Universitat Oberta de Catalunya (UOC)

Master's degree in Data Science

MASTER'S FINAL PROJECT

Area: Human Behaviour

Predicting human behaviour in PGGs using a game theoretical approach

---

Author: Ignasi Vilarasau Antolín

Tutor: Julià Vicens Bennasar

Professor: Albert Solé

Barcelona, September 18, 2020

FINAL THESIS SHEET

Title of the Thesis:	Predicting human behaviour in PGGs
Name of the author:	Ignasi Vilarasau Antolín
Name of the teaching collaborator:	Julià Vicens Bennasar
Name of the PRA:	Albert Solé
Delivering date (mm/aaaa):	06/2020
Titulation or program:	Data Science
Area of work of the Final Thesis:	Human Behaviour
Language:	English
Key-words	Public Goods Games, Machine Learning, Explainability

Declaration

The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.

Isaac Asimov (January 2, 1920 - April 6, 1992)

Aknowledgements

I would like to thank my advisor Julià Vicens Bennasar for his constant support, accurate guidance and confidence. I would also like to thank my parents for always being there and encouraging me. And at last but not least, give a special thank to my friends that supported me through all this work.

Abstract

Social interactions are present on every daily situation, and social situations that involve strategic behaviour are the keys of every social interaction. Those situations can be extrapolated to *Evolutionary Game Theory* through *Public Goods Games*, (2). On this sense, studying the behaviour of individuals from the theoretical and experimental point of view can lead to an understanding how individuals would react and act to those strategic situations and it could be very important in order to be able to anticipate the outcomes of every kind of conflict or social situation (4).

In order to be able to understand that behaviour we will try to identify discrete behavioural types (clusters) of individuals in experimental data and try to classify each individual behaviour to one of the types of discrete behaviours identified. In addition, we will also apply more machine learning supervised models in order to be able to classify the individuals with the maximum accuracy.

Speaking of machine learning models allows us to also speak about *eXplainable Artificial Intelligence (XAI)*. *XAI* creates a suite of machine learning techniques that enables humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners (5). So, in addition to applying all the machine learning models that we previously talked about, we will also introduce the main aspects of *XAI* techniques/methods and apply them to our ML models. Thus, we will finally apply a set of algorithms, based on game theory, that correspond to the contribution of each feature of the model towards pushing the prediction away from the expected value.

Key-words: Public Goods Games, Machine Learning, Explainability.

Contents

List of Figures

Chapter 1

Introduction

1.1 Context and justification of the interest of the proposal

The interest of our proposal is based on two things. The first one, is the creation of supervised ML models that, with the maximum accuracy, classifies individuals into discrete behavioural types and it could be applied into simulating societies, policy-making scenario building, and even a variety of business applications. The second one, is based on the applicability of the eXplainable Artificial Intelligence (XAI) methods in the post-hoc analysis in order to understand our ML models' results with the maximum information about the features making those results as much human comprehensible as possible, (1). And another interesting point is that to apply this XAI techniques we will use a solution concept of fairly distributing both gains and costs to several actors working in coalition, based on game theory. This new sort of techniques are in the trend because humans do not want to accept the models "as-is" we want to know what is happening in their core (20).

1.2 Main objectives

The main goal of this work is to cluster the participants identifying discrete behavioural types. Then, the goal is to create different ML supervised models to predict classify the individuals on those discrete behavioural types depending on their behaviour during the PGGs studying the accuracy and effectiveness of all models. The secondary goal is to apply to those supervised ML models XAI techniques (focusing on SHapley Additive exPlanations (SHAP)) to explain and better understand the results and see if it can help us getting to much more interesting results and conclusions.

1.3 Methodology

In order to be able to accomplish the goals that we previously set, we will count on the experimental data from previous public studies of *Public Goods Games*. We will also work with the interpreted, high-level, general-purpose programming language, *Python* in order to create the ML models. We will firstly create an *Exploratory Data Analysis* to extract some first conclusions about the experimental data. Later, we will only focus with the most relevant ones and we will try to cluster the participants of each PGG into different types of participants related to their behaviour during each game. We will then apply supervised machine learning models to classify all the participants into the clusters previously found and apply the explainability and interpretability methods to better understand the results of our models.

1.4 Planning followed

The planning followed can be summed up into 6 different steps that we went through during this work. These steps are the following:

1. Select the experimental data from *PGGs* to train and test the ML models.
2. Create the first clustering and classification models (First two models, unsupervised and supervised)
3. Create all the remaining ML models (supervised)
4. Evaluate the accuracy of all the ML models
5. Implement the explainable and interpretable methods *Model-Agnostic Methods* ((3)) to the ML models previously created
6. Evaluate the suitability of all the XAI methods and extract conclusions and possible next steps

1.5 Summary of the final results

We have been able to obtain very interesting results from a machine learning perspective applying a first clustering of the PGGs' participants, with the later classification and final calculation of the shapley values and feature importance related to the initial conditions of the different Public Goods Games and the features of each game analyzed.

We have also obtained from our previous machine learning analysis the main features and initial conditions that positive or negative influence individual and group contributions in PGGs that very much in coincidence with many work already done in the field of PGGs.

1.6 Summary of the rest of the Chapters

Firstly, we will introduce the first concepts related to this work, which are *Game Theory*, *Public Goods Games*, *eXplainable Artificial Intelligence (XAI)* and *Shapley Values* (see chapters 2 and 3).

Secondly, we will start with a *literature review* of some papers that are related to *Public Goods Games* and then, if we find the experimental data and results interesting for our future work we will start with an *exploratory data analysis* to do a first overview of the experimental data (see chapter 4).

Then, we will apply unsupervised and supervised machine learning models to cluster and classify participants from experimental data from two of the PGGs analyzed before and we will apply model-agnostic methods to better understand the impact in the output value of the classifications of each feature and initial conditions of each of both PGGs analyzed (see chapter 5).

Finally, we will extract the pertinent conclusions of all the work done in the previous section and propose future lines of work and possible issues and difficulties of the subject (see *Conclusions* 6).

Chapter 2

Game Theory and PGGs

2.1 Game Theory

The importance of the study of game theory and human behaviour in games has been present in our lives since the publication of *The Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern in 1944 (Princeton University Press), the study of human interactions by what since is known as "game theory" has revolutionized the sciences, especially the social sciences and biology. In economics, for instance, game theory changed the way equilibrium concepts are understood and eleven Nobel Prizes have since gone to game theorists. Game theory provides a sharp language to formulate mathematical models of underlying interactions that promise clean predictions, now integral parts of the social sciences toolbox (2). Later on, with the work of John Nash, all the work in the field focused on the study of the mathematics of game theory framework (a general framework for decision making in uncertainty when payoffs depend on the actions taken by other players, (3)), and, in particular, with the *Nash equilibrium*. But with his work, John Nash also did research on the behavioral or experimental research and stated that *the study of experimental games is the proper route of travel for finding "the ultimate truth" in relation to games as played by human players*, (4).

The concept of game theory is interesting because any time we have a situation with two or more players that involve known payouts or quantifiable consequences, we can use this theory to help us determine the most likely outcomes from that situation. Furthermore, it can also be extrapolated to other many behavioural sciences because of its versatility of its min features which are the following:

1. **Game**: Any set of circumstances that has a result dependent on the actions of two or more decision-makers (players)

2. **Players**: A strategic decision-maker within the context of the game

3. **Strategy**: A complete plan of action a player will take given the set of circumstances that might arise within the game

4. **Payoff**: The payout a player receives from arriving at a particular outcome (The payout can be in any quantifiable form, from dollars to the abstract concept of *utility*.)

5. **Information set**: The information available at a given point in the game (The term information set is most usually applied when the game has a sequential component.)

6. **Equilibrium**: The point in a game where both players have made their decisions and an outcome is reached

As we previously said, with the great versatility and universality of the main features of game theory, during the late 1990's experts tried to use game theory as a universal language for the unification of all the behavioral sciences. The main fields or areas that had a direct application from the behavioural sciences are Economics and business, Psychology, Political science, Biology, Computer Science and Philosophy.

Specifically, during the application of the mathematical theory of games to biological contexts, a new theory arised (it was originated in 1973 with John Maynard Smith and George R. Price's formalisation of contests, analysed as strategies, and the mathematical criteria that can be used to predict the results of competing strategies) from the realization that frequency dependent fitness introduces a strategic aspect to evolution. That new theory was the *Evolutionary Game Theory*. With this new theory, the understanding of equilibrium changed. Now the focus is centered on discovering which equilibrium is the most stable and how they change through time (through the different rounds of the game).

From this new theory, plenty of new theoretical approaches were created in order to deeper study proposed in order to study the different. Evolutionary game theory encompasses Darwinian evolution, including competition (the game), natural selection (replicator dynamics), and heredity. Evolutionary game theory has contributed to the understanding of group selection, sexual selection, altruism, parental care, co-evolution, and ecological dynamics. Many counter-intuitive situations in these areas have been put on a firm mathematical footing by the use of these models.

The new equilibrium concept is related to the population of individuals that play each evolutionary game. The theorists started pointing that this population can always be segregated with three different types of individuals playing together:

1. *Fairmen* always demand exactly half the resource.

2. *Greedies* always demand more than half the resource. When a greedy encounters another greedy, they waste the resource in fighting over it.

3. *Modests* always demand less than half the resource. When a modest encounters another modest, they take less than all of the available resource and waste some.

Evolutionary game theory and its interacting population of individuals has been used to explain a number of aspects of human behavior. And one of the aspects of human behaviour that has been of a higher interest is a branch of experimental economics called **Public Goods Games**.

2.2 Public Goods Games

The evolution of cooperation among unrelated individuals in human and animal societies remains a challenging issue across disciplines.

Public goods games are social dynamics that involve three main components: they involve resources that are jointly provided; the resource is non-excludable in nature; it is also non-rivalrous. Economic theory describes the problem of public-good provision through individual contributions as a game in which agents have to decide how much of some resource to contribute to the creation of a public good and how much to spend on private goods. Owing to the non-rival and non-exclusive nature of the public good, the prediction made by equilibrium analysis is that players have incentives to free-ride on others and therefore individual contributions will be suboptimal. This effect is known as a social dilemma in the broader literature.

In this context, two models have attracted most attention: the prisoner's dilemma for pairwise interactions and the public goods game for group interactions. The two games share many features as demonstrated by the close linkage of their cores. In well-mixed populations with random encounters between individuals, cooperators are doomed and vanish quickly. However, in spatially structured populations with limited local interactions, cooperators are able to survive and co-exist with defectors in a stable equilibrium. Spatial extension enables cooperators to form clusters and thereby reduces exploitation by defectors. The geometry (square versus honeycomb), i.e. the connectivity, has pronounced and robust effects on the fate of cooperators. For example, in pairwise interactions cooperators thrive more easily on honeycomb lattices but for group interactions including all neighbors, it becomes increasingly difficult to promote cooperation in larger groups.

As we can see, this behaviours are very much related to the three types of individuals from the evolutionary game theory (*greedies*, *modests* and *fairmen*).

In a mathematical formulation, the payoffs for cooperators P_C and defectors P_D in a group of N interacting individuals are then given by:

$$P_D = \frac{rn_c c}{N}, \tag{2.1}$$

$$P_C = P_D - c, \tag{2.2}$$

where r denotes the multiplication factor of the public good, n_c the number of cooperators in the group and c the cost of the cooperative contributions, i.e. the investments in the public good. Thus, the total value of the public good is given by the number of cooperators n_c times their investment c and multiplied by r. From this total each player gets an equal share but cooperators have to additionally bear the costs of their contributions.

Such public goods interactions are abundant in human and animal societies. Consider for example predator inspection behavior, alarm calls and group defense as well as health insurance, public transportation, the fight against crime or environmental issues, to name only a few. Fortunately, and undermining the basic rationality assumptions in economics, human subjects do not always follow the rational reasoning and, of course, fare much better by doing so. From a theoretical viewpoint, the reasons for this outcome are not fully understood but likely involve issues related to voluntary interactions or reward, punishment and reputation.

Chapter 3

Explainable Artificial Intelligence

3.1 Concept of Explainable Artificial Intelligence

From then on, many researches had been made related to predicting human decisions with behavioral theories, but it was not until the mid-2010 when the revolution of Artificial Intelligence entered the field of Social Sciences, and specifically the field of human behaviour in the framework of game theory, (5; 6; 7; 8; 9), where many scientists applied supervised ML algorithms in order to predict human behaviour/decisions. It has also been demonstrated that it is possible to cluster (using unsupervised ML models), identifying discrete behavioural types on dyadic games and evolutionary games, (10; 11).

It is also need to be said that another most than used method for predicting human behaviour on games has been *Deep Reinforcement Learning* and it has produced surprising advances. Deep RL involves training an "agent" to become adept in given "environments," enabling algorithms to meet or surpass human-level performance on a diverse range of complex challenges, including Atari video games, the board game Go, and subtle hand-manipulation tasks (12).

Nowadays, the research on human behaviour inside the game theory framework is mainly focused on trying to implement deep learning methods using generative modeling, such as convolutional neural networks and *Generative Adversarial Networks (GANs)*. Specifically GANS are the new path to follow because of its potential. *Generative Adversarial Networks* are an unsupervised Deep learning model that its architecture is based on two types of ML models, a supervised one that is the responsible of creating the generator model and an unsupervised model responsible of the discriminator model. The main idea of adversarial networks is based on a theoretical scenario of games in which the generating network has to compete against an adversary. The job of the generator network is to produce random samples. On the other hand, its adversary, the discriminating network, seeks to distinguish between samples taken from the

training data and samples extracted by the generator model, (13; 14).

On this sense, in our work, we will try to implement as many ML models as possible in order to compare the results obtained for each one of them and be able to conclude which methodology could be the best when speaking of predicting human behaviour inside the game theory framework.

Nowadays, Artificial intelligence (AI) is at the forefront in changing the world and the way we live. Yet, as AI becomes more sophisticated, more and more decision making is being performed by an algorithmic 'black box'. To have confidence in the outcomes, it may be necessary to know the rationale of how the algorithm arrived at its recommendation or decision – *Explainable AI*. Yet opening up the black box is difficult and may not always be essential, *Explainable AI* may reveal the aspects of the decision making process that provide a meaningful explanation to humans, Figure 3.1, (15; 21).

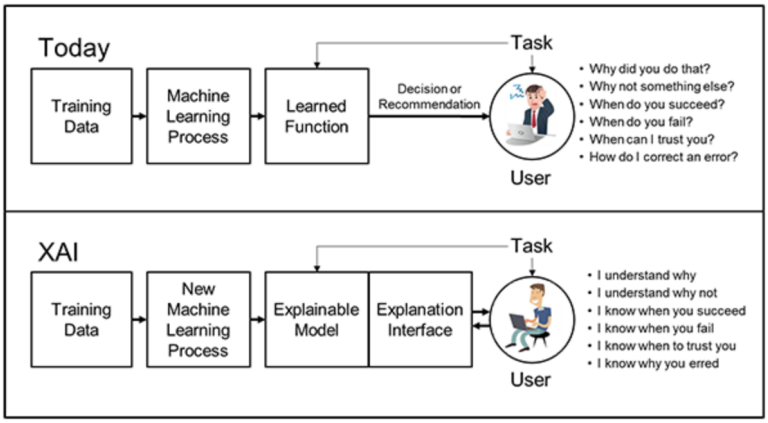

Figure 3.1: XAI Concept.

From that necessity, the *XAI* program was born, (15). The *XAI* program is focused on the development of multiple systems by addressing challenge problems in two areas: (1) machine learning problems to classify events of interest in heterogeneous, multimedia data; and (2) machine learning problems to construct decision policies for an autonomous systems. *XAI* is one of a handful of current *Defense Advanced Research Projects Agency (DARPA)* programs expected to enable "third-wave AI systems", where machines understand the context and environment in which they operate, and over time build underlying explanatory models that allow them to characterize real world phenomena , (15).

Taking all of this in hand, we will focus on how to make supervised machine learning models interpretable and self explainable.

