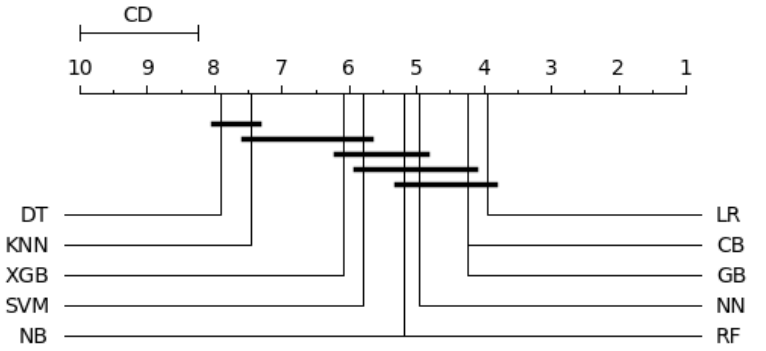

Figure 4 CDD for participation models

The calculated critical difference was 1.749. For vote, the modes with the superior performance were Naïve Bayes, Logistic Regression, Cat Boost, Neural Network, Support Vector Machine, Gradient Boosting, and Random Forest, of which Naïve Bayes appeared to be the most effective. Considering conventional participation, Gradient Boosting performed best, with no statistically significant differences from Logistic Regression. For non-conventional participation, Gradient Boosting, Cat Boost, Logistic Regression, Support Vector Machines, and Neural Networks were the ones with the highest performance, the best one was Gradient Boosting. Relatively to participation, Logistic Regression, Cat Boost, Gradient Boosting, Neural Networks, Random Forest, and Naïve Bayes were the ones with the highest performance, with Logistic Regression being the best. The best model for each dependent variable will be analysed in the following phase.

4.3 Models' Interpretation

One method for model interpretation involves computing the SHAP (SHapley Additive exPlanation) values. It measures each variable's importance in the model, allowing the

identification of those that most influence the target behaviours. The variables are presented according to their importance. The first variable on the plot is the one with more impact on the outcome.

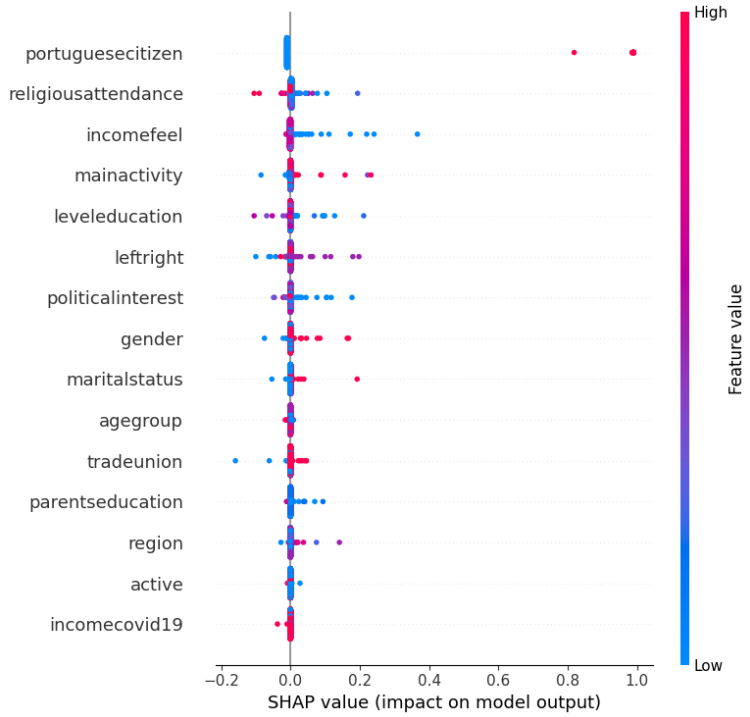

Figure 5 SHAP values for vote

The variable "portuguesecitizen" is the one that has the highest contribution to vote outcome. Specifically, individuals who are not Portuguese citizens tend to have a higher chance of abstaining from voting. The second feature with more impact is "religiousattendance", which positively relates to vote. The "religiousattendance" variable is coded from 1 to 7, where 1 indicates that the individual never attends religious events, and 7 that attends every day. In the figure, the further to the left the dots are, the individuals have more chances to vote, and the opposite to the right. In this case, the red dots, which signify the highest feature values, are on the left. In "religiousattendance" variable, the higher values mean more frequent attendance at religious events. Regarding the variable vote, the individuals who frequently attend religious events are likelier to vote.

