Chapter 3

Exploratory Data Analysis

The following section illustrates the different steps undertaken to obtain a prediction model for pro-environmental behaviour.

In particular, the results of unsupervised machine learning algorithms are shown. Correlational Class Analysis and Partitioning Around Medoids Clustering are performed and the created classes are inserted in the final prediction models.

Performing these two algorithms I have the final dataset, and therefore I can continue with the second step: Exploratory Data Analysis. This analysis concerns a quick description of the complete dataset’s explanatory variables and then of the two distinct subsets according to the climate change risk perception level. As I have already mentioned in the Methodology (Chapter 2), in the first place the prediction is fitted on the entire dataset, and then it is computed separately on two subsets: with the individuals that have a high-risk perception (qb2 greater or equal to 6) and with the individuals that have a low-risk perception (qb2 less than or equal to 5).

The last section of the chapter shows some characteristics at the macro-level. I have discussed that some of the factors that shape pro-environmental behavior are regulations or habits of countries. For this reason, some descriptive analyses grouped per country are done.

**3.1 Unsupervised Machine Learning Algorithms: finding new patterns among citizens**

**Correlational Class Analysis**



Figure 8: CCA

CCA partitions the dataset into classes depending on the similarity of correlation patterns to apprehend unobserved heterogeneity (Rossoni et al., 2021). The algorithm is performed through the *corclass* package in the R software using climate change items (qb4\_3, qb4\_5, qb7, qb8, qb9). CCA has divided the sample into five correlational classes: group 1 includes 5704 cases (26%), group 2 4415 (20%), group 3 6106 (28%), group 4 3271 (15%), group 5 2482 (11%). As previously stated, CCA aggregates individuals who agree about the cardinal arguments, although they do not coincide with the answers given or they do not concur with each other (Rossoni et al., 2020). Classes combine individuals who perform similar decision patterns between each pair of variables. Figure 5 illustrates the individual modules or classes created as a network. Each node coincides with one item, while the edges reveal the statistically significant correlation between variables (Rossoni et al., 2021). The more line is thicker, the more significant is the correlation. [[1]](#footnote-1)

In group 1 the correlations between all pairs of variables are set to 1. Actually, group 1 is called “zero class”, due to these rows get 0 variance and to default, R sets the correlations between all pairs of 0 variance rows to 1. Therefore, there is no sharing cultural meaning in this group.

Group 2 obtains a high positive correlation (0.8) mainly between qb4\_3 and qb4\_5. These two questions concern the EU's role in climate change. Group 3 has a slight correlation among all variables, except for qb4\_5. In group 4 I find a strong positive correlation among qb7, qb8, and qb9 and separately between qb4\_3 and qb4\_5. The first block of variables mainly concerns the importance of the government decisions to fight climate change while the second block of variables, as we have already seen, concerns Europe’s role. Lastly, group 5 has a strong correlation between the following pair of variables: qb4\_5 and qb7; qb4\_5 and qb8; qb4\_5 and qb9. In this case, qb4\_3 is completely isolated.

Figure 9: CCA and Country

In the second place, it is interesting to analyse how CCA’s groups are distributed according to country, as shown in figure 6. The aim is to understand whether there are some shared cultural meanings among citizens of the same country. I analyse the percentage of individuals belonging to the CCA’s groups for each country. Group 2, 4, and 5 are composed of a similar proportion of citizens from each country. Instead, I find that over 30% of citizens from Cyprus, Ireland, Malta, Portugal, Slovakia, Spain, and the United Kingdom belong to group 1. Additionally, over 30% of citizens from Austria, Bulgaria, Czech Republic, Estonia, Finland, France, Germany, Greece Italy, Latvia, Luxemburg, and Sweden belong to group 3. In both cases, I can not find some specific patterns, such as belonging to the same part of Europe. It is interesting to find similar shared meanings schemas among subjects in nations extremely different and opposite such as Italy and Estonia.

**Partitioning Around Medoids Clustering**

PAM clustering is an unsupervised method that looks for patterns without any information of the classification target. The aim is to partition citizens into classes (clusters) according to similar *green-identity*. Clustering groups the similar observations within each group, while the observations in different groups are different from other clusters. Grouping the attitudes of citizens in clusters means dividing them according to similar opinions about the governance of climate change. I remember that also in this case, like CCA algorithm, the questions about climate change (qb4\_3, qb4\_5, qb7, qb8, qb9) are fitted in that algorithm. Also for this PAM algorithm, I use the *cluster* package in the R software.