When ML models do not meet any of the criteria imposed to declare them transparent (16), a separate method must be devised and applied to the model to explain its decisions. This is the purpose of post-hoc explainability (post-modeling explainability), which aim at communicating

understandable information about how an already developed model produces its predictions for any given input. The great advantage of model-agnostic interpretation methods over model-specific ones is their flexibility, they can be applied to any kind of supervised ML model, (17).

On this sense, during this thesis we will study model-agnostic methods in order to derive our ML learning models into explainable and interpretable supervised ML models. This model-agnostic methods are *SHAP* and *LIME*, but we will finally only use *SHAP* in our Machine Learning Models' section.

3.2 Model-agnostic methods

SHAP is a local surrogate model approach to establishing feature importance. It is based on the game theoretically optimal *Shapley Values*. *SHAP* assigns each feature an importanec value for a particular prediction. The Shapley value of a feature's importance is its average expected marginal contribution after all possible feature combinations have been considered. The *Shapley value* guarantees to perfectly distribute the marginal effect of a given feature across the feature values of the instance. Thus *SHAP* currently produces the best possible feature importance type explanation possible with a model agnostic approach, (17; 18; 20).

SHAP is based on the *Shapley values* (named after Lloyd Shapley) which are a solution concept of fairly distributing both gains and costs to several actors collaborating. The Shapley values apply in situations when the contributions of each actor (in our case, actor refers to each feature of the dataset obtained from the PGG) are unequal, but they collaborate to obtain the payoff (the solution value of the machine learning model, in our case). Shapley values also ensure that each feature gains as much or more as they would have from acting independently. For each variable (payout), the Shapley value, is basically trying to find the correct weight such that the sum of all Shapley values is the difference between the predictions and average value of the model. In other words, Shapley values correspond to the contribution of each feature towards pushing the prediction away from the expected value.

On the other hand, *LIME*, is a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. It does so by performing various multi-feature perturbations around a particular prediction and measuring the results. It then fits a surrogate (linear) model to these results from which it gets feature importance, capturing local feature interactions, (16; 18; 19).

Instead of training a global surrogate model, *LIME* focuses on training local surrogate models to explain individual predictions. Thus it is inherently local. On the contrary, shapely values 'decompose' the final prediction into the contribution of each feature - this is what some mean by 'consistent' (the values add up to the actual prediction of the true model, this is not

something you get with *LIME*).

That is why we will focus on *SHAP* to explain our individual predictions of the machine learning models we will create to better understand how individuals' strategies affect the rounds of the PGG and which features are the most importants when contributing more or less to the common fund.

Therefore, the goal of *SHAP* is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. The *SHAP* explanation method computes Shapley values from cooperative game theory. The feature values of a data instance act as players in a cooperation (as we previously mentioned). Shapley values tell us how to fairly distribute the "payout" (value predicted in our model) among the features. In our case, a player will be an individual feature value. Also, one of the great innovations that *SHAP* brings is that the Shapley value explanation is represented as an additive feature attribution method, a linear model.

$$f(x) = \beta_0 + \beta_1 x + ... + \beta_n x_n, \tag{3.1}$$

where x is the instance for which we want to compute the contributions. Each x_j is a feature value, with $j = 1, ..., p$. The β_j is the weight corresponding to feature j.

$$\phi(f) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j), \tag{3.2}$$

where $E(\beta_j X_j)$ is the mean effect estimate for feature j. And if we sum all the contributions the result we get is the following:

$$\sum_{j=1}^{N} \phi_j(f) = \sum_{j=1}^{N} (\beta_j x_j - E(\beta_j X_j)) = f(x) - E(f(X)), \tag{3.3}$$

where this result is the predicted value for the data point x minus its average predicted value. In the previous equation 3.3 of all contributions we can see that feature contributions can be negative and this gives us clear interpretability of the feature importance on each data point on each feature, as we will see later on.

It is also important to say that the linear model 3.1 has three important properties needed for the correct explaniability of the predictive machine learning model. These properties are: *Local accuracy*, *Missingness* and *Consistency*. This properties ensure that the algorithm is efficient, meaning that it needs fewer observations than a less efficient one to achieve a given performance, that it will also take care of missing features giving them a Shapley value of 0 and it also will be consistent, meaning that if our model changes so that the marginal contribution 3.2 of a feature value increases or stays the same (regardless of other features), the Shapley

value also increases or stays the same.

To sum up, we have seen that model-agnostic methods in Social Sciences are a new path that started not so many years ago. Nowadays, we are constantly using deeper and deeper machine learning models and neural networks for all tasks, thus this makes it very much harder to interpret and explain its results.

So, that's why we propose the applicability of model-agnostic methods in order to correctly interpret our predictions and start creating an inherent dependence of understanding what is happening inside the black boxes models that we build. Because, as we previously said, the ability to correctly interpret a prediction model's output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled and it turns even much more important when the model is predicting the human behaviour (18; 19).

Chapter 4

Public Goods Games' Datasets

In this section we will review some papers that are related to *Public Goods Games* and that have experimental results in order to see which one of them could be helpful in our future work plan. We will start with a *literature review* and then, if we find the experimental data and results interesting for our future work we will start with an *exploratory data analysis* to do a first overview of the experimental data.

4.1 Hunger affects social decisions in a Public Goods Game but not an Ultimatum Game

4.1.1 Paper review

This paper investigated the effects of hunger (breakfast vs non-breakfasts) on cooperation via 2 different experiments. The first one being an *Ultimatum Game* and the second one being a multi-round *Public Goods Games* (with 10 rounds of non-punishment possible and 10 rounds of possible punishment). The results for experiment 1 revealed no significant effects of hungry. However, experiment 2 showed that hungry has effects on cooperation. The first interesting result from experiment 2 was that the non-breakfast participants contributed more to the group in the non-punishing game (10 rounds). The second finding was even more interesting showing that for the non-breakfast groups in the punishing game (10 rounds) the higher the contribution amount was made to the group, the lesser punishment was established.

With the data of experiment 2 available we could try to cluster (non-supervised learning) groups from both games (punishment and non-punishment) using many of the data available (how hungry, breakfast/non-breakfast, time contribution, gender, group share, group contribution, initial payoff, group nº, etc.). We could also try to predict the desired amount of contribution (e.g. >10 or =<10) and apply classification models, or the time spent on deci-

sions for a train and test sets of the data and therefore be able to apply Python libraries for model explainability in order to know the feature importance of the parameters.

4.1.2 Exploratory Data Analysis

Since we want the maximum amount of features that could influence individual behaviour and change the individual contribution, we will analyze the information on the Experiment_time dataset since we have information about the hunger, elapsed time between answers and punishment information. For that purpose we will also only work with the *Public Goods Games* dataset and not with the *Ultimatum Game* dataset (23).

First of all we take care of the *Data Cleaning* step. Secondly we divide the *Exploratory Data Analysis (EDA)* with two independent datasets. We obtain the *punishment game* and the *non-punishment game* to be able to study them independently and see if there are different behaviours depending on the initial features of the PGG.

For the *Punishment game* we saw that Hunger is a relevant feature. The contributions, total payoff and final payoff differ from the total mean of those features which are 7.48, 23.14 and 393.46 respectively. And we may also see that the elapsed time is also bigger than the total mean (9.01) on those participants have values between 1-5 on How hungry.

On the *Non-Punishment game*, the *hungry* participants took less time to participate, had a lower *final payoff* and contributed less.

During the *EDA* we have studied that Hunger feature is a relevant feature. The individual contributions', *total payoff*, *final payoff* and *punishment sent* and *punishment received* differ from the total mean of those features which are 11.58, 16.20, 393.46, 2.72 and 6.83 respectively. And we may also see that the elapsed punishment time is also bigger than the total mean (17.13) on those participants have values between 1-5 on How hungry.

On the *Non-Punishment game*, the hunger participants took greater time to participate, contributed less, punished more, received less punishment and had lower payoffs.

We can also see that the *contribution* and *group share* features were 4 points and 6 points bigger during Punishment rounds than in no Punishment rounds, but the *Final Payoff* was greater on the no-punishment rounds. The elapsed time was also bigger, consistent with a greater amount of time needed to think on the answer. So, as we have seen on these visualizations, punishment creates the contrary effect of hunger.

Then we applied a linear regression feature importance analysis. From that analysis we can see that there are no clear linear correlations with the features (because the coefficient of determination of all the linear regressions were <0.25). However, we may see that with a first linear approximation which of the features play a more important role on the contribution of an individual which are the amount of punishment received and the hunger and the most

important features on the group share and payment sent/received are the time features. So, the amount of punishment that an individual may receive is a fear factor that could be related to a greater amount contributed to the common benefit.

4.2 Group size effects and critical mass in public Goods games

4.2.1 Paper review

This paper studied the relevance and importance of punishment on stable cooperation in Public Goods Games and in particular, they study ostracism (27) as a particular way of punishment.

In order to be able to study its relevance they prepare an experimental setup which consists of a player choosing between a standard public Goods game (game A) and a public Goods game with an option to exclude members from the group (games B10 and B8, where the difference depends only on the initial endowment). In addition to these two games, and in order to distinguish between the effect of self-selection and the effect of the institution, they also conducted two additional treatments, B10-exo and B8-exo, in which groups played games A and B over the same number of rounds but, unlike the groups in the endogenous treatments, these groups could not vote on the two games but had to play the game that was announced by the computer.

The final results showed that the exclusion institution (B10 and B8, game B) increases contributions to the public Goods. Subjects who had been excluded or received a vote for exclusion adjusted their contributions closer to the group average in later rounds and subjects who vote for the exclusion institution contribute significantly more than those who vote against it. But if the number of social players is too low to implement the exclusion institution, the contributions of the supporters and the opponents of the institution are similarly small, just as the social preferences models predict. They also stated that the players' institutional choice can be better explained by assuming social preferences than by the standard model of purely selfish players (because in the standard model the threat of exclusion is not sufficient to sustain cooperation).

To sum up, they name two factors that reduce the chances for cooperation: first, if the share of social players within the group is smaller and, second, if the social players do not have the exclusion institution available to exclude the other players from the group.

Finally, they showed that that effect of the institution itself is more important than the sorting and signaling that comes with the endogenous choice.

With all that being said and with the data available, we could apply a clustering method

to identify selfish and social preferences groups of people (or individuals) to explain better why the results obtained in the paper were largely inconsistent with standard economic theory, where the threat of exclusion is not sufficient to sustain cooperation and a Nash equilibrium (0 contribution) should be reached. We could also classify data from a test/train set with different ML models taking an special care to the feature importance of the parameters that influence more the economic behaviour or social preferences behaviour (resembling the two layers from the thesis (26)).

4.3 Voting on the threat of exclusion in a public Goods experiment

4.3.1 Paper review

On this paper they studied the possible effects on cooperation of the group size with a critical mass (being the constant vale of the return after a critical mass is reached). They ran an experiment participants played one round of a N-person general public Goods game with Critical Mass and sessions differed in the size of the interacting group.

Their results stated that critical mass (maximum number of cooperators) has no effects on cooperation the difference is not statistically significance theoretically nor experimentally. That group size has a positive effect on cooperation. The last finding was that the most played strategy was *All Defection* followed by *All Cooperation* and where the rest have a slight trend to switch preferentially from defection to cooperation as the group size increases. The main discussion on the paper is why the inverted-U effect of group size on cooperation (a public Goods game with a piece-wise linear-then-constant return for cooperation should give rise to a inverted-U effect of group size on cooperation, if the cooperative equilibrium model is assumed) is not observed. The reason why they did not get that effect was because it would require a significant proportion of subjects that change strategy twice: from defection, to cooperation, and then back to defection in order to observe the inverted-U effect that appears.

With the data available and knowing that there is a high percentage of dropout and also that the results were very much in contrast with theory we may think that this data would not be the best to use in our study.

4.3.2 Exploratory Data Analysis

We start our *EDA* for the Voting on the threat of exclusion in a public Goods experiment (24) dataset. Previously to the analysis, we may say that we only used the information about

the first experiment, the stacked results from the 8 sessions run in June 2016 (endogenous treatments). That's because, as it is stated on the original paper, contribution rates are very similar in the endogenous treatments and the exogenous treatments. Contributions in game B are slightly higher in the endogenous treatments than in the exogenous treatments (in both B10 and B8). For that reason we think that we may find more interesting results regarding individual's behaviour in the endogeneous treatment, (24).

To start our *EDA* we will first we take care of the *Data Cleaning* step. Secondly, we start selecting only the main features that are also selected on the original paper (24). We compared the behaviour depending on the treatments 1 or 2 (choosing between B10 and A games, or B8 and A games). And we can see that for *Treatment 1* all the values are greater than on treatment 2. *Contributions*, *profit*, *ostracism*, they all have greater values on the *treatment 1*. This behaviour can be explained because players, the majority of time, chose to play B10 game, as stated on the original paper (24). We can also see the exact same behaviour if we compare the two different endowments, 10 and 8 with the kind of treatment. The explanation that for treatment 1 we saw greater values is that a greater amount of endowment is offered on those games. Therefore, the greater also are the contributions and the greater also is the sum of ostracism observed (because more money is involved in the game). At this point, we might start observing the appearance of a concept of fear of being voted to abandon the group and get no final profit.

We now group the numerical attributes for the *Ostracized* feature and we can see that as voting rounds get passed, more individuals are voting for ostracism and the mean profit of the groups voting in favour of ostracism increases until it reaches its maximum. We also see that the subjects who vote for the exclusion contribute significantly more than those who vote against it.

We finally apply the linear regression feature importance model and we can see that the linear regression states a great linear relation (coefficient of determination > 0.97 between the contribution and the votes of the members of the groups (after round 1). We can see that the support becomes stronger over time, specially after the first voting round, when players are getting used to play with their group and it is also when the game methodology becomes better known for all players. This results are also related to the great variance of the dependent variable (*contribution*) that can be explained with the independent variables in our regression model (EndowmentB, Ostracized[i]).

4.4 Large scale and information effects on cooperation in public goods games

4.4.1 Paper review

This paper studied the effects of societal large scale and information on cooperation in *Public Goods Games*. They setup an experiment consisting of three different treatments with two different experimental populations, 100 and 1000 respectively. The difference between the treatments was the information provided to each participant before each round. The first treatment was based on giving each participant the following information about the their own contribution, their earnings in the past round, their cumulative earnings, and the average group contribution in the past rounds. The second treatment was where the distribution of the other people's contributions to the pool in the past round was provided along with the same information that participants received in the first treatment. The third treatment was identical to the second one except that people's average contribution is omitted. All experiments lasted 14 days, a round per day. Thus they had 24 hours to make a decision on how much to contribute in that round, based on the information they had. In all PGG treatments, the endowment was the same, 10 points, and they had to decide how many of them (0, 2, 4, 6, 8, 10) wanted to contribute to the common pool. Finally their marginal per capita was 0.1 in the 100 experiment and 0.01 in the experiment of 1000 participants.

The data and therefore the results obtained from this experimental setup were very much interesting. They found that the distribution of contributions per round in PGG100 and PGG1000 were rather similar. That they both were centred around the average value of the endowment, closely following the average contributions of the group. That the average levels of cooperation of these two treatments are indistinguishable, suggesting that a group size of 100 individuals was a Goods representative of large groups. And that their potential differences were not statistically significant. Thus, they concluded that there were no relevant differences between the experiments with 100 and 1000 subjects, at least, as far as collective behaviour was concerned.

Related now to the information effects they found no differences on the treatments nor population by studying the average contributions, but they found interesting results with the distribution of players' contributions and players' behaviour. They found different individual behaviours depending on the amount of information provided. From this result they decided to get further insights on individual behaviour. They conducted a series of statistical analysis concluding that informing players on how many people contribute what may have been a promoter of cooperation. Finally they also tried an alternative approach to finding hidden

patterns of behaviour or *types of people*, where they applied a hierarchical clustering to the time series of decisions (contribution per round) and they found that three clusters, for all the treatments, were the most appropriate number of *types of people*.

So, from the results obtained and from the data available we could try to cluster the participants and study which parameters have more feature impact on the cluster decision. And then implement more ML models in order to classify nor predict a train/test set of data focusing on the feature importance of the information given to the subjects.