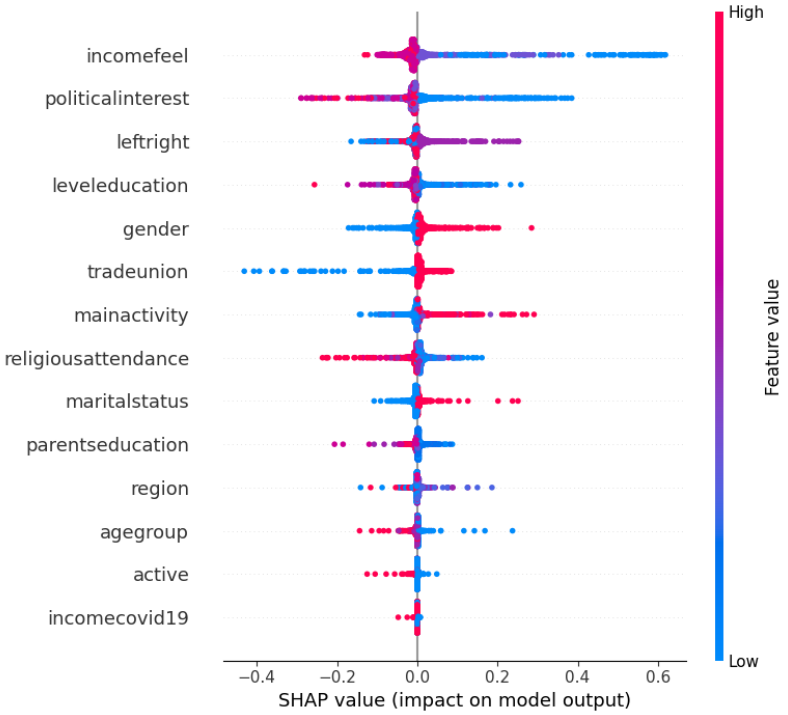

Figure 6 SHAP values for vote without "portuguesecitizen" variable

Since the variable "portuguesecitizen" has a strong influence on the outcome, 98.3% of the participants are Portuguese, it seemed important to analyse the results without this variable. The previous figure presents the shap values for vote without "portuguesecitizen" variable. The variable that has more impact on voting is "incomefeel" someone with a higher income is more likely to vote than someone with a lower income. Also, if a person has an interest in politics, the chance of voting is higher. Individuals who position themselves in the centre of the political spectrum have a higher possibility of not voting. Relatively to gender, there is an apparent effect: men are more likely to vote than women.

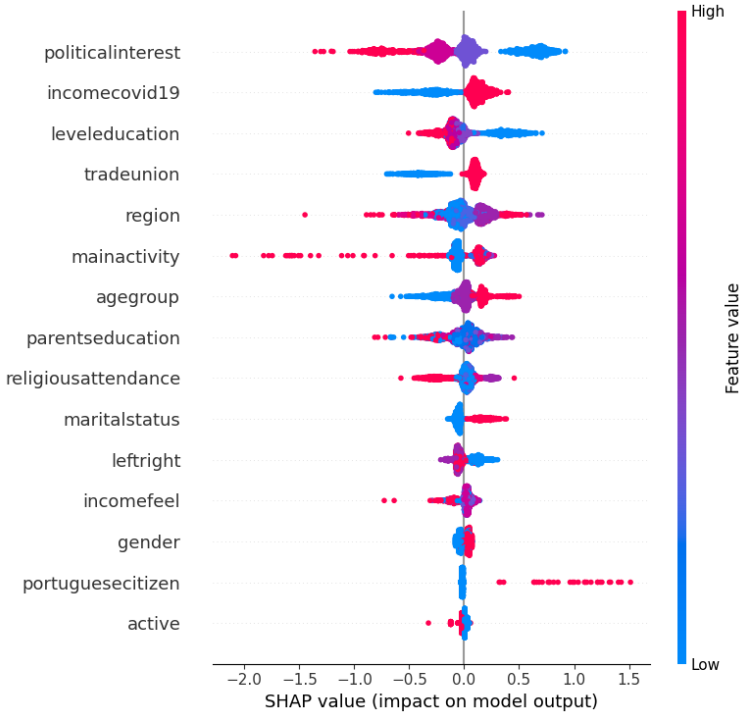

Figure 7 SHAP values for conventional participation

Relatively to conventional participation, the profile of an individual who conventionally participates in politics is a Portuguese citizen who is interested in politics and is more likely to have higher levels of education. Additionally, affiliation with a trade union increases the chance of participation. In terms of age, is someone young, whose income has reduced with the covid-19 pandemic.