PAM requires, as a parameter, the number of clusters. “One of the most commonly applied methods for assessing cluster validity is silhouette width which encompasses two clustering criteria: separation (i.e., the average distance to the closest other clusters) and compactness (i.e., average within‐cluster distance)” (Lengyel & Botta‐Dukát, 2019, pag. 13232). Silhouette width indicates how well each cluster divides observations, as shown in figure 7. The best choice is 2.

Figure 10: Clustering silhouette

Observations are then divided into two clusters with PAM. The summary results of PAM clustering are shown in figure 8. The distribution is balanced: 11171 (51%) individuals belong to cluster 1, and 10807 (49%) belong to cluster 2. Although overall the level of agreement or importance of these questions is elevated (the mode of qb4\_3, qb4\_5 is 2, while the mode of qb7, qb8, qb9 is 1), I can see that 2 clusters resemble two different types of green-identity, which I call “moderate green identity”, “extreme green-identity”.

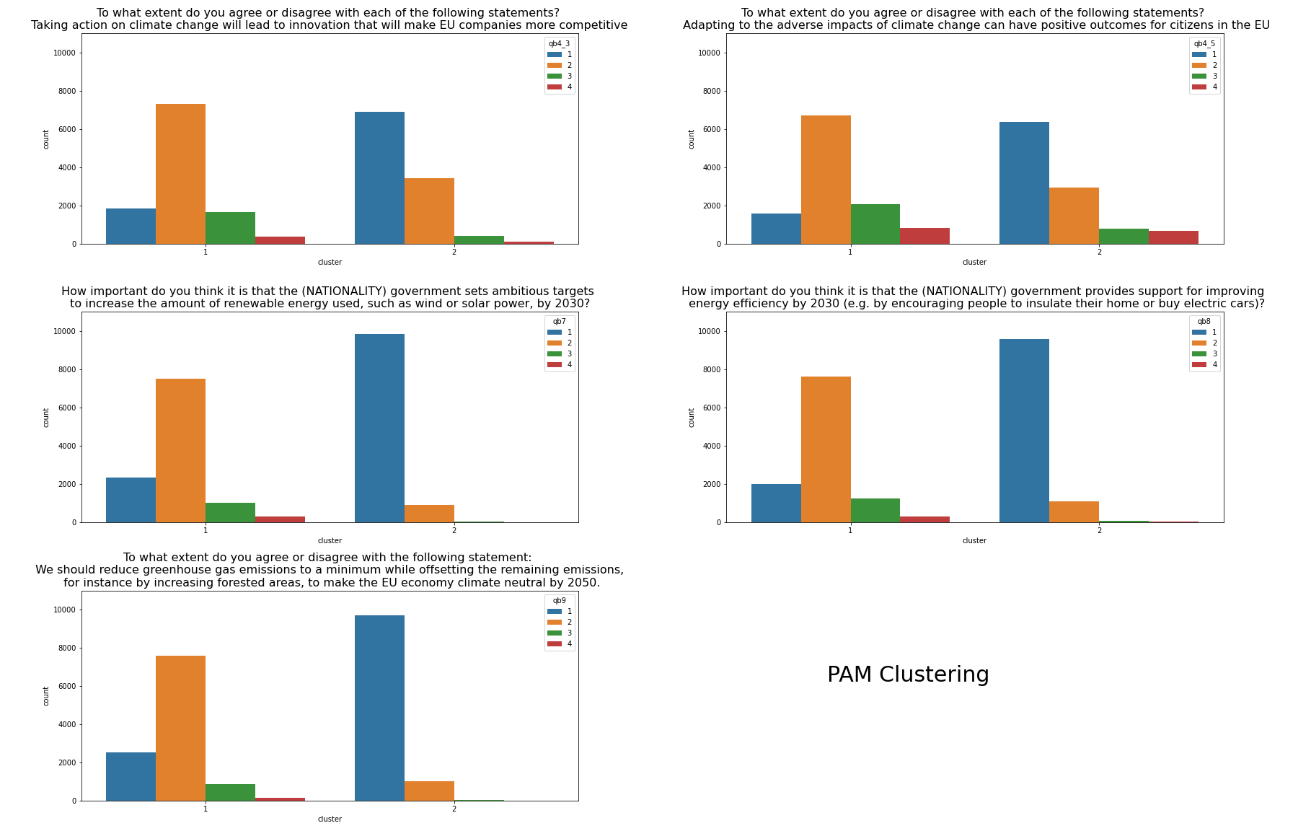


Figure 11: Summary of PAM Clustering

***Cluster 1: Moderate green identity***

* More than 75% answered questions (qb3\_4,qb4\_5, qb7, qb8, qb9 ) with moderate answers (option 2 or 3);
* There is no difference among genders;
* 70% of citizens belong to centre, centre-left or centre-right;
* 57% stopped studying at the latest at 19, 35% stopped studying after 20, 6% are still studying;
* More than 65% of the citizens of Czech Republic, Estonia, Finland, Latvia, Poland belong to this cluster;
* the average age is 51 years old;
* the average climate change risk perception is 7.2;
* 56% take place a pro-environmental behaviour.

***Cluster 2: Extreme green-identity***

* More than 60% answered qb3\_4 and qb4\_5 questions with the maximum answer (1), and about 90% answered qb7, qb8, qb9 questions with the maximum answer (1).
* There is no relevant difference among genders;
* 66% of citizens belong to centre, centre-left or centre-right
* 49% stopped studying at the latest at 19, 41% stopped studying after 20, 7% are still studying;
* More than 65% of the citizens of Cyprus, Denmark, Spain, United Kingdom belong to this cluster;
* The average age is 50 years old:
* The average climate change risk perception is8.6;
* 75% take place a pro-environmental behaviour.

To sum up, cluster 1 is composed of citizens that are more moderate in the answers to climate change items than the citizens of the opposite cluster. Additionally, cluster 2 is composed of slightly younger, more educated, more worried, and more active citizens.