4.4.2 Exploratory Data Analysis

First of all we take care of the *Data Cleaning* step. Secondly, we start grouping the information about each round in order to compare the information and conclusions extracted about the original paper Large scale and information effects on cooperation in public good games (22) to extract further conclusions.

From the first analysis (grouping the information of each round) we can see that the *group_avg_contribution* reduces when the nº of rounds is decreasing, *player_contribution* is also decreasing and it has its peak on middle rounds. But on average, the contribution remains constant with ± 0.75 in comparison with the mean values of the *group_avg_contribution*.

If we now focus specifically on the large scale and information features and group the contributions around the size feature, we can see that the size of the groups has influence only in the games where more information was given to the participants (H and HM). It also may be seen that that the individual player contribution and group average contribution increases more than 20% on these two games.

To sum up, we can see that group size only becomes a relevant feature when much more information is given to the participants, otherwise it is not.

From the previous results, we will now classify the data between the games PGG_100 and PGG_1000 and the other two; PGG_HM, PGG_HM2, PGG_H, PGG_H2.

We will now apply two feature importance (linear regression) algorithms to further understand the impact that features have on the player's individual and group contribution.

From the linear regression analysis we can see that there is no linear relation between the variables studied (coefficient of determination <0.1). However, we can see that the impact of group average contribution is greater on the PGG_100(0) dataset, because on the PGG_HM the importance of a larger group of participants decreased the player's individual contribution in accordance with the previous results observed.

We then next analyze the feature importance using a Tree Classifier and PCA algorithms on the categorical *Treatment* feature.

The results obtained with this new analysis are not surprising at all and are in concordance

with all the previous conclusions. From the classification of the random forest we can clearly see with a great correlation coefficient that there is a high relation (coefficient of determination >0.99) between the group average contribution and individual contribution depending on the information gave to the participants on each game.

So, as we can see, information clearly becomes an important feature on the individual's cooperating behaviour to the common fund.

4.5 Resource heterogeneity leads to unjust effort distribution in climate change mitigation

4.5.1 Paper review

On this paper they studied the influence of inequality on cooperation in collective risk dilemmas (climate change mitigation). This inequality was studied both from the collective (reaching the common goal) and from the individual (how different individuals behave under different circumstances) visions of the problem. This approach let them identify how agents reacted to resource heterogeneity and what actions must had been taken to promote environmental justice. The game that the subjects had to play was based on an initial distribution of six different endowments to the participants of each group and its final goal was to contribute 120 on a common fund in order to avoid a climatic catastrophe. On the paper they also prove that from theory it was shown that there was not a clear theoretical prediction about what should happen in their heterogeneous version of the game, with an initial distribution of six different endowments to the participants. During the game the subjects had to contribute into the common fund between 0 and 4 during 10 rounds. At the end of each round all players saw the information of how much money had been contributed to the common fund, and they also saw the individual contributions of the six players in the previous round and the total amount contributed by each player up to this round. Then, if the goal was reached at the end of the game, all the participants kept the money that they had not contributed in the form of a gift card, but if they did not reach the goal, the participants only kept the remaining money with a probability of 10%. So, on this game the participants were also playing with another condition which was information about the individual's contributions of each of the players of the group at the end of each round.

As we previously said, on this paper there were two analysis. They started with a group analysis, where they studied the collective climate action of the groups and they found that all groups reached the goal and disadvantaged individuals were contributing much more than a fair share of the mitigation, and that the richest ones were contributing less. It is telling

that all hyper-generous behavior is observed in the two poorest types of individuals, while a large majority of those endowed with the largest amount behaved selfishly (irrespective of what they claim to believe about fair contributions, as we have seen). They also found that subjects with lower education level were predicted to make higher contributions in equal condition This result may be related to the one found in (23) where the non-breakfast participants (related to the most deprived communities) contributed more to the group, or to the common goal. The second analysis was from the individual's behaviour. They applied a hierarchical clustering and they found 2 clusters from a homogeneous treatment and 3 clusters of individuals from the heterogeneous treatment. The final conclusion was the suggestion of establishing concrete thresholds to be reached in a particular time period and rewarding the population may work to address people towards a general cooperation. As they observed in their experiments.

From the data available it can be seen that there are 3 different studies that can be done. The first one is the one addressing the pre-game questions, the second one is referring the collective behaviour and the individual behaviour. From this data we could create an unsupervised model to cluster individuals or group behaviour and then apply ML models to classify/predict train/test populations focusing on the conditional social features that are the initial endowment of the group/individual, the amount contributed and the pre-game questions.

4.5.2 Exploratory Data Analysis

We will start the *EDA* for the Resource heterogeneity leads to unjust effort distribution in climate change mitigation (25) dataset. We previosly need to say that we will only use information about the user's contribution during the games, we will use the *userround* and *user* datasets (25). That is mainly because we want to stick with each individual's behaviour.

First of all we take care of the *Data Cleaning* step. Secondly, we start grouping the information about the categorical feature *study_level*. And as we may see we cannot infer a direct relationship between the winnings and the study levels because the group of participants with the higher level of studies also has the greater amount of starting capital.

We may now aggregate the information about each round of each player, because it could be interesting to see if the amount contributed varies severely depending on the round. The contribution's option most selected was 2, which makes sense being the mean vale of contribution options. However, the amount contributed that was the most selected for the players with a greater starting capital was proportional to the amount contributed.

Next we will study the education level of the participants with the novel feature of the contribution on each round. The variation of the amount contributed for each study level is on the confidence interval of 99%. So, there is no apparent relation between both features.

We may now study the feature importance with the feature selection methods that are

intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. We will focus on the option_selected and the features that are the most important on predicting the contribution's output. We will then start with the categorical variable study level and continue with the winnings and contributions.

With the fit of a linear regression, we may see that the starting capital is the feature that has a greater influence on the individual contribution, as it was stated on the original paper (25). However, there is a 76% of variability for the dependent variable *winnings* that is still unaccounted for. For that exact reason we will base our study on the contribution of each round rather than focusing on the winnings.

From that statement, we aggregate by *start_capital* and see that disadvantaged individuals (lower starting capital) are contributing much more than a fair share of the mitigation, and that the richest ones are contributing less. With this information we may now create a new variable that will account on the proportion of the amount contributed/starting capital.

And if we finally analyze the same aggregation (for *start_capital* but using the new feature *contribution_scaled* than before, we can clearly see what we were stating before. Disadvantaged individuals (lower starting capital) are contributing much more than a fair share of the mitigation, and that the richest ones are contributing way less (from 7.15 on disadvantaged individuals to 4.66 on rich individuals).

4.6 The public Goods game on multiplex networks

4.6.1 Paper review

This paper implemented two different models to increase cooperation on groups. The first model was the simplest. It consisted on the frequency of cooperation on multilayered networks (specially on multiplex networks, where each node is present on many layers, but the relationships between them may be different on each, and different strategies may be played on each layer). This first model studied the effect on cooperation of strategy updates across layers finding that how often players chose whether to cooperate or defect in each of a number of connected economic networks alters how cooperative the entire population is. The second model included a layer of social pressure on top of the standard economic (where economic standard model layer. It consisted on a model where strategy updates occurred by maximising both economic payoff and adhering to local social behaviour. According to the author there are three factors that determine how cooperative each player is: the amount that players have previously donated; how much they have accumulated in the game; and how cooperative others are. And he found that the inclusion of both conditional strategies and social pressure resulted

in higher levels of cooperation, and that the initial cooperativity of the population was a crucial factor.

Through the thesis, the author extracts many results and conclusions from his work that echoed the empirical results observed on the state of art of the cooperation in social dilemmas. That was one of his main motivations.

One of the conclusions that can be extracted from its work is that it shows that the inclusion of social considerations in a player's strategy in the public Goods game can have a significant effect on the amount contributed to the public Goods. Meaning that even if a player has a rational goal (in this case a large payoff), a very small amount of social pressure may take them far away from this. Another conclusion is that in finite populations each player updates its strategy by moving towards the strategy of those that are performing better, and the strategies that perform better are heavily dependent on the group composition. So, this means that the information given to a participant may be a crucial factor in order to make him perform better.

The conclusion extracted from the two different layers in model 2 is that the dominant effect is the influence from the social layer, where for even a small probability of considering the social neighbour's strategies the mean strategy very quickly moves towards the population mean. So, that the economic layer (following the pure strategic reasoning) is not the dominant.

Finally, the author computed the models into communities where he created each community with three different strategies (clusters): free-riders, weak conditional cooperators and conditional cooperators.

And the conclusion extracted from the computation was that in order to increase the population strategy to a point where the amount donated to the public Goods increases after each iteration, a large number of very cooperative individuals must be introduced. The results in this section showed that these communities were resilient to this intervention, and that if the system is to be shifted to a more cooperative state additional mechanisms are required, punishment of free-riders, for example (ostracism or punishment in hunger in PGG).

With all of the empirical results supported by the models, there was an empirical result that was not echoed by the models. That was the choice of population structure to understand the dynamics of the game and the effect of social influence, but it did not model the empirical populations well. And the author explains that to achieve a further understanding, time scales involved must be important (time response of each participant, hunger PGG), and that investigating this would be a Goods next step to understand the behaviour of these populations.

With all that being said we may think that this thesis might be beneficial for our work in order to be able to explain all the conditions (from the social layer) that may have effects on cooperation in PGGs from the other papers. The data available for this thesis might not be beneficial in order to compute any machine learning model because they are numerical results

from computations of the models that try to explain the behaviour of the empirical results from the state of art. Thus, it may be great for explaining the results that we may obtain from our ML models and from our ML explanations using XAI libraries.

Chapter 5

Machine Learning Models

From the previous Public Goods Games analyzed we will focus on the *Large scale and information effects on cooperation in public Goods games* dataset and *Resource heterogeneity leads to unjust effort distribution in climate change mitigation* dataset.

We will focus on these two datasets in order to create the pipeline of models with the final explainability libraries, because they have more features that can raise greater discussion in feature impact on the final machine learning model.

5.1 Large scale and information effects on cooperation in public goods games

5.1.1 Unsupervised Clustering

In this section will focus on the features that are more relevant related to cluster the behaviour of individuals on the PGG (22). The main goal of this section is to obtain the differences the change in players' contribution behaviour depending on the features used on their unsupervised ML model for clustering.

And we will start by utilizing the previous preprocessed data, where each individual was assigned to a row and each round's contribution was stored into a column (feature). Then, we will apply clustering for the individual round's contributions, followed by the round group contributions and followed by the information effects (the type of PGG treatment). We are also going to distinguish the analysis between the PGG treatments, meaning that, as we previously said, we are going to study the PGG treatments where some information is given to the participants (PGG_H, PGG_H2, PGG_HM, PGG_HM2 treatments) separately from the PGG treatments where no information is given to the participants (PGG100, PGG1000 treatments). We obtained 3 as the best number of clusters (through the elbow and silhouette methods) as

we can see in 5.1. We may see that the cluster with the greatest amount of participants is

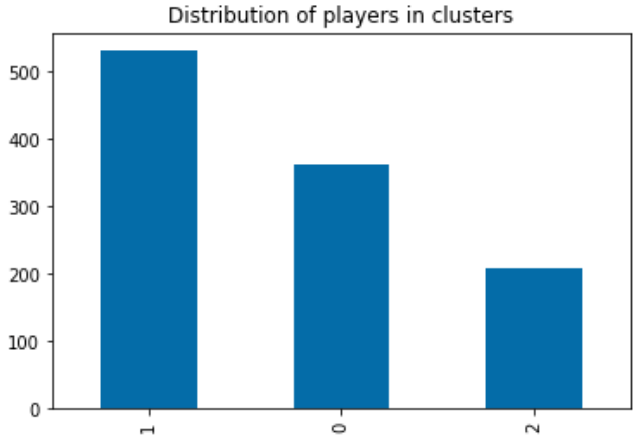

Figure 5.1: # Players in each cluster.

Cluster 2 (Cluster=1), followed by Cluster 3 (Cluster=2) and Cluster 1 (Cluster=0), but we do not know yet what kind of behaviour participants/individuals of each cluster will have. For that reason we will continue plotting the mean contribution of the participants of each cluster against the mean contribution of all the participants. From that plots A.1, A.2, A.3 we can see that participants in cluster 2 are contributing, on average, like the mean contribution value of all participants. And we can finally see that we found three different "types of participants" on the non information treatments and focusing only on round individual contributions, *greedies*, *fairmen* and *modests* from the evolutionary game theory 2. We find, a first cluster of low contributors (*greedies*), with an average contribution that decreases in time; a second cluster of people who, on average, follow the mean contribution *modests*; and a final group formed by full cooperators and generous participants (*fairmen*). If we now group the participants in the different *treatments* we can see in figure 5.2 that in PGG1000 treatment there is the greatest amount of high contributors, meaning that group size, as we stated in the (see chapter 4) is always affecting the collective behaviour on contributions and final payoff to the group.

After this previous clustering, we also applied clustering in the average group contributions and individual contributions separating also the original dataset into 3 smaller datasets from the three different information game treatments. Once we obtained the previous results, analyzing all the different combinations, we can state that giving information about the contribution (histogram or histogram and average contributions) of the other participants in the group to each participant on each round, translates directly into an increase in higher contributors because the clusters with the greatest amount of participants contributing more than the average contribution are located on treatments PGG_H, PGG_H2, PGG_HM and PGG_HM2. We can also see that in PGG100 and PGG1000 the greatest amount of participants were located in the

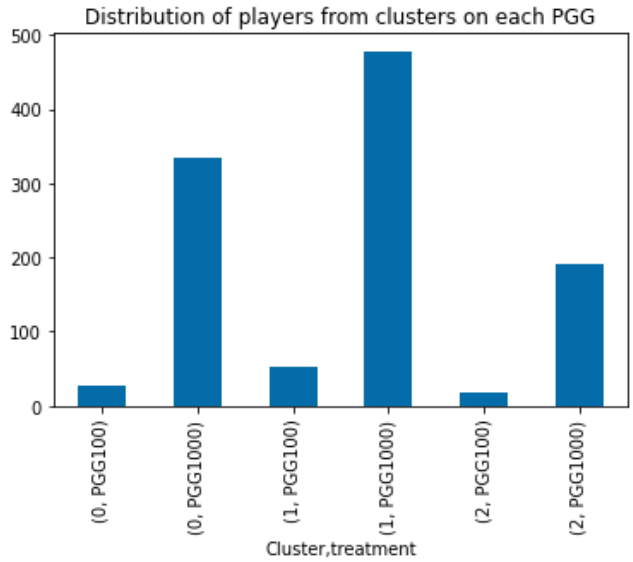

Figure 5.2: # Players from each PGG distributed in clusters.

average contribution behaviour. Furthermore, the lowest average contribution of the lower contribution clusters of all combinations was found to be on the information treatments' clusters, supporting the idea that the information given, gives power to the ongoing/average contribution of the players of the group. We may also see that, as it could have been expected, the results were found to be more interesting when round individuals contribution was analyzed.

Finally, we can also see that group size effects are more visible on the PGG100 and PGG1000 where no information was given to the participants.

Now, we want to study the behaviour in each round, trying to know when the behaviour of the participants changed to a higher contribution or to a lower contribution. For this new section, we will use supervised ML models and we will recall the game theoretic approach (Shapley values, see chapter 3) to explain the output of those machine learning models.

5.1.2 Supervised Clustering

As we previously said, we will now focus in the individual contributions per round of each participants of each of the three information treatments to study and understand the behaviour of the participants of each cluster in each round.

We will start the creation of supervised models with a *random forest classifier* to investigate in the feature importance on each participant classification from each clusters, in order to better understand the behaviour of the participants. We will also apply the *SHAP*'s python library to be able to calculate the *shapley values* for each feature, where the features will be the rounds and the different treatments and we will see which one of them influences the most to the

individual and group contribution.

We have seen that the accuracy shows that for the no information dataframe we obtain better results, maninly because we also have a greater population for the test set. The accuracies obtained are the following:

1. PGG100 and PGG1000 accuracy: 0.956
2. PGG_H and PGG_H2 accuracy: 0.881
3. PGG_HM and PGG_HM2 accuracy: 0.896

The accuracy values are comfortable for us to proceed with the explainability library and calculate the Shapley values (each feature contribution to the prediction class, weighted and summed over all possible feature value combinations) for all features in the test set. This values will help us to study, with a game theory approach, the contribution of a feature value to the difference between the actual prediction and the mean prediction for each feature.