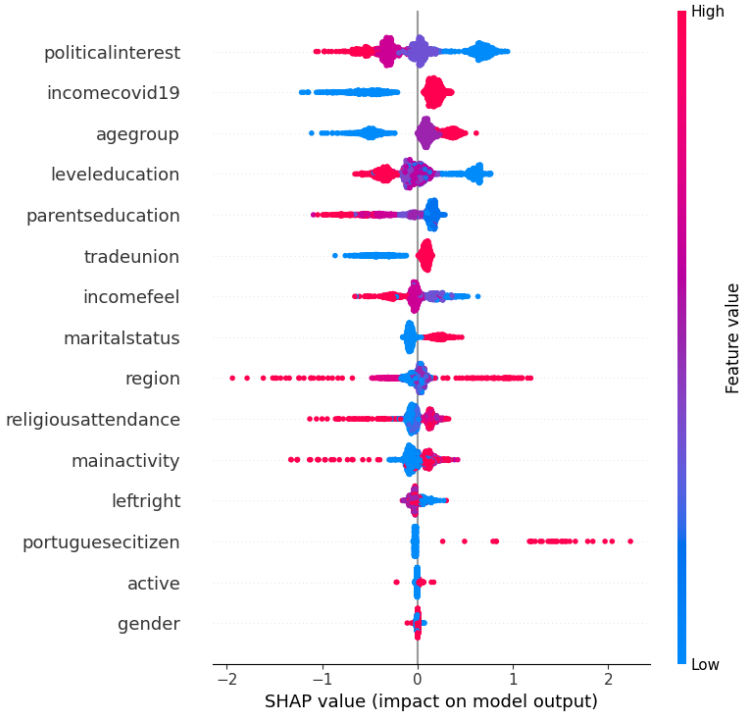

Figure 8 SHAP values for non-conventional participation

The lack of engagement in non-conventional political activities in Portugal is linked to factors such as non-Portuguese citizenship and disinterest in politics. Older individuals with lower levels of education also demonstrate a propensity for not participating in a non-conventional way.

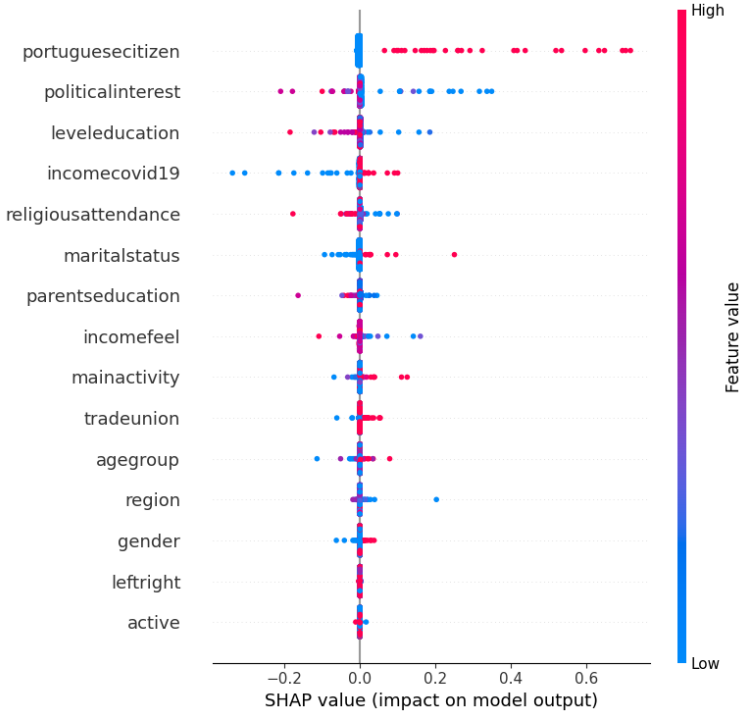

Figure 9 SHAP values for participation

As participation is the junction of vote, conventional participation, and non-conventional participation, SHAP values for participation will also be a conjugation of SHAP values of the three forms of participation. Portuguese citizenship is the variable with more impact, individuals who are not Portuguese citizens are more likely not to participate, and their interest in politics influences positively their participation in politics. Also, the level of education has a high impact, people with higher levels of education have a higher chance of participation.