**3.2 Explanatory Data Analysis**

**Entire dataset**

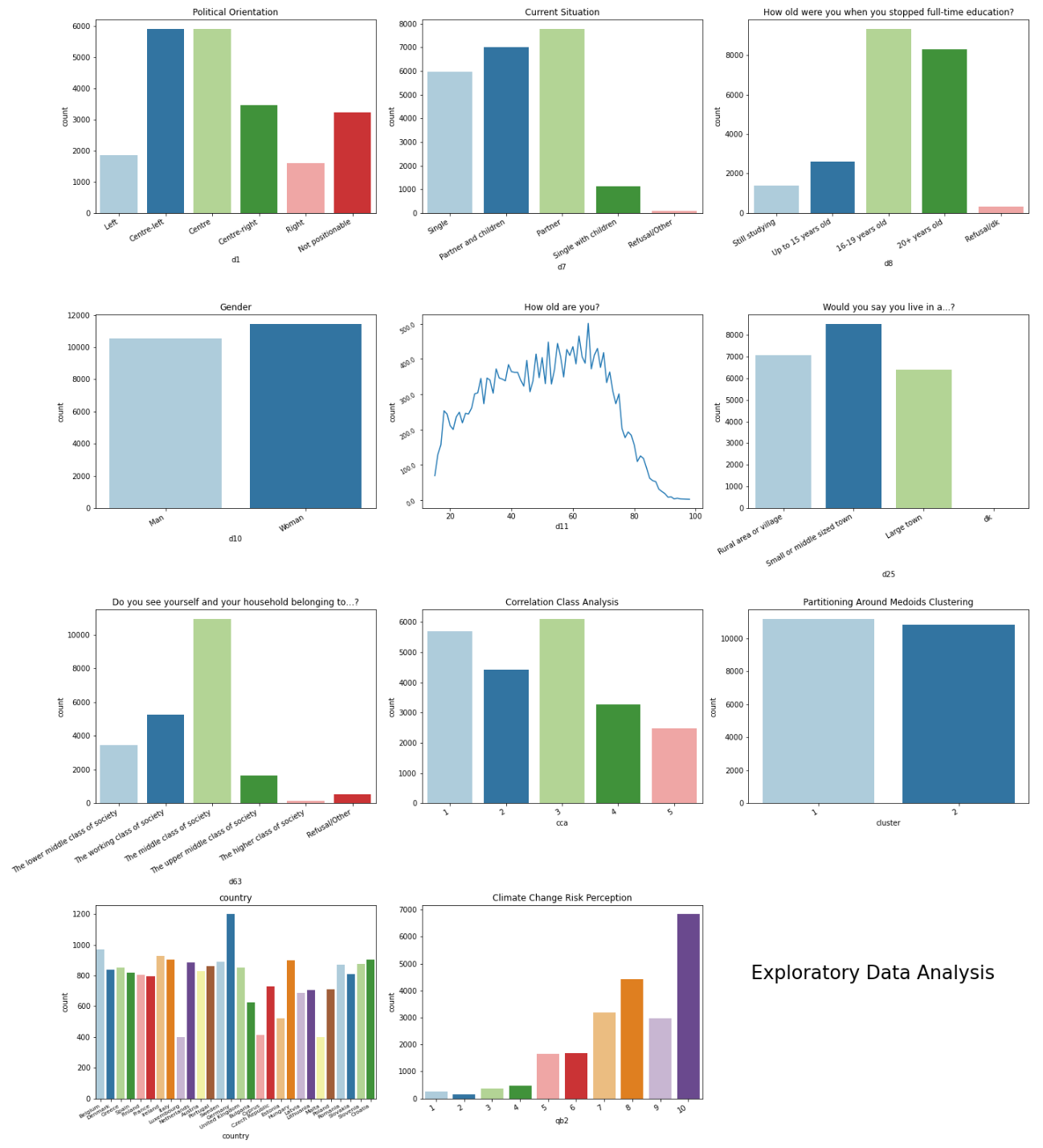
****

Figure 1: Countplot of explanatory variables

Before starting with the analysis it is important to describe the dataset. Figure 1 shows a quick summary of the explanatory variables. Dataset is composed by:

* 80% of citizens have a central political orientation (between centre, centre -left, and centre -right);
* 70% of individuals live with partner and/or children;