We will continue the supervised models with a *multi layer perceptron classifier*, a class of feed-forward artificial neural network where we apply a previous normalization of the features. We will use this supervised model, as well as the previous random forest to classify the participants to each cluster and better understand the importance of the features on that classification.

We may see, as from the previous supervised model that the accuracy obtained on *MLPC* is better for all datasets. The accuracy for the no information dataframe is still the best in accuracy, but the other two datasets are also very high. The accuracies obtained are the following:

1. PGG100 and PGG1000 accuracy: 0.978
2. PGG_H and PGG_H2 accuracy: 0.970
3. PGG_HM and PGG_HM2 accuracy: 0.955

The accuracy values are also comfortable, even more comfortable than the values obtained in the random forest classifier for us to proceed with the explainability library and calculate the Shapley values (each feature contribution to the prediction class, weighted and summed over all possible feature value combinations), as we did before, for all features in the test set. This values will help us to study, with a game theory approach, the contribution of a feature value to the difference between the actual prediction and the mean prediction for each feature. We are going to focus in the clusters with higher percentage of high contributors, where the average contribution amount is higher on each of the three PGG treatments.

5.1.3 Explainability

PGG100 and PGG1000:

We will start with the first dataset where participants (100 participants and 1000 participants) do not receive much information about the other's contribution. The information that each participant receives after each round is the standard one: her own contribution, her earnings in the past round, her cumulative earnings, and the average group contribution in the past rounds.

We will start the analysis by creating the collective force plot. This plot combines all of the force plots for each participant on the X_test set (on the horizontal axis) where we can see the SHAP values for all features. Force plots give us information about the output value, which is the prediction for that participant and the direction on which the features are pushing the prediction, to higher output values (red), or to lower output values (blue); meaning if one feature has positive, or negative impact on the prediction. We can see that the force plot using

Figure 5.3: Force plot with the MLPC supervised model.

the Neural Network Classifier (figure 5.3 is less noisy than the Random Forest Classifier due to the normalization of the features' values. We can also see a higher average predicted cluster membership than the previous supervised model. A common point between both supervised models is the fact that we can clearly see two outstanding groups of participants where the positive influence of features clearly increases the output value. We next study an individual force plot from the outstanding group of participants, red cluster of participants (participant n=300). And we are able to see, as expected, that the behaviour in participant 300 is clearly positive and has increased the output value from 0.547 to 0.73. This participant had positively contributed in all rounds clearly being a higher average contributor.

We will then continue analyzing the role that each feature plays on the group with higher membership to the cluster of higher average contributors in the last 50 participants.

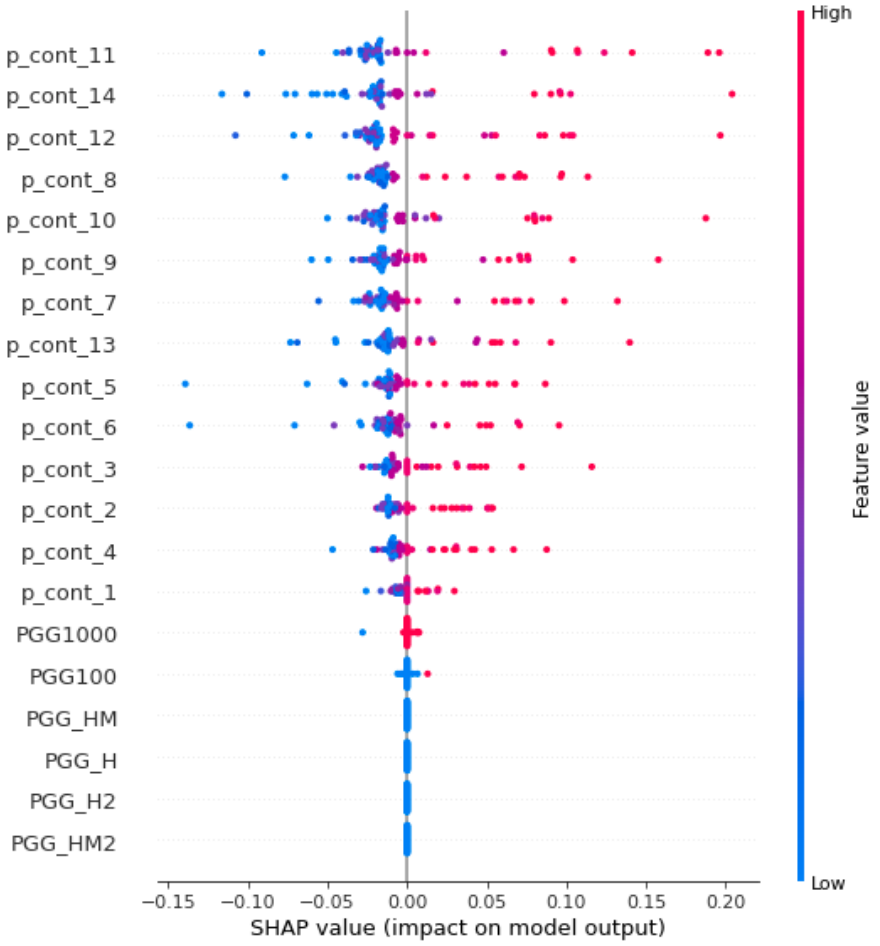

Figure 5.4: Mean absolute value of the SHAP values for the last 50 rows.

We can see that the higher the contribution in all rounds, the higher the increase in the output value (see figure 5.4. Furthermore, we can see that the vast majority of contributions are decreasing the output value, but those contributions (colored in blue) are almost in the null *SHAP* value (the impact on the model output). That means that with only a few high contributions and specifically in some of the last rounds, a participant has more probability of being classified for the algorithm as a higher average amount contributor.

Finally, we will create the *SHAP* feature importance plot for all clusters and all features to compare the relevance of the features on each cluster on both supervised models.

From figure 5.5 we can see that in agreement to the previous statements, for Class 2 the rounds with more feature importance on the classification are rounds 7, 10, 9, 12, 13. But in general, rounds 8, 7 and 6 are the most important when predicting which cluster the participant

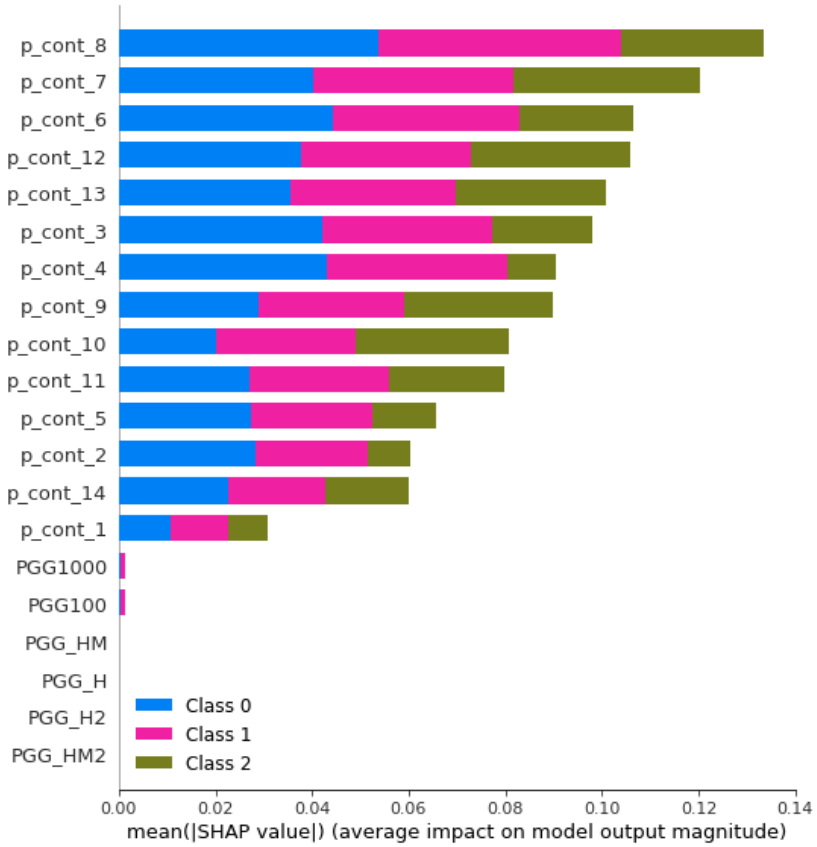

Figure 5.5: Mean absolute value of the SHAP values for each feature and each cluster in PGG_100 and PGG_1000.

should belong. Those are just the middle game rounds, which means that for a game that is lasting 14 rounds, the behaviour of its participants will be better decided by the middle rounds of this game, when no information is provided.

PGG_H and PGG_H2:

We will follow the study analyzing the higher average contributors cluster from the dataset of participants where the information of the distribution of other's people contributions to the pool in the past round is also provided to the participants (100 participants and 1000 participants, PGG_H and PGG_H2). We also run two different supervised machine learning models, as we saw in the previous section and we decided to continue analyzing the results from the *MLPC*, where we obtained better results with better accuracy.

In this case we also applied a force plot and we can see from figure A.4 a very similar behaviour in comparison with the model from the random forest classifier and from the previous dataset analyzed. In this figure we also obtained two different groups of features' contribution to the output value.

We will continue with the plot of the SHAP feature importance for all different clusters and features in order to compare the importance of each round from the neural network classifier to the random forest classifier for each cluster.

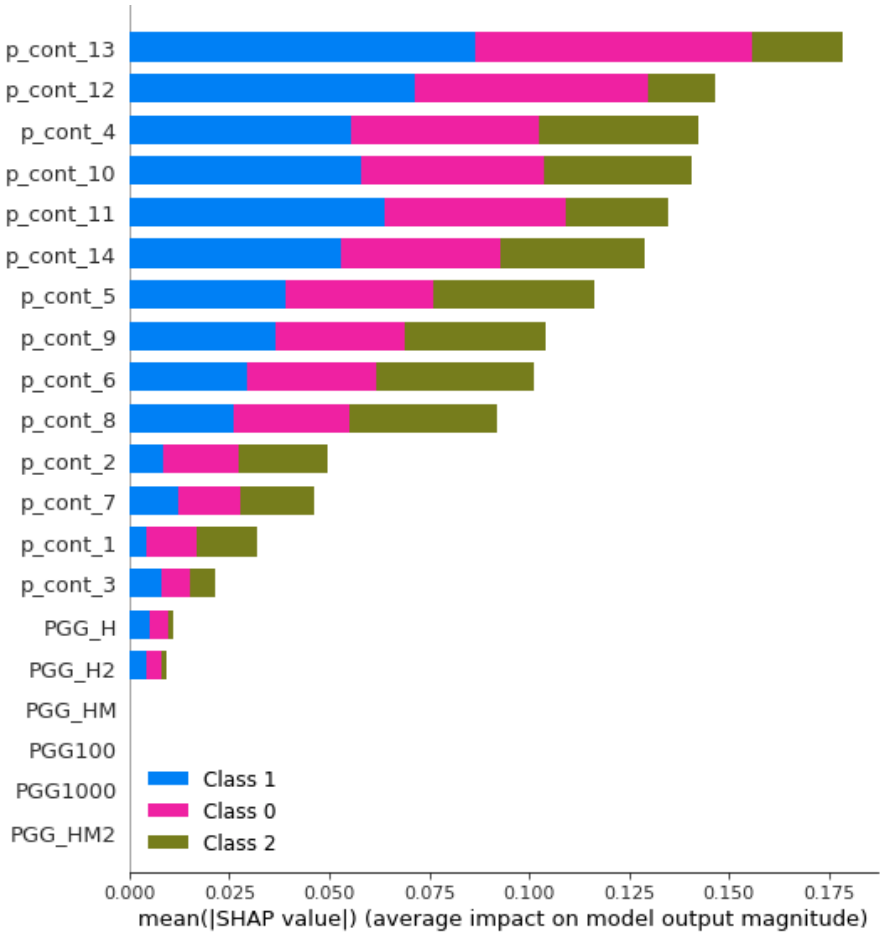

Figure 5.6: Mean absolute value of the SHAP values for each feature and each cluster in PGG_H and PGG_H2.

From figure 5.6 we can see that for this PGG, the distribution of the other people's contributions to the pool in the past round was provided to the participants the rounds where there is a lower distribution of values and the feature importance is higher than the other to determine the cluster's belonging of each participant are clearly the last rounds for the Class 1 (higher average amount contributed). Another interesting result is that for Class 2, the class with lower average amount contributed, the rounds that have more feature importance are the middle rounds. That could mean that this generous behaviour is much more difficult to be assigned to participants from a PGG where the average contributions is also higher than the previous PGG (as we have seen in the previous unsupervised learning section).

PGG_HM and PGG_HM2:

We will finally end the study for this PGG with the final dataset of participants, which is the PGG where the distribution of the other people's contributions to the pool in the past round is provided to the participants along with the their own contribution, their earnings in the past round, their cumulative earnings, and the average group contribution in the past rounds (100 participants and 1000 participants in PGG_HM and PGG_HM2). And we will also be focusing on the cluster with participants with the higher average amount contributed.

In the force plot A.5 we can observe the common two distinguishable groups of individuals regarding the output value where the last participants (red group) are positively contributing to the final prediction from the first base value and in the contrary, the blue participants are negatively contributing to the final prediction value.

In order to finish the study of the effects of the rounds on the predictions we will also create the feature importance plot 5.7 for all the clusters and we will be able to compare it with the results of the previous datasets and the results.

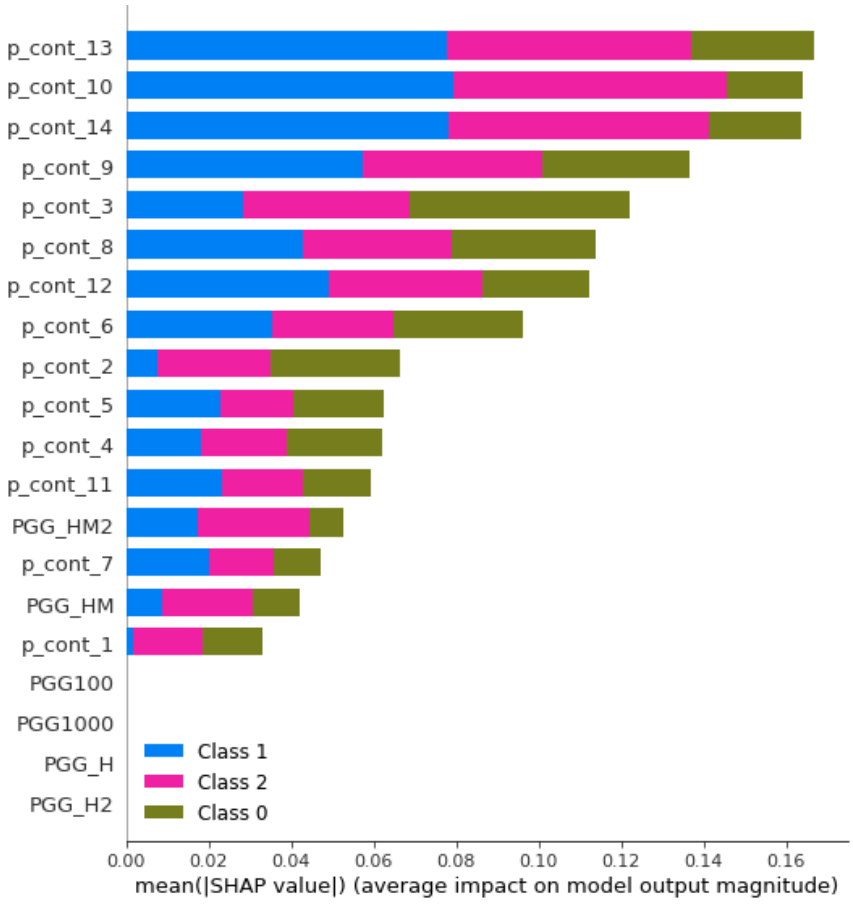

Figure 5.7: Mean absolute value of the SHAP values for each feature and each cluster in PGG_HM and PGG_HM2.