5. Conclusions

This dissertation aimed to study the determinants of political participation in Portugal by applying machine learning algorithms. Political participation was divided into three distinct forms: vote, conventional participation, and non-conventional participation. Participation was also a variable, it was the aggregation of vote, conventional participation, and non-conventional participation.

The data was from the European Social Survey conducted in 2020 and the "The Political Participation of Youth in Portugal, 2020" survey, funded by *Fundação Calouste Gulbenkian*. There was a need to uniformise the datasets to merge them. The first step was to define which variables were common to both and then do some transformations so they would be coded as the same. Then, having just one dataset, it was necessary to deal with missing values, and since they were not missing completely at random, all the missing values, except those from the variable "leftright", were deleted.

Ten machine learning algorithms were applied: Decision Tree, Random Forest, Cat Boost, Neural Network, Gradient Boosting, Naïve Bayes, XGBoost, Logistic Regression, Support Vector Machine and Nearest Neighbor. These models were built to make predictions for the four target variables. Then, the models were compared by performance metrics: accuracy, precision, recall, F-score, sensitivity, and area under the curve. Also, non-parametric tests were conducted to identify the best model for each case.

Regarding vote, the model that performed better was Naïve Bayes. The variables that had more impact on the outcome were: "portuguesecitizen", "religiousattendance", and "incomefeel". Individuals who are not Portuguese citizens, with low attendance to religious events, and who feel that their current income is not allowing them to live comfortably are more likely not to vote.

Relatively to conventional participation, the algorithm chosen was Gradient Boosting. "politicalinterest", "incomecovid19", and "leveleducation" were the three variables with more influence in predicting the outcome. Someone interested in politics, whose income was reduced with covid-19 and who has the highest levels of education, is more likely to conventionally participate in politics than someone who is not interested in politics, whose income has increased during the pandemic and with a lower level of education.

Considering non-conventional participation, the algorithm chosen was also Gradient Boosting. The three variables with more impact were "politicalinterest", "incomecovid19", and "agegroup". This means that an individual with political interest, whose income was reduced due to the pandemic and who is younger, has more probability of participating non-conventionally than someone without interest in politics, whose income has increased with covid-19 and older.

Lastly, to participation, the algorithm chosen was Logistic Regression. "portuguesecitizen", "politicalinterest," and "leveleducation" were the three variables that most contributed to the outcome. The main differences between who participates in politics and who doesn't are that while who participates is probably a Portuguese citizen with political interest and with a higher level of education, the second one is a non-Portuguese citizen without interest in politics and with a lower level of education.

This study has some limitations, the volume of data was not high, the sample after data treatment only had 2475 answers. Also, the data was unbalanced, especially for vote, only 16.7% answered that they didn't vote in the national election of 2019. Consequently, participation was also unbalanced, the proportion of respondents who never participated in politics was 10.2%. This was expected since the abstention rates in surveys are always low. Some variables were also unbalanced, for example, "agegroup", almost 50% of the respondents was aged between 35 and 64.

Despite the limitations, this study helped develop the knowledge acquired during the master's. Also, this is the first study considering the determinants of political participation in Portugal using Machine Learning.

References

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series, 1142*.

Bass, L., & Casper, L. (2001). Impacting the Political Landscape: Who Registers and Votes among Naturalized Americans? *Political Behavior, 23*, 103-130.

Chaplot, N., Pandey, D., Kumar, Y., & Sisodia, P. S. (2023). A Comprehensive Analysis of Artificial Intelligence Techniques for the Prediction and Prognosis of Genetic Disorders Using Various Gene Disorders. *Archives of Computational Methods in Engineering, 30*, 3301-3323.

Chong Hua, K., Mohd Jamil, J., & Mohd Shaharanee, I. (2021). A Decision Tree Classifier for Predicting Voter Turnout in Malaysian General Election. *Central Asia and the Caucasus, 22*, 816-830.

Costa, P. (2022). *Portugal em 2020.* Fundação Calouste Gulbenkian.

Costa, P., Nogueira, A. R., & Gama, J. (2021). Modelling Voting Behaviour During a Genereal Election Campaign Using Dynamic Bayesian Networks. In G. Marreiros, F. Melo, N. Lau, H. Lopes Cardoso, & L. Reis, *Progress in Artificial Intelligence* (pp. 524-536). Springer.

Da Silva, F. F., & Costa, P. (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*, 117-140.

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies, 7*, 1174-1179.