* 42% of individuals stopped studying between 16 and 19 years old, while 38% after 20 years old;
* 52% of women;
* the average age is 50.5;
* 39% live in a small or middle-sized town;
* Germans are over-represented when compared to the number of other citizens of other countries;
* Climate change risk perception average is 7.9, while the mode is the answer 10.

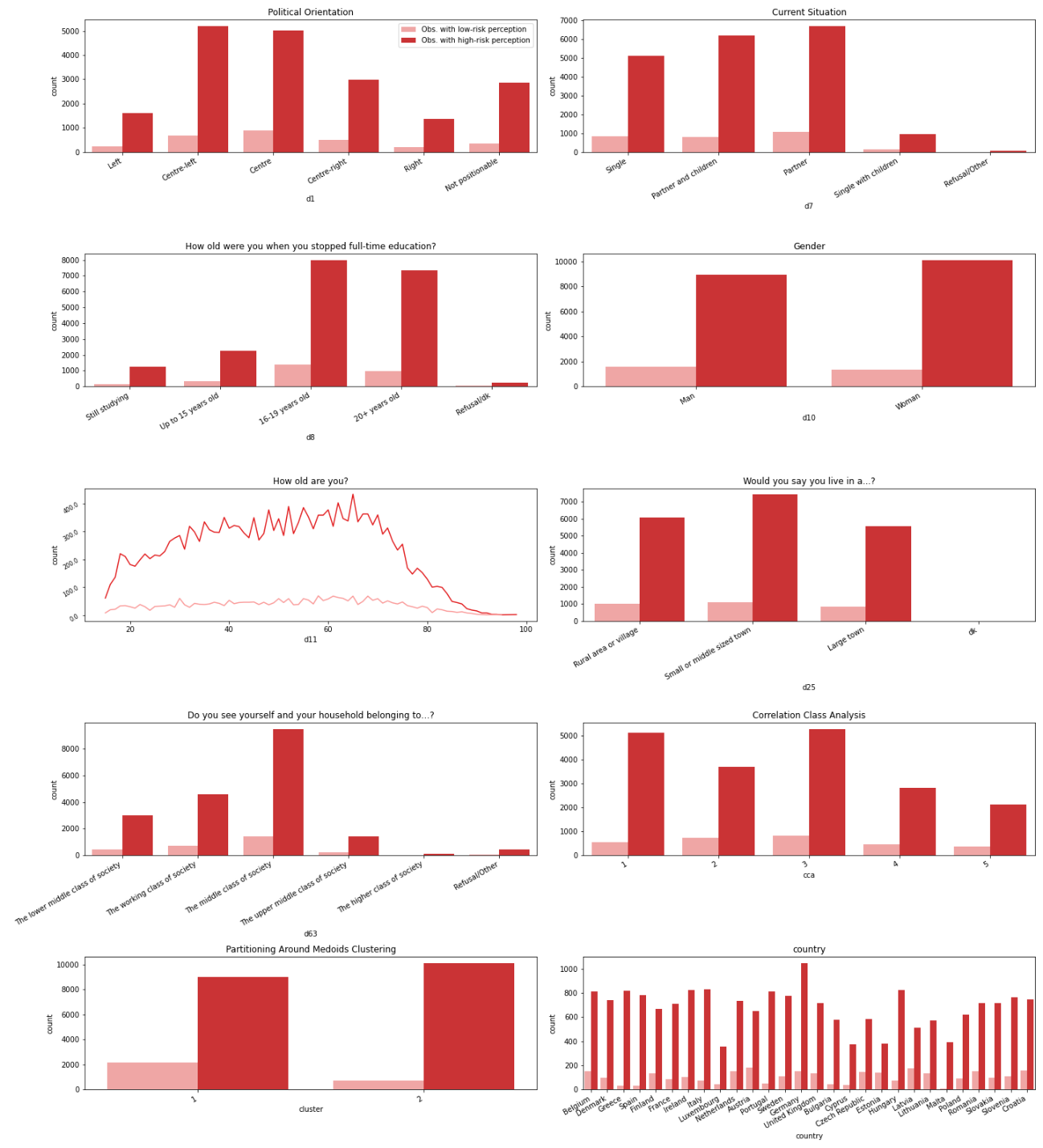
**Individuals with high-risk perception against individuals with low-risk perception**  