From the previous figure 5.7 we can also see that the rounds that are more influencing when predicting an individual's belonging to class 1 (high average contributors) are clearly the last two rounds. This behaviour is related to the behaviour observed on the previous PGG, PGG_H and PGG_H2, for both supervised models and is also related with the behaviour observed on the previous PGG_HM and PGG_HM2 with the other supervised model. This behaviour, as it has also been sated before, is also related to the increase of the average contribution of the information PGGs in contrast with the lesser-information PGG. Meaning that until the last rounds there cannot be a clear decision on the participants behaviour between higher average contributor and average contributor. We can also see that for the lower average contributors (class 0), the most important features are the first rounds, meaning that if the participant does not contribute generously, the individual will less probably end up being a higher average contributor.

5.2 Resource heterogeneity leads to unjust effort distribution in climate change mitigation

5.2.1 Unsupervised Clustering

In this section we will be using the dataset from the Climate Change mitigation PGG (25). First of all, we will preprocess the final dataset from the previous Climate Change Dataset EDA section (4) in order to be able to get all the information of each participant in a row, treating every row of the dataset as a participant and treating each round contribution as a feature stored in a column. Then, we will focus on the features that are more relevant related to cluster the behaviour of individuals on the PGG (25).

The main goal of this section is to obtain the differences in the change in the behaviour of the players' contribution depending on the features used in our later unsupervised ML model for clustering. We will then start by clustering only for the *individual round contributions*, the *final winnings* and the *level of studies* of each participant. It has to be said that we thought about not including *end capital* and *starting capital*, because they are directly related with the scaled contribution of each participant (calculation made on the previous section).

With figure 5.9 we can see that the second Cluster (Cluster=1) is the cluster with the greatest amount of individuals, followed by the third one (Cluster=2) and the first one (Cluster=0). We will now proceed study the characteristics of this clusters.

In figure 5.9 we can see that the majority of richest participants are the ones that fall into clusters 1 and 2. Those clusters, and we will explore that with more detail in the next sections, are the average contribution participants and the lower average contribution participants.

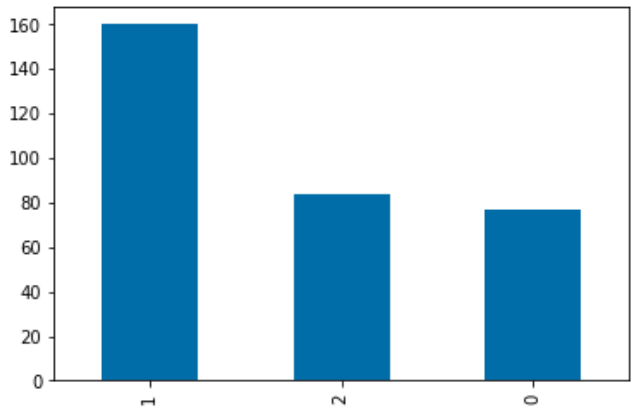

Figure 5.8: # Players in each cluster.

Meaning that as it seems on this heterogeneous PGG, richer participants (with higher initial endowment) tend to continuous rich through the game. We can also see that Cluster nº 2 is the Cluster where the great majority of the participants fall into, being the cluster with participants with average final winnings. The third cluster, on its side, is the cluster with participants with lower average final winnings compared to the average final winnings of all participants. With all that being said, we have clearly identified three different behaviours on participants depending on the winnings feature (as we did with the previous PGG (22)). In this case have the majority of the participants classified in the average final winnings cluster (we will later study how did they play during the game, in order to know why they have such final amount of winnings), followed by participants with less final winnings that could be an indicator of great contributions during the game and we finally have the amount of participants with greater average final winnings, also related with lower contributions during the game. We have again seen the three types of individuals playing together from the evolutionary game theory that we studied in section 2.

We will now analyze all three clusters based on the behaviour of each participant contribution on the different rounds of the PGG and we can see the plots in figures A.6, A.7 and A.8.

From the previous figures (A.6, A.7 and A.8) we can see that the cluster with the greatest amount of participants is the cluster where the participants contributed on average, the average of all participants, followed by the cluster where participants contributed almost, on average, 3 points more than the total average. This results are not in contradiction with the observations that we made earlier where we plotted the clusters depending on winnings' feature (see figure 5.9. Both results agree on that the second more populated cluster is the lower final winnings for individuals and the higher round scaleted contributions for participants, meaning that the lower final winning implies, on average, a higher scaled contribution during the game, a more

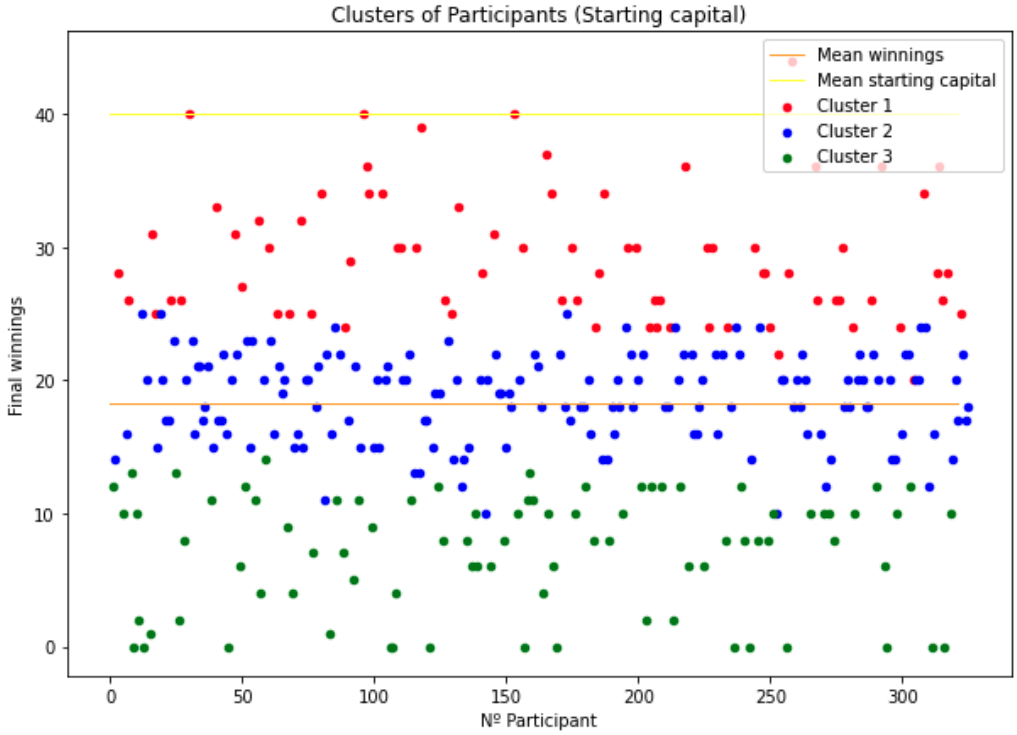

Figure 5.9: Participants plotted against their final winnings with the mean winnings plotted and classified in 3 clusters.

generous participants' behaviour. Later on we will also study, with the help of explanability libraries, once we have trained supervised ML models to predict the group for each participant, the importance of each round and the study level on the process of classifying each participant to each cluster.

Now, we will get into more detail about the importance of feature, each individual round contributions to the behaviour of the participants on each cluster, as well as the final winnings feature related to the starting capital (level of starting wealth) and the study level (level of education) importance. For this study, we will use some supervised ML models and we will recall a game theoretic approach to explain the output of those machine learning models, named SHAP, (SHapley Additive exPlanations).

5.2.2 Supervised Clustering

As we previously said, we will now focus on the individual scaled contributions per round of each participant, the final winnings and their study level to better understand the behaviour of each of participants of each cluster. We will now create our supervised model in order to classify the participants into the three previous clusters and be able to calculate the Shapley Values of each feature using the SHAP library later on.

We will start with a random forest classifier as a supervised learning model. With this model we obtain a 96.22% of accuracy on predicting the output class in the test set (where we applied a 33% of the total population of participants.

We will continue the supervised models with a multi layer perceptron classifier, a class of feed-forward artificial neural network. We will use this supervised model, as well as the previous random forest to classify the participants to each cluster and better understand the importance of the features on that classification. But this time we will also use min-max scaler to avoid the local minima solutions of the model and to avoid higher discrepancies on the results between features due to the magnitude differences between feature's values. With this model we obtain a lower accuracy value than with the random forest classifier. We obtain a 80.19% of accuracy on predicting the output class on the test set. We may want to try to identify if some outliers are present in our predictions and we will use SHAP decision plots for that function.

5.2.3 Explainability

With the level of accuracies reached in the previous section we can now explore how much and how, each feature contributes to the final prediction calculating the shapley values.

We will first start the explanability analysis with the random forest classifier classification of participants and we will also start creating the collective force plot. This plot combines all of the force plots for each individual on the X_test set (on the horizontal axis) where we can see the SHAP values for all features. We will focus on the cluster that has the greatest amount of higher contributors in average in order to better understand what could trigger a participant to contribute more to the common goal (which in this case was their own wealth and the climate change mitigation). For that purpose, we will start with the collective force plot, as previously said. This plot will give us information about the output value, which is the prediction for that participant, the direction on which the features are pushing the prediction to higher output values (red), positive impact on the prediction, or to lower output values (blue), negative impact on the prediction.

In figure 5.10 we can see that for the cluster with greater percentage of higher contributors, we can see that from row 26 to the final, winnings has always a negative impact on the prediction, it is pushing the class label lower, but from the 0th row to the 25th row, winnings converts itself into a feature with positive impact, pushing the class label higher. This means that from participant 1 to the 25th one (in the test set), the winnings are higher than average for those participants. This subset of participants are the participants with higher model output value to be part of cluster nº 3 because they are the players that have lower final winnings and contributed less, but not that less to make the other features more important on the classification step than winnings' feature.

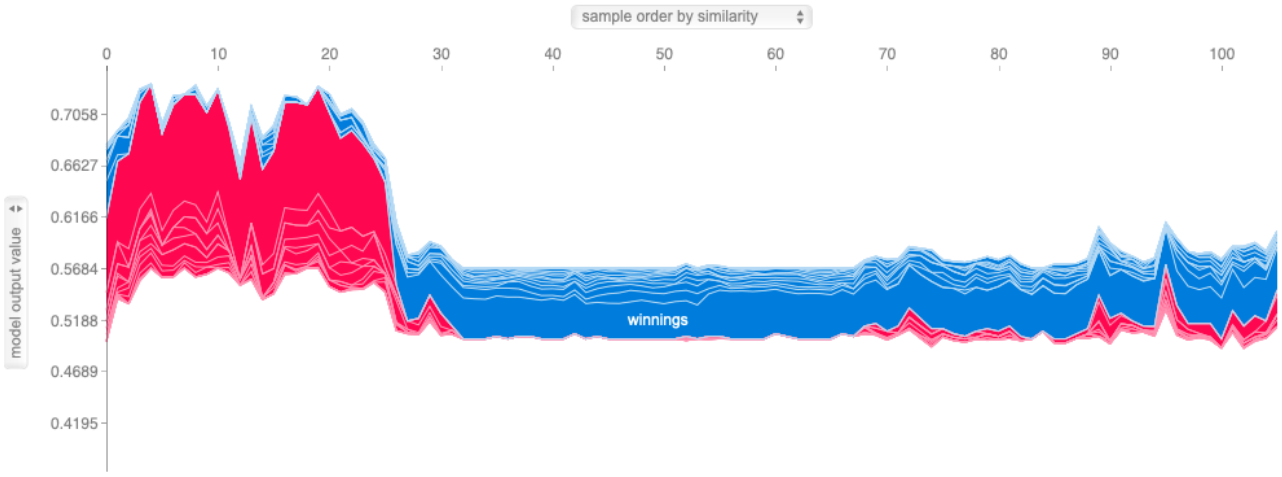

Figure 5.10: Force plot with the Random Forest Classifier supervised model.

Due to the interesting initial 25 participants, we will analyze the SHAP values for all first 25 participants for each feature and deeper explore their base value, which is the value that would have been predicted if no other features where available, meaning that we are going to be able to better understand the difference in importance of each feature.

From figure A.9 we can see that all values are almost falling on the range [-0.1, 0.1] except from winnings' feature. This feature has extreme values for negative impact and positive impact. We can start realizing that this feature will be the most important when talking about feature impact on the model's output, negative and positive. We can also see what we have been saying during the analysis, the most influence feature, and not only for class 1, is the winnings feature. This result could be polarized by the slightly difference on the values of the contributions and the final winnings, but we think that this difference cannot be translated to this huge difference on the average impact on model output value. That's why we are now going to apply a Multi Layer Perceptron Classifier and normalize the data in order to reduce the possible impact of the difference on magnitudes on the values of the features. With that being said, we can still notice that winnings feature has more importance on class 1, the cluster with average contributions, followed by the higher average contributions. For further detail into the rounds relevance into the output values we continue the analysis with the Neural Network Classifier supervised model.

For the MLPC classification, we will try to identify why we obtained a lower accuracy, we will investigate for possible outliers that could make the model more sensible to miss-classifications. In this sense, decision plots from SHAP library can help identify outliers. On these plots (see figure A.10, each observation's prediction is represented by a colored line. At the top of the plot, each line strikes the x-axis at its corresponding observation's predicted value. This value

determines the color of the line on a spectrum. Moving from the bottom of the plot to the top, SHAP values for each feature are added to the model's base value. This shows how each feature contributes to the overall prediction. At the bottom of the plot, the observations converge at expected value. And as we can see in figure A.10 There are no predictions that stand out. There are no prominent effects. We could also use a decision plot to expose a model's typical prediction paths to see what high-scoring predictions have in common.

So, despite the fact that this accuracy value is lower than the accuracy in the other supervised ML model, it is still more than a 80% of accuracy on the test set. So, after the analysis, with no outliers spotted, we consider that it still is a good accuracy value to continue with the model-agnostic method to explain the output and obtain reasonable conclusions.

We will then start with our analysis with the collective force plot on the cluster with higher contributors on average, Cluster 3.

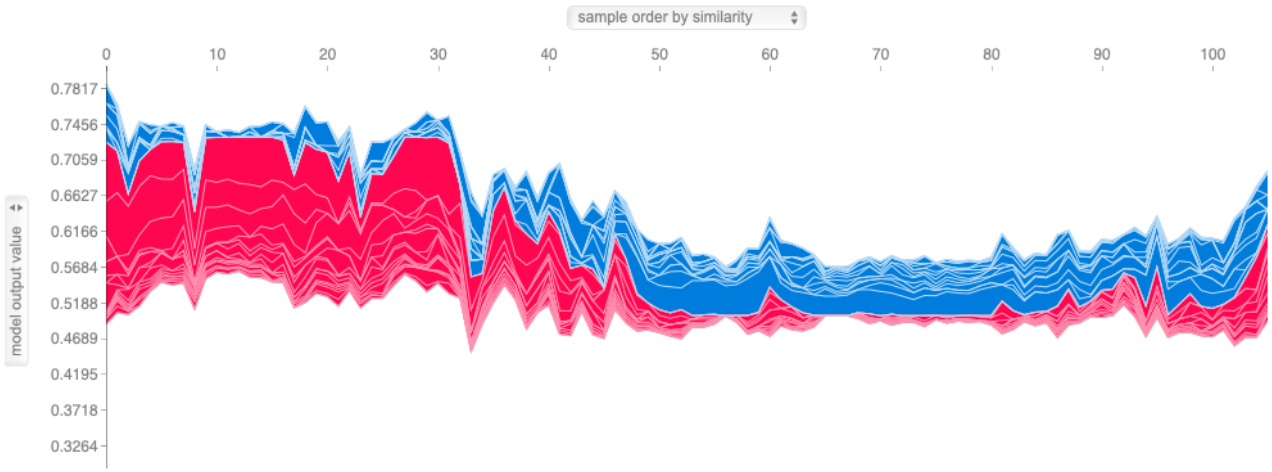

Figure 5.11: Force plot with the MLPC supervised model.