Ekman, J., & Amnå, E. (2012). Political participation and civic engagement: Towards a new typology. *Human Affairs, 22*(3), 283-300.

Fawcett, T. (2006). *An introduction to ROC analysis, Pattern Recognition Letters* (Vol. 27).

Herbold, S. (2020). Autorank: A Python package for automated ranking of classifiers. *Journal of Open Source Software, 5*(48).

Kim, S.-y. S., Alvarez, R. M., & Ramirez, C. M. (2020). Who Voted in 2016? Using Fuzzy Forests to Understand Voter Turnout. *Social Science Quarterly, 101*, 978-988.

Magalhães, P. (2022). *Um retrato comparativo e longitudinal, 2002-2019.* Fundação

Caloust Gulbenkian.

Reichert, F. (2016). How internal political efficacy translates political knowledge into political participation: Evidence from Germany. *Europe's Journal of Psychology, 12*(2), 221-241.

Smets, K., & Van Ham, C. (2013). The embarrassment of Riches? A Meta-Analysis of Individual-Level Research on Voter Turnout. *Electoral Studies, 32*, 344-359.

Teorell, J., Torcal, M., & Montero, J. (2007). Pol. Em J. W. Van Deth, *Citizenship and Involvement in European Democracies: A Comparative Analysis* (pp. 334-357). Milton Park: Routledge.

Theocharis, Y., & Van Deth, J. (2018). The continuous expansion of citizen participation: a new taxonomy. *Political Science Review, 10*, 139-163.

Van Deth, J. (2016). What is political participation. Em *The international encyclopedia of political communication* (pp. 349-367).

Van Deth, J. W. (2001). *Studying Political Participation: Towards a Theory of Everything?*

Welcome to the shap documentation. (25 de July de 2023). Obtido de Welcome to the SHAP documentation - SHAP latest documentation: https://shap.readthedocs.io/en/latest/

Appendix

1. Variables common to both data sets

European Social Survey	The Political Participation of Youth in Portugal
gndr	A1
agea	A2.1
region	NUTSII
edlvdpt	A5
polintr	Q9
mbtru	Q11_2
lrscale	Q17
mnactic	Q28
emplrel	Q29
wrkctra	Q30
hincfel	Q31
hapirc19	Q32
ctzcntr	Q34
maritalb	Q36
rlgatnd	Q40
vote	Q19

2. Variables that compose conventional participation

European Social Survey	The Political Participation of Youth in Portugal
contplt	Q11_1
donprty	Q14_3
badge	Q14_4
	Q14_5
	Q14_10
	Q14_14

3. Variables that compose non-conventional participation

European Social Survey	The Political Participation of Youth in Portugal
sgnptit	Q14_1
pbldmna	Q14_2
bctprd	Q14_6
pstplonl	Q14_7
volunfp	Q14_8
	Q14_9
	Q14_11
	Q14_12

4. Little's MCAR test results

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1-active																			
2-agegroup	*																		
3-conventionalpar-ticipation	*	0.0001																	
4-gender	0.1736	0.11	0.1973																
5-incomecovid19	*	*	*	0.8797															
6-incomefeel	*	*	*	*	*														
7-leftright	*	*	*	*	*	*													
8-leveleducation	*	*	*	0.0008	*	*	*												
9-mainactivity	*	*	*	*	*	*	*	*											
10-maritalstatus	*	*	0.9649	0.0394	*	*	*	*	*										
11-nonconventional-participation	*	*	*	0.0055	*	*	*	*	*	0.0116									
12- parentseducation	*	*	*	0.3373	*	*	*	*	*	*	*								
13-participation	*	*	*	0.0205	*	*	*	*	*	*	*	*							
14- politicalinterest	*	*	*	*	0.3549	*	*	*	*	0.2743	*	*	*						
15- portuguesecit-izen	*	*	*	0.4208	0.0003	0.1153	0.0065	0.0418	*	*	*	*	*	0.2537					
16-region	0.0001	0.0019	0.0188	0.1289	*	*	*	*	*	*	*	*	*	0.0239	*				
17-religiousattend-ance	*	*	0.282	*	*	*	*	*	*	*	0.0001	*	*	0.1558	0.0261	*			
18-tradeunion	*	*	*	0.0027	0.844	0.0007	*	0.015	*	*	*	0.0007	0.0003	*	0.6343	0.0174	0.9704		
19-vote	*	*	*	0.0323	*	*	*	*	*	*	*	*	*	*	*	0.0063	*	0.001	

* $p<0.0001$