Figure 2: Countplot of explanatory variables according to risk perception level

Instead, figure 2 displays the description of the two datasets, according to the climate change risk perception level. As you can see, the two subsets are not balanced. The subset with low-risk perception observations (in the graph it is represented with pink color) has 2906 cases (13%), while the high-risk perception one (red color) has 19072 cases (87%). The trends of all variables of the high-risk perception subset follow those of the complete dataset. Instead, the low-risk perception subset is slightly different. For example, individuals of this latter subset are slightly older and mostly men. Additionally, most of the citizens do not come from Germany but Austria. Lastly, as I have already explained, the distribution of clusters between the two subsets is opposite: on one side mostly individuals who belong at cluster 2 (extreme green-identity) have a high-risk perception, and those who belong at cluster 1 have a low-risk perception.

**3.3 Macro-level**

Climate change attitudes do not vary only between countries but also between citizens in the same country (Xie et al., 2019). As you can see in figure 3, the percentage of those who believe that climate change is the single most serious problem varies significantly according to country. For example, Bulgaria and Croatia obtain a smaller percentage, that is 11% of citizens who think climate change is the single most serious problem. On the contrary, about 1 out 2 of Sweden’s citizens indicated climate change.



Figure 3: Single Most Serious Problem per Country

Another interesting example is the difference in behavior. As you can see in figure 4, more than 60% of the citizens of Romania declared that they are not behaving in favor of the environment. Conversely, Malta has 95% of citizens that take place pro-environmental behavior. These differences may be due to a multitude of reasons, differences in economic incentives and taxes, way of socialization, social desirability that vary according to country. These possible characteristics are not taken into account, but I want to remind the reader of the hypothetical influence that these could have on the outcome.



Figure 5: Pro-environmental Behaviour per Country

Attitudes among countries could so vary since they are influenced by different contextual factors (Echavarren et al., 2019; Krajhanzl, 2010). According to Echavarren and colleagues (2019), opinion, perception, and behavior could change due to different natural hazards and political contexts. For example, water deficit, temperature growth, or the level of environmentalism in the political activity of a given nation may shape citizens’ attitudes or behaviours (Echavarren et al., 2019). These macro-variables should be significant mediators in explaining risk perception or pro-environmental behavior. Some indexes are considered with the sole purpose of remembering that they could affect and moderate the phenomenon studied. Then, they are not inserted in the final models since only the multilevel analysis could be adopted. Besides, the aim of the research is not to evidence national or cultural differences, but on the contrary, it is to find patterns at individual levels, regardless of the place of origin. However, these differences at the macro levels are presented.