With figure 5.11 we can see that in comparison with the Random Forest Classifier, the average positive output value has increased as well as the importance of the other features. Some interesting results may be obtained with this supervised model.

We will follow the analysis with the SHAP values plot for all the features and its importance for the first 30 participants.

With figure 5.12 we can see now that, as we were previously saying, winnings feature has reduced its difference on the impact on the model output. We can see that it still has the same negative effect when it has a positive SHAP value and positive impact with negative SHAP values. Furthermore, we can see that the last rounds have a big impact on the model output as well (positive impact with positive SHAP values). and that r6, the level of studies has also a very big impact on the model output and on the same direction that winnings has. When

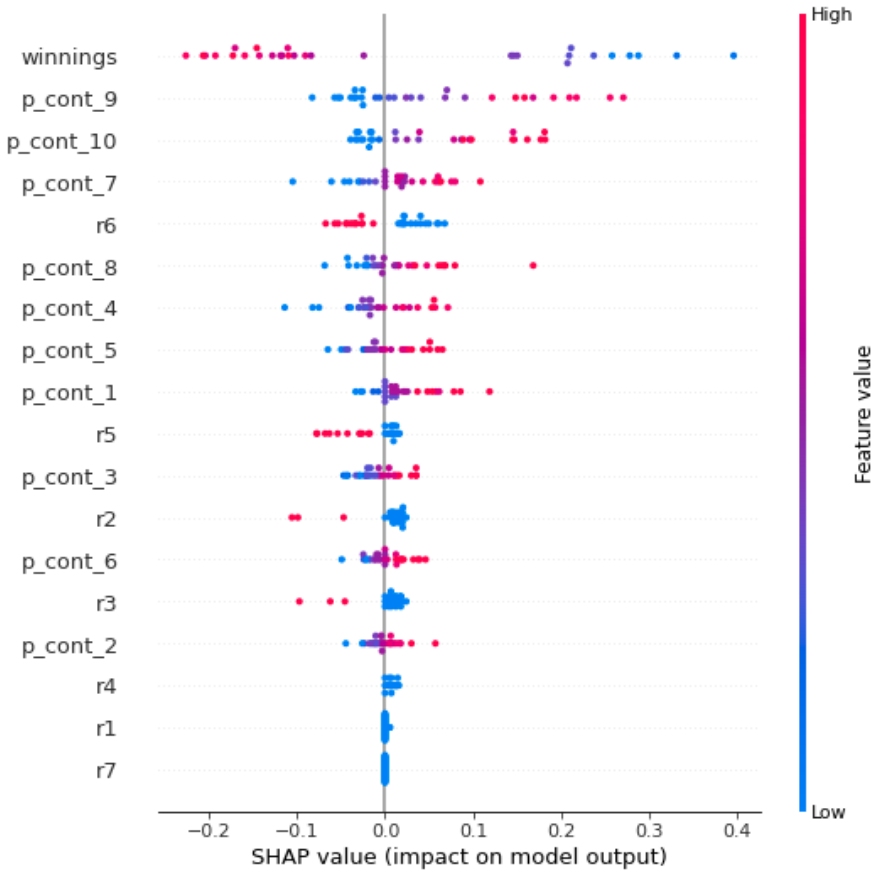

Figure 5.12: Mean absolute value of the SHAP values for the first 30 rows (for cluster=2) with the MLPC supervised model.

participants have a level of studies it seems that the influence is negative towards the model output, and the belonging of that participant into the higher average contributors cluster.

And if we finally apply the SHAP feature importance plot for all three clusters (see figure 5.13) we may see that that the impact of the features has been re-scaled and the average impact on the model output can be now studied with more precision than with the previous Random Forest Classifier model. We can see that with this supervised model, and as we have been identifying on the previous plots, the most influent feature is still winnings. And the cluster in which this feature has more impact is in cluster 3 (class 2), the higher average contributors. This cluster is also importantly affected by the last two rounds' contribution and the level of studies. If we now recall the previous section's plot Distribution of players from clusters for each starting capital we can also see that the contributions that are more influent in this cluster are the contributions of participants that had a fewer amount of starting capital (less starting level of wealth). Referring to the education level now, we can observe now that it has a great average impact on the model output and in previous plots we have seen that this impact is

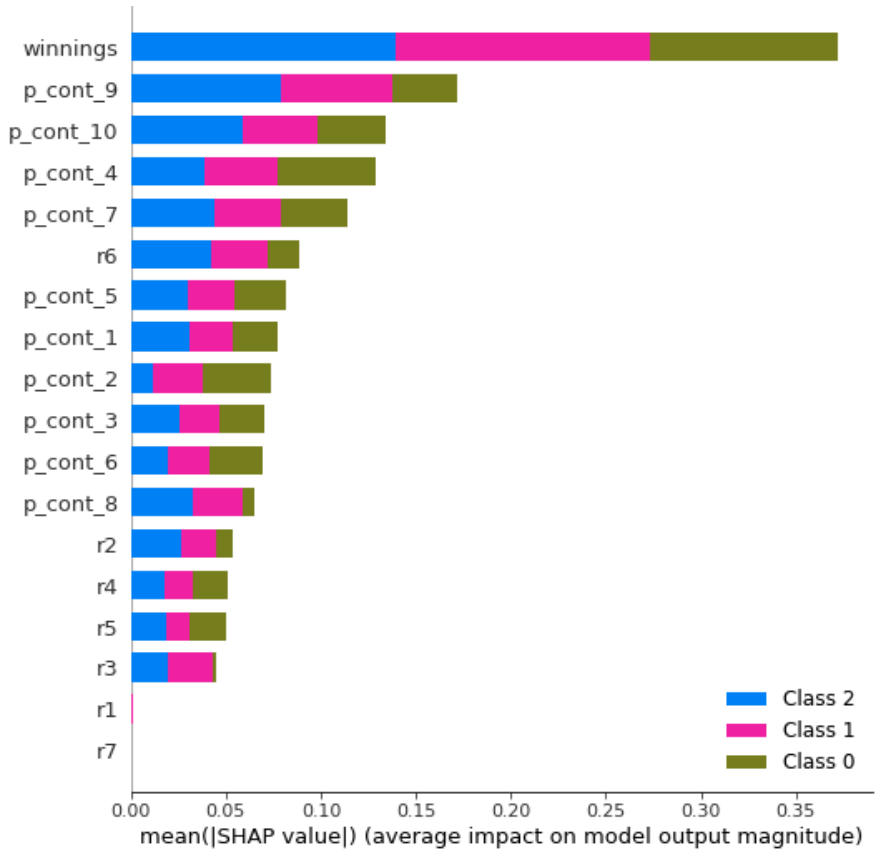

Figure 5.13: Mean absolute value of the SHAP values for each feature in each cluster.

negative. The negativeness of this impact means that subjects with lower education level were predicted to make higher contributions in equal conditions.

Chapter 6

Conclusions

6.0.1 Large scale and information effects on cooperation in public good games

The main goals of the original paper (22) were to study PGGs with larger scales than usual (beacause it is reported that all the previous studies were made on small groups of people, largest reported is 100) and study the effects of giving different type of information about the other participants on each round to each participant (three types of information treatment). They studied and analyzed the contributions of the participants on the PGGs on each of the group sizes (100 and 1000 participants) and found that the average levels of cooperation of these two treatments are indistinguishable, suggesting that a group sizes of 100 individuals was a good representative of large groups with differences in contributions being not statistically significant. This result is in agreement with the results earlier found on our first part of the EDA section for the original's paper dataset. Regarding the information effects, they found that they have a drastic effect on subjects' contributions. Specifically they found that as much information provided to players on how people contributed on each round had been a promoter of cooperation. They also used an unsupervised learning algorithm to cluster the participants' contributions into types of behaviour and they found three types of individual decisions.

Our results are in total agreement with theirs because we also showed that the average contribution of participants was higher in the treatments where more information was provided to the players. We also showed that the two last rounds of the PGG were the most influent when deciding the participants' belonging to the average or the above average contributors. Regarding the lower average contributor, we concluded that their non generous behaviour started on the first rounds for the PGG100 and PGG1000 and the middle rounds for the PGG_H, PGG_H2, PGG_HM and PGG_HM2, and consequently, those rounds were the most influents when assigning the participants to the non-generous behavioural cluster (or free-riders cluster).

The difference observed in rounds' importance might be explained with the introduction of the information effects. This effects produces a higher contribution in all participants and it can specifically affect and delay the assigning of those individuals to the non-generous cluster.

Furthermore, this result can also be observed on the conclusions on Fraser, S., Nettle, D. (23). In this paper, that we also studied for the PGG approach and constructed an Exploratory Data Analysis, they stated that for the PGG experiment there is a greater generosity of hungry participants and the punishment treatment, the results were highly, not marginally, significant. We also showed that for the punishment treatment the average contributions of all participants were significantly higher than in the no punishment treatment. This behaviour can be somehow related to the results found on our dataset of the Large scale and information effects where we had three different treatments related to the information given to each participant on each round, which could also be seen as a possible spread of reputational information, through gossip, and consequently whom might fear being gossiped about). Also possibly related to this behaviour we find Dannenberg, A, et, Al., 2019. (24) where a public goods experiment is used to study whether groups choose to implement an institution that allows for the exclusion of members and how this relates to individuals contributions. In this paper they found that exclusion institutions (B10 and B8, for game B) increases contributions to the public goods. Subjects who had been excluded or received a vote for exclusion adjusted their contributions closer to the group average in later rounds and subjects who vote for the exclusion institution contribute significantly more than those who vote against it. We also found similar results on our previous EDA about the endogenous treatment (treatment with slightly higher contributions) indistinguishable on which treatment (exclusion with or without cost) we were focusing, the support for the institution increased over time. We also observed that average contributions were higher when the exclusion option was available and participants who voted in favor of the exclusion institution contributed more than those who voted against it.

Other related studies (33), (34) also show that the establishment of this fear factors may enhance cooperation in repeated public-goods games and that availability of costly peer sanctioning can have a large positive impact on cooperation in social dilemmas, but defining more participants that are able to punish, may be more efficient enhancing cooperation to common-pool resources. However, it is also demonstrated that the beneficial role of punishment possibilities for cooperation success is highly fragile, thus the correct and precise decisions must be made when punishing or ostracising group members.

All these papers and specially the papers (22), (23), (24) whose experimental PGGs we have studied during this work are related somehow through a kind of fear factor that induces to cooperation. So, this fear factor in our main studied PGG (22) may be related to the fear that some sort of reputational information might spread through gossip and therefore this is

what can induce non-cooperators and all participants to abandon or reduce egoistic behaviours and cooperate to the common-pool (33).

6.0.2 Resource heterogeneity leads to unjust effort distribution in climate change mitigation

The original paper's (25) goal was to better understand the effect of resource inequality when diverse actors interact together towards a common goal. Their results showed, as we previously said, that the effort distribution was highly inequitable, with participants with fewer resources contributing significantly more to the public goods than the richer and that with an unsupervised learning algorithm they classified the subjects according to their individual behavior, finding the poorest participants within two "generous clusters" and the richest into a "greedy cluster". Our results are strictly in accordance. We also found 3 clusters (generous, slightly above average/average and selfish contributors) using an unsupervised learning algorithm and we also stated that richer participants were classified in the greedy cluster with the higher educated participants. Our results in the Model Explainability also showed that the rounds with more influence for classifying the participants on each cluster were the last two of the collective-risk dilemma. That is because the participants that contributed above average are classified into two different groups of generous behaviour having only a difference of 1.5 points in contribution and those last two rounds are the rounds where the difference between both clusters will be critical in order to classify the participant into his/her belonging cluster due to their contributions. The participants from the selfish cluster are classified with their contributions on their first to middle rounds, meaning that pure selfish participants are more likely to stay behaving that way.

The final effects and conclusions stated on the paper Climate, related to the features that we have studied are that (on heterogeneous treatment) disadvantaged individuals contributed much more than a fair share of the mitigation, that the richer ones contributed less. It appeared that, contrary to the expectations of the poor exploiting the rich in a public goods context, here we found the opposite situation. This result can also be observed on the conclusions of another paper, (23). In this paper, that we also studied for the PGG approach and constructed an Exploratory Data Analysis, they stated that for the PGG experiment there is a greater generosity of the no-breakfast participants in the no-punishment treatment (we will only analyze the no-punishment treatment because our dataset of the Climate Change Mitigation did not have any kind of punishment/exclusion treatments). So, if the effects of hunger on cooperation in richer, repeated interactions that we observed in experiment 2 prove robust, this could have real-world implications for social inequalities in cooperation and anti-social behaviour.

With all that being said, it has been stated that on our experimental PGGs with heterogeneous initial conditions related to basic needs, such as hunger, wealth and education the most disadvantaged individuals have been the most generous when contributing to the common wealth, even if the goal was related to climate change mitigation actions or if it was purely focused on cooperation.

This behaviour observed with the contributions with the PGGs can be thought to be related somehow to the concept of reciprocal altruism (28), (29). This concept was first introduced by the American evolutionary biologist and sociobiologist Robert Ludlow "Bob" Trivers in the field of evolutionary biology in 1971. But this concept and the kind of relationships that it contains are exaclty analogous to certain type of PGGs. The benefits and the power of this theory that introduces the concept of reciprocal altruism were dramatically demonstrated by a pair of tournaments held by the American political scientist, Robert Axelrod around 1980. He applied the Game Theory to study the mechanism of reciprocal altruism (30). In this paper, Axelrod and Hamilton (an English evolutionary biologist), revealed that reciprocating the assistance from another individual is stable in evolution as long as there are enough altruists in the population. One of the most popular explanations of why we tend to observe this behaviour of reciprocal altruism on PGGs (31) is based on the systematic strategy based on the principle of reciprocal altruism, TIT-FOR-TAT (32). This is based on the observation that the participants in a PGG tend to make a cooperative act itself -or a reputation for being a cooperative person- may with high probability be reciprocated with cooperation, to the ultimate benefit of the cooperator (31). Consequently, the disadvantaged individuals could feel to be more cooperative than the wealthier participants because there would be a higher possibility of a future reciprocity from others.

6.0.3 Possible future work and limitations

In our work we have introduced two main aspects. The first one is the mix of machine learning models (with unsupervised and supervised models) to study and analyze data from Public Goods Games. This models allowed us to classify the participants on those games depending on their behaviour (contributions to the common fund). Then, we also have been able to interpret and explain these classifications of the participants on each group of common behaviour, studying which features were contributing more or less to the output value of the supervised machine learning model. The other interesting thing was the applicability of a game theoretical concept, as shapley values, to analyze these feature importance in the outputs of the different models.

In this sense, we propose to extend this application of model-agnostic methods applying algorithms to explain the outputs of the machine learning models used to classify participants

depending on much more features. Where the final goal is to understand which features better influence higher or lower contributions to the common goal based on a wider range of initial conditions and features, not only the features that we had analyzed during our work.

On the other hand, apart from possible future lines of work, one of the biggest issues with game theory is that, like most other economic models, it relies on the assumption that people are rational participants that are self-interested and utility/payoff-maximizing and in reality, in most of the experimental results obtained people does not act like rational participants at all. Of course, we are social beings who do cooperate and do care about the welfare of others, often at our own expense. Game theory (see section 2) cannot account for the fact that we may not always fall into Nash equilibrium situations. It highly depends on social context of each individual and these features are the most relevant to future study because they will be, in the majority of the situations the trigger for the higher or lower contribution of each individual to the common goal. In this sense, if we now recall the initial declaration, we can see a direct relationship between this issue from game theory and the global issue with science and society and how it becomes a greater issue when science tries to understand and explain society and how society behaves.

Bibliography

[1] C. F. Camerer, 2011. *Behavioral Game Theory: Experiments in Strategic Interaction.* January 2011. Russell Sage Foundation

[2] N. Harald, 2015. *Behavioral Game Theory.* ETH Zürich.

[3] Bounau, 2017. *Journal of Game Theory 2017, 6(1): 7-14 A Case for Behavioural Game Theory.* doi:10.5923/j.jgt.20170601.02

[4] F. Nash, 2008. *The Agencies Method For Modeling Coalitions And Cooperation In Games.* doi.org/10.1073/pnas.1216361109

[5] Plonsky, et, Al., 2018. *Predicting human decisions with behavioral theories and machine learning.* arXiv:1904.06866

[6] H. Tembine, 2015. *Deep Learning Meets Game Theory.* doi:10.1109/TCYB.2018.2886238

[7] Hartford et. Al., 2016. *Deep Learning for Predicting Human Strategic Behavior.* doi:10.14288/1.0319323

[8] Schuurmans, Zinkevich, 2016. *Deep Learning Games.* doi:10.1109/TG.2019.2896986

[9] Cotla, 2015. *Learning in Repeated Public Goods Games - A Meta Analysis.* doi: 10.1371/journal.pone.0016836

[10] Poncela-Casasnovas, et, Al., 2016. *Humans display a reduced set of consistent behavioral phenotypes in dyadic games.* doi:10.1126/sciadv.1600451

[11] Fallucchi, et, Al., 2018. *Identifying discrete behavioural types: a re-analysis of public goods game contributions by hierarchical clustering.* doi:10.1007/s4088

[12] Sikver, et, Al., 2018. *Mastering the game of Go with deep neural networks and tree search.* doi:10.1038/nature16961

[13] J. Goodfellow, et, Al., 2014 *Generative Adversarial Nets.* arXiv:1406.2661

[14] Ali el Hassouni1, et, Al., 2018. *Using Generative Adversarial Networks to Develop a Realistic Human Behavior Simulator.* doi:10.1007/978-3-030-03098-8_32

[15] Turek, 2018. *Defense Advanced Research Projects Agency, DARPA Explainable Artificial Intelligence (XAI)*

[16] Arrieta, et, Al., 2019. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.* doi:10.1016/j.inffus.2019.12.012

[17] Cristoph Molnar, 2020. *A Guide for Making Black Box Models Explainable Interpretable Machine Learning.*

[18] Lundberg, Lee, 2017. *A Unified Approach to Interpreting Model Predictions.* doi:10.1007/BF025512743

[19] Ribeiro, et, Al., 2016. *"Why Should I Trust You?" Explaining the Predictions of Any Classifier.* doi:10.1145/2939672.2939778

[20] Georgios Sarantitis, 2018. *-gsarantitis.wordpres.com- Interpretable Machine Learning with SHAP*

[21] PwC - UK, 2018. *Explainable AI - Driving business value through greater understanding*

[22] Pereda, M.,et al. *Large scale and information effects on cooperation in public good games.* Sci Rep 9, 1502. October 21. doi:10.1038/s41598-019-50964-w

[23] Fraser, S., Nettle, D. 2019. *Hunger affects social decisions in a Public Goods Game but not an Ultimatum Game* PsyArXiv. October 2. doi:10.31234/osf.io/67abq.

[24] Dannenberg, A, et, Al., 2019 *Voting on the threat of exclusion in a public goods experiment.* Exp Econ 23, 84–109. April 9. doi:10.1007/s10683-019-09609-y

[25] Vicens, J., et, Al., 2019. *Resource heterogeneity leads to unjust effort distribution in climate change mitigation* 13(10): e0204369. October 31. doi:10.1371/journal.pone.0204369

[26] Allen, J. 2018. *The public goods game on multiplex networks.* Doctoral thesis, University of Surrey.

[27] Pereda, M., et, Al., 2019. *Group size effects and critical mass in public goods games.* Sci Rep 9, 5503. April 2. doi:10.1038/s41598-019-41988-3

[28] Trivers, R.L., 1971. *The evolution of reciprocal altruism. Quarterly Review of Biology.* 46: 35–57. doi:10.1086/406755

[29] Pizarro D., et, Al., 2003. *The evolution of Reciprocal Altruism.* doi:10.4135/9781412956253.n437

[30] Axelrod, R., Hamilton, W. D., 1981. *The evolution of cooperation.* Science, 211, pp. 1390-1396.

[31] Dawes R., Thaler R., 1988. *The Journal of Economic Perspectives.* Vol. 2, No. 3.

[32] Rapoport A., 1974. *Game Theory as a Theory of Conflict Resolution.* doi:10.1007/978-94-010-2161-6.

[33] Feinberg, M. et, Al., 2014. *Gossip and ostracism promote cooperation in groups.* doi:10.1177/0956797613510184.

[34] Leibbrandt, A. et, Al., 2015. *Incomplete punishment networks in public goods games: experimental evidence.* doi:101177/0956797613510184.

[35] Guererk, O. et, Al., 2010. *The Effects of Punishment in Dynamic Public-Good Games.* doi:10.2139/ssrn.1589362.

Appendix A

Appendix

A.1 Large scale and information effects on cooperation in public goods games

A.1.1 Unsupervised Clustering

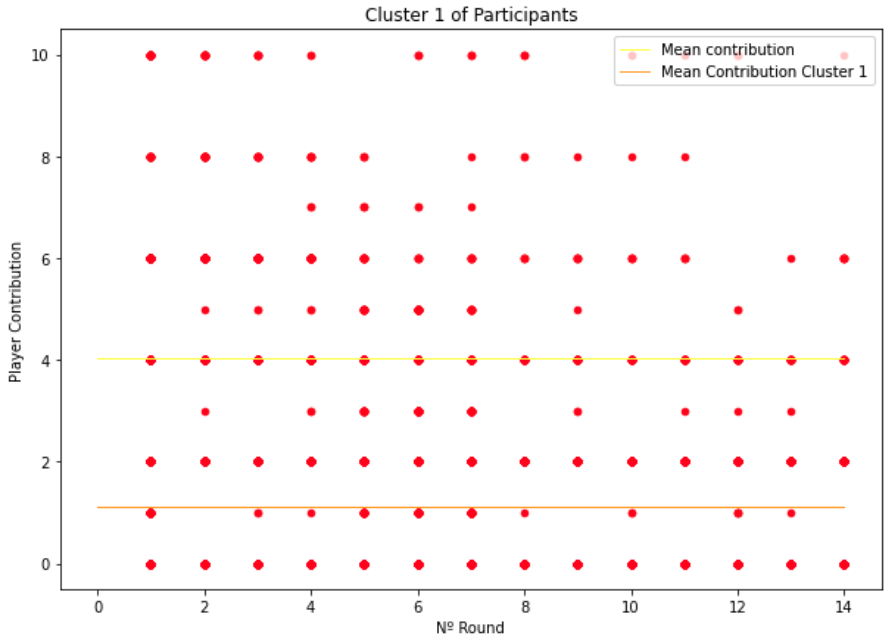

Figure A.1: Player contribution per round in Cluster 1. Frequency is plotted with bigger representative dots.

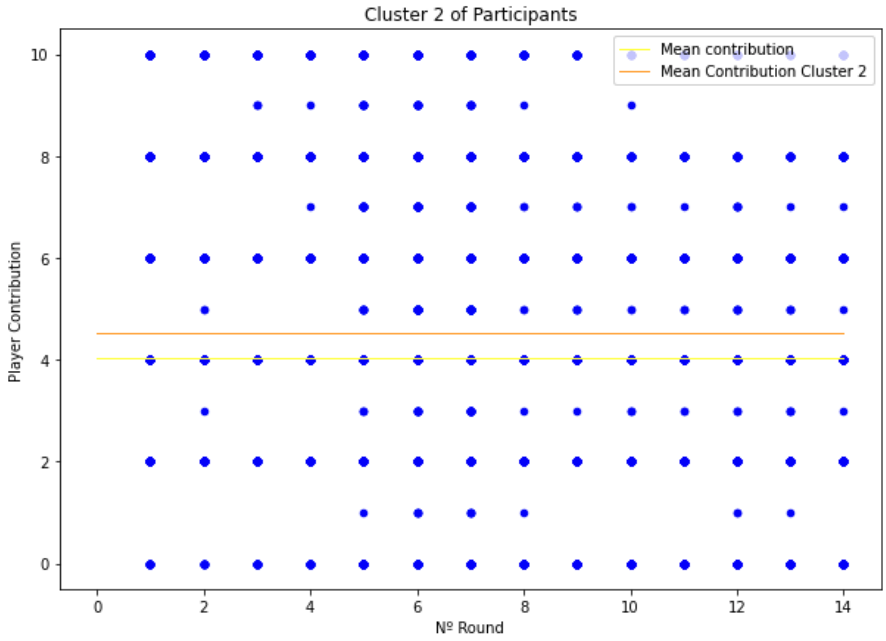

Figure A.2: Player contribution per round in Cluster 2. Frequency is plotted with bigger representative dots.

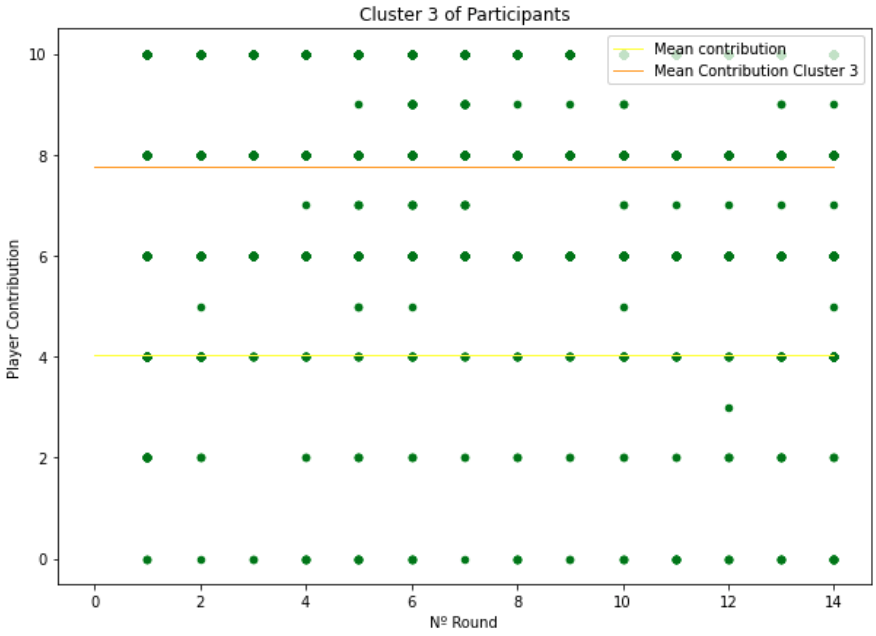

Figure A.3: Player contribution per round in Cluster 3. Frequency is plotted with bigger representative dots.

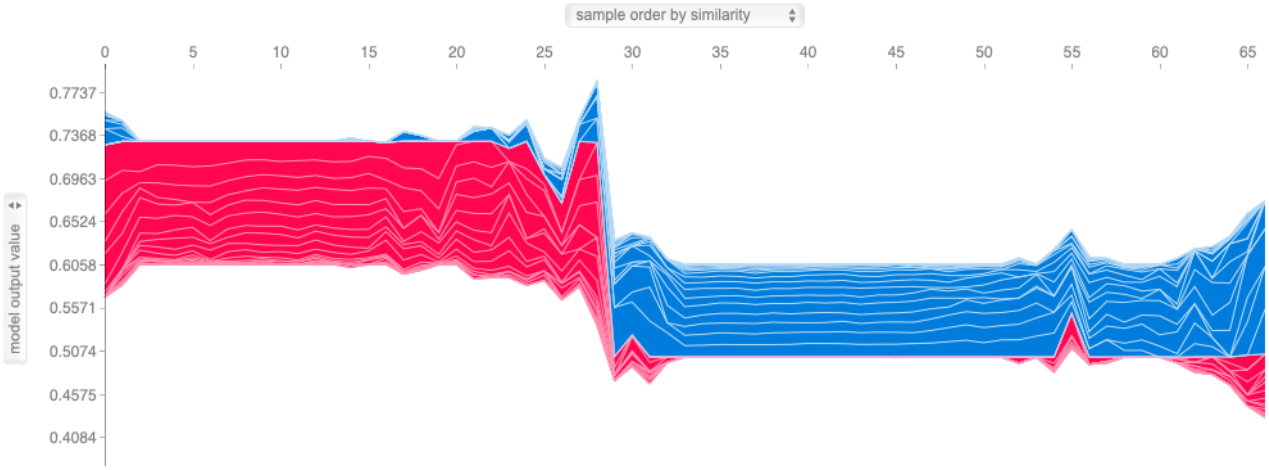

Figure A.4: Force plot with MLPC supervised model for PGG_H and PGG_H.

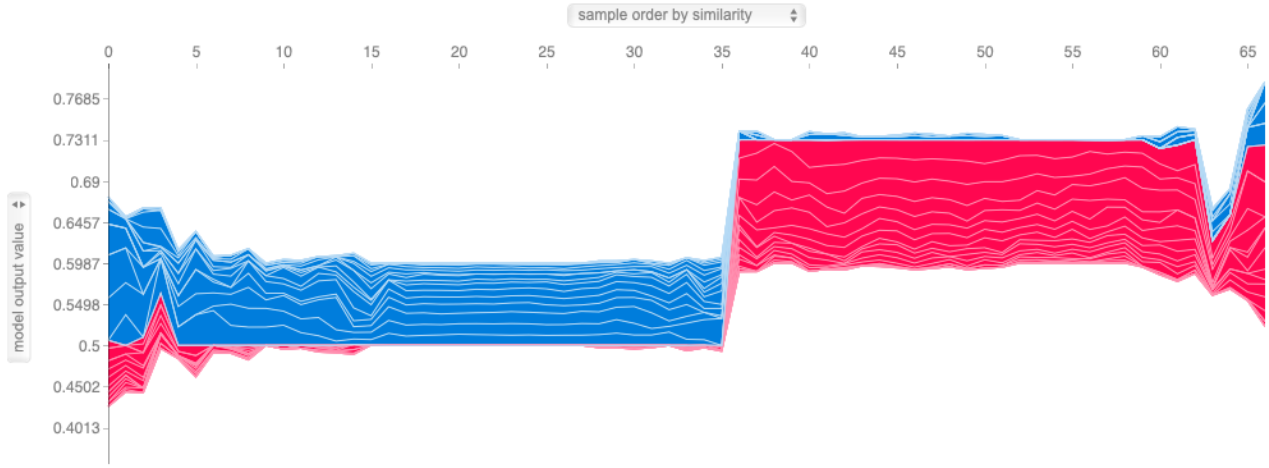

Figure A.5: Force plot with MLPC supervised model for PGG_HM and PGG_HM2.

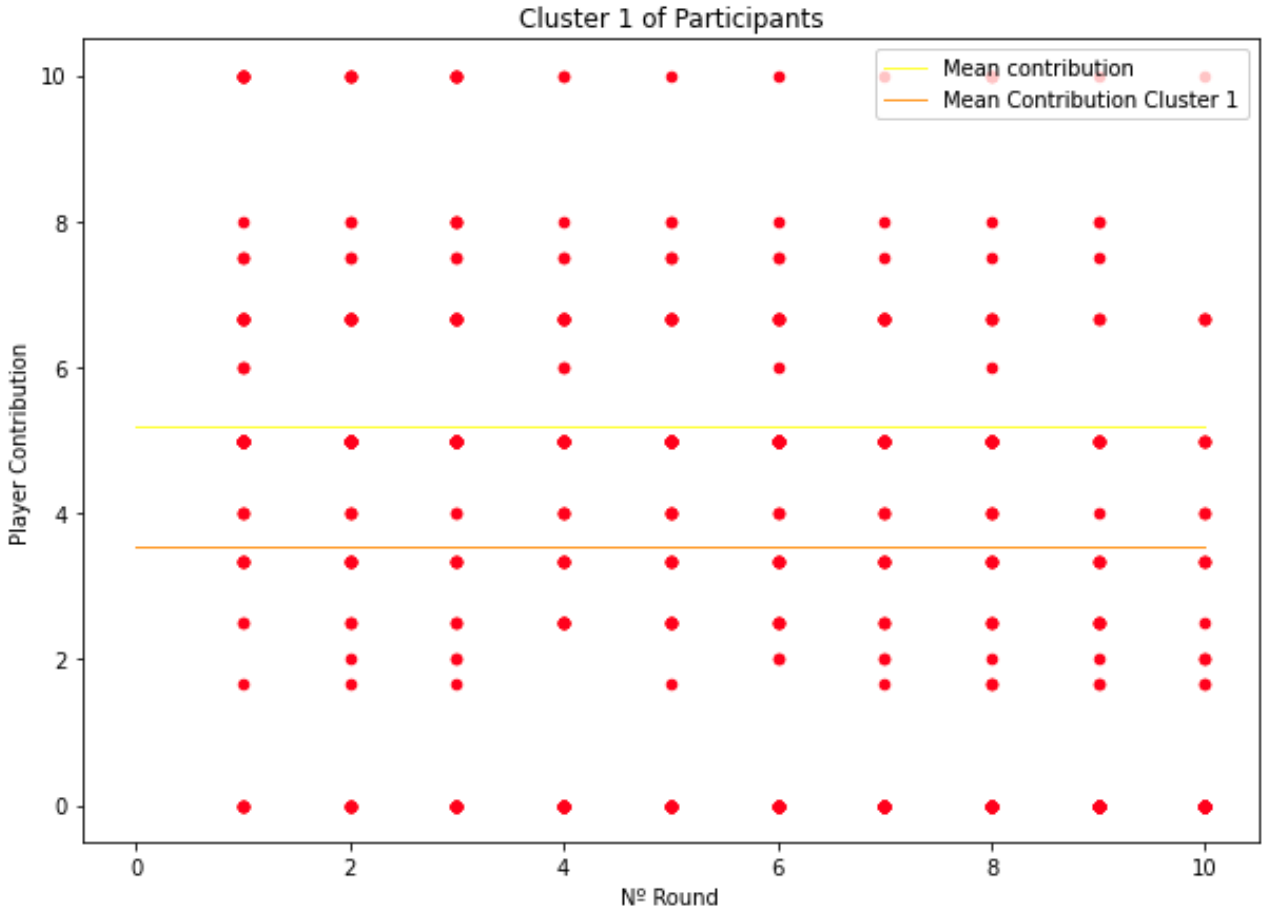

Figure A.6: Force plot with MLPC supervised model for PGG_HM and PGG_HM2.

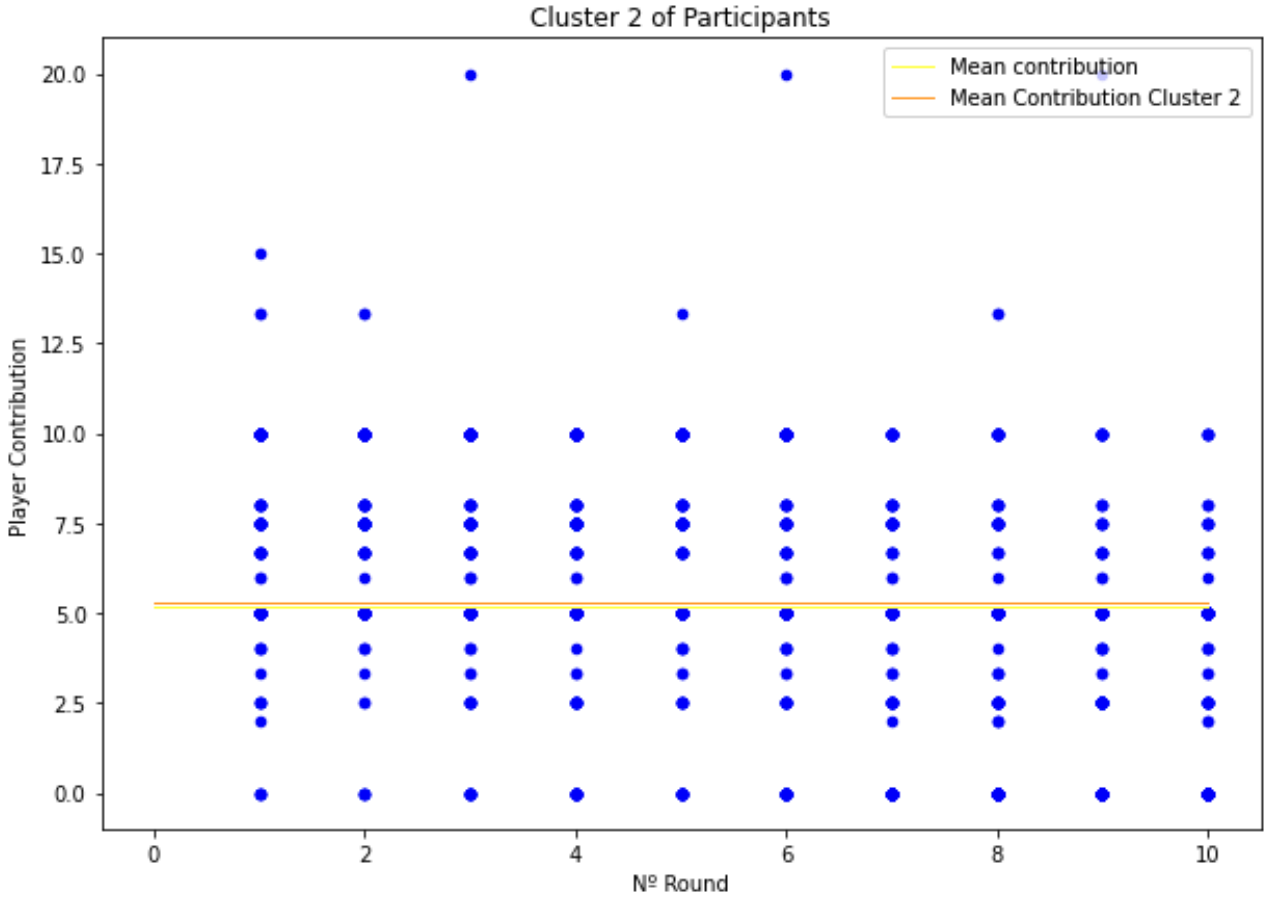

Figure A.7: Force plot with MLPC supervised model for PGG_HM and PGG_HM2.

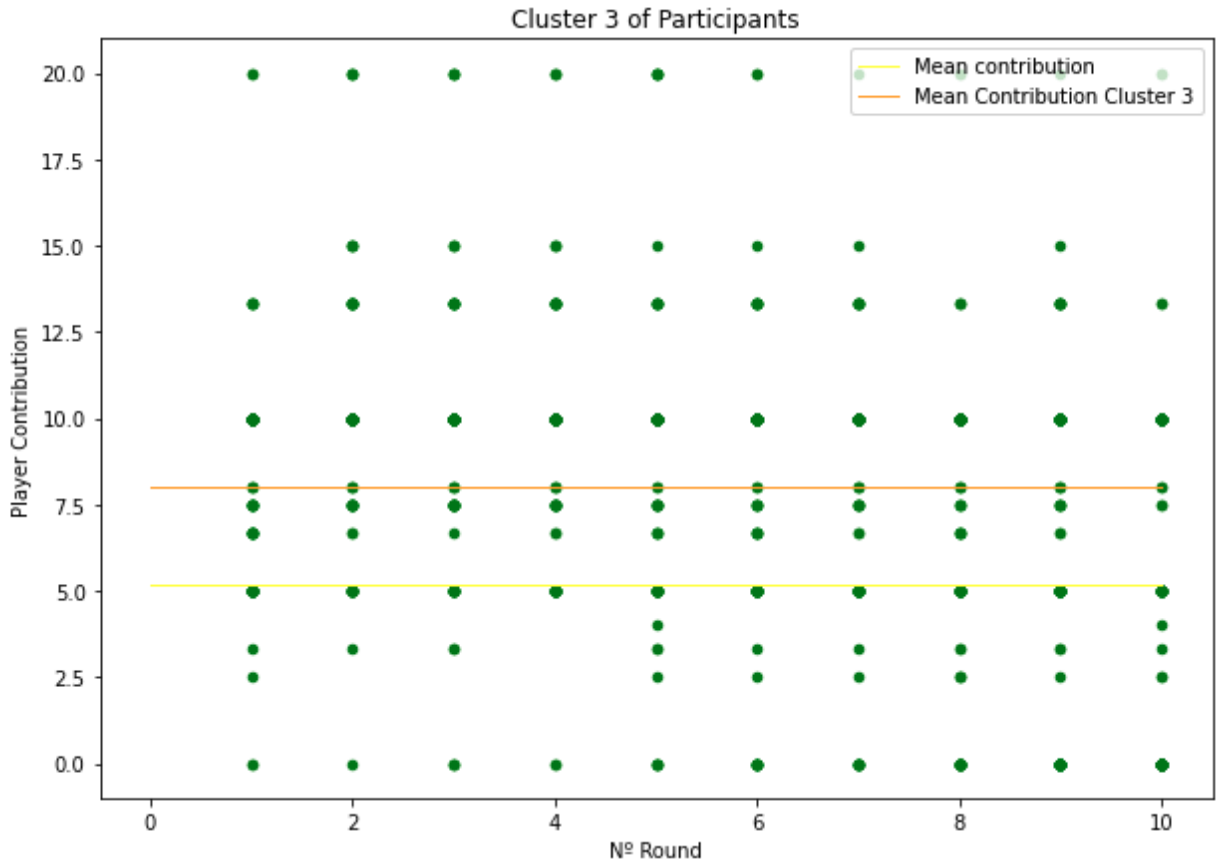

Figure A.8: Force plot with MLPC supervised model for PGG_HM and PGG_HM2.

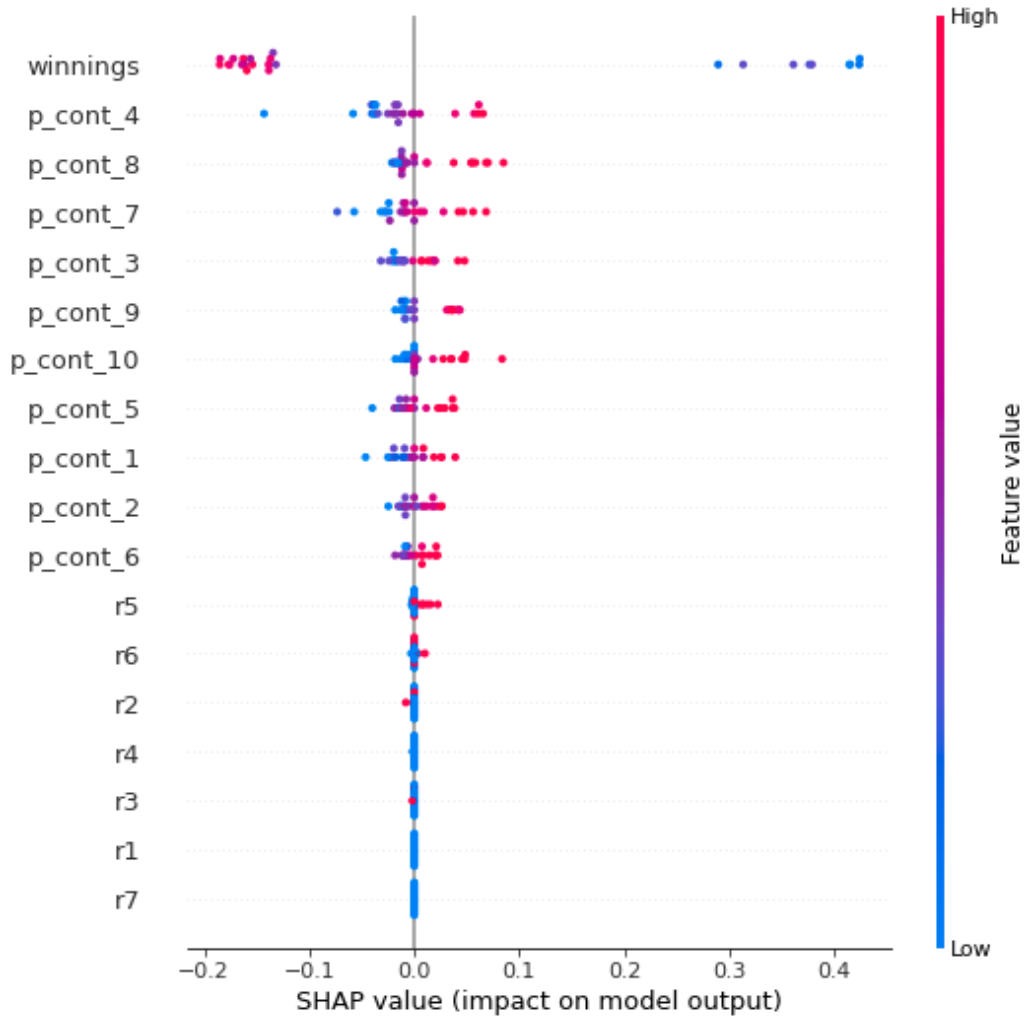

Figure A.9: Mean absolute value of SHAP values for the first 25 participants.

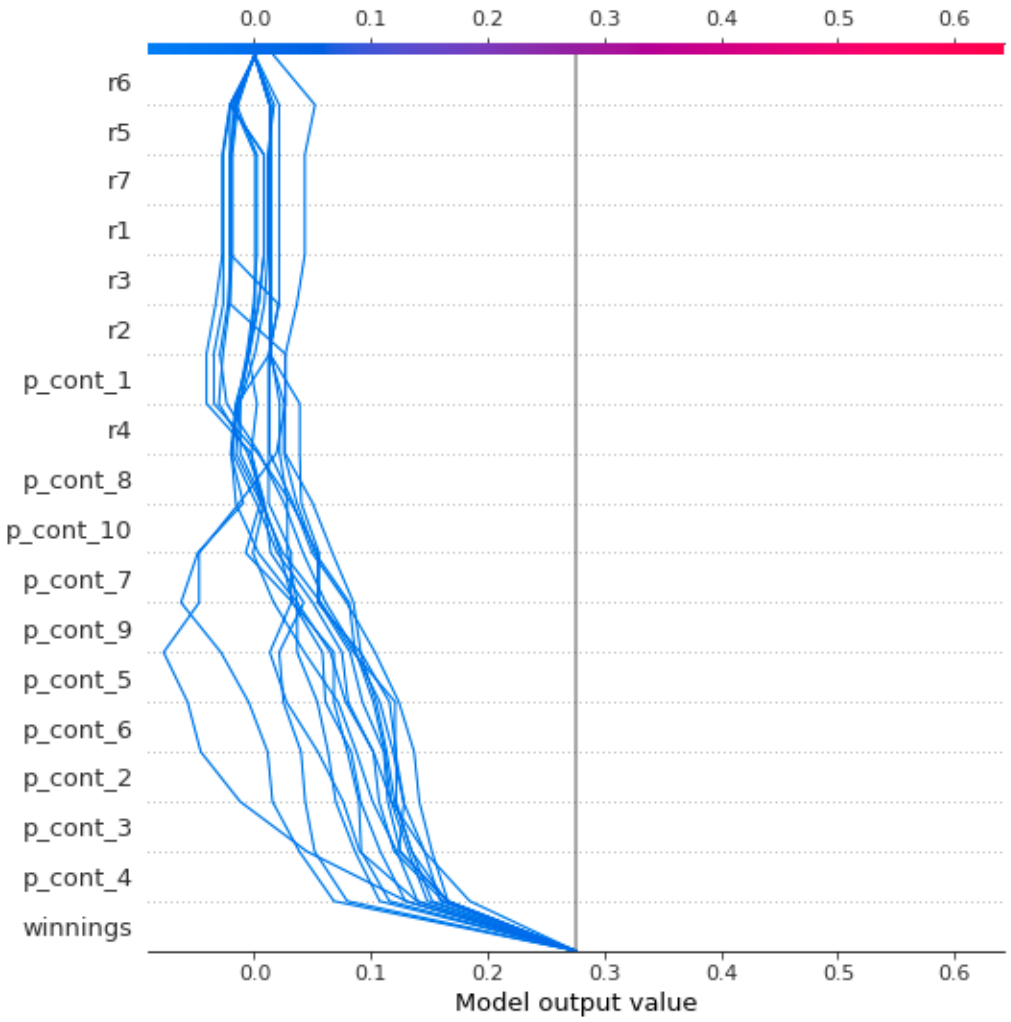

Figure A.10: Desicion plot in the probability range [0, 0.1] to correclty classify Cluster 3 (cluster=2).