For the natural hazards, the 2020 Environmental Performance Index (EPI) is used (the 2019 EPI is not available in order to use the same data of year of this survey) (Yale Center for Environmental Law & Policy, 2020). EPI quantifies numerically environmental health and ecosystem vitality around the world. This index helps us to describe with a score the severity of environmental problems in each country. Some indicators that composed the index are air pollution, drinking water quality, species protection. These phenomena could positively affect climate change concerns and opinions (Echavarren et al., 2019). Citizens should perceive biodiversity loss or temperature increases, leading to greater apprehension. Figure 6 shows the score across European Union (EU). The best score is obtained from Denmark, while the worst from Bulgaria.

Figure 6: The 2020 EPI

For the political context, the 2019 Climate Change Policy Performance is selected, which is a measurement of national and international climate policies (Burck, 2018) developed by organization Germanwatch. It is one of the indicators that belong to the Climate Change Performance Index (CCPI). The indicator constitutes the measurements taken by governments to reduce the current level of GHG emissions per capita or the use of renewable energy. Briefly, it is defined as a measure of countries’ progress and their capacity for climate protection (Burck, 2018). In the Climate Change Policy, the record goes to Portugal and Bulgaria gets the lowest score in all European Union, as figure 7 shows.

According to scholars (Echavarren et al., 2019; van der Linden, 2015), socio-cultural context influences individual attitudes towards climate change concerns. Therefore, the notable differences in attitudes across countries should be also due to these indicators. In fact, “sociological research suggests that contextual factors and processes can be powerful forces shaping how individuals and communities engage with the issue” (Lee et al., 2015, p. 1014). There are different ecological tax reforms or cultural habits that affect and shape individual climate change attitudes and behavior.

Figure 7:The 2019 Climate Change Policy

In this way, It is important to remember that these macro-factors should affect also individual preferences.

Chapter 4

Analysis and Prediction

As mentioned in the literature review (Chapter 1), the main predictor of pro-environmental behaviour is climate change risk perception. However, other factors can shape the outcomes. The main focus is now on predicting the behaviour of citizens in a dummy outcome in the entire dataset, thus using all the observations. This part amis to undesterstand the actual role of climate change risk perception.

**4.1 Evaluating Classification Models**

Before proceeding with the analysis, it is important to explain the ways the evaluation will be measured, since one of the main factors according to which I will choose the best model is the maximization of these metrics. I consider two measurements of evaluations:

* **Accuracy**: the proportion of correct predictions to the total number of predictions (true positives and true negatives) given by the classifier (Battiti & Brunato, 2014). Accuracy is the most used metric that generally describes the goodness of a model. The formula is:

The range is 0-1, where 0 is the worst score and 1 the best score, all the inputs are predicted in the right way.

* **Macro F1-score average** (macro-F1 for short): combine precision and recall, where the precision is the fraction of the number of true positives among the total number of items labeled as positive, and the recall is the fraction of numbers of true positives among the total number of features that belong to the positive class (Shmueli, 2019). macro-f1 computes the harmonic mean, as shown in the formula:

The scoring range is 0-1, where 0 is the worst score and 1 is the best one. It is used, especially, when for the class imbalance problem due to it is more sensitive to data distribution, as in this case.

To select the best model for each classifier, I decide to compromise: a fair balance between accuracy and macro-f1. The aim is to classify and predict as well both classes. On one side, this compromise can sometimes lead to losing a few percentage points of accuracy but, on the other side, it improves accuracy within the classes.

In closing, I explain how to understand what variables have a fundamental role in our models. I compute feature importance by the different algorithms in *scikit-learn*. Logistic regression finds a set of coefficients to use in the weighted sum to produce a prediction. These coefficients can be used directly as a feature importance score. Instead, tree-based models can measure feature importance in two ways: Gini Importance or Mean Decrease Accuracy. Gini Importance counts the times a variable is employed to split a node. Mean Decrease Accuracy estimates how much accuracy the model losses by omitting each feature. However, *scikit-learn* implemented Gini Importance, also in this analysis, it is adopted.

**4.2 Prediction Behavior In the Entire Dataset**

In this part, the research study aims to identify and discover the most significant variables and offer a good prediction of pro-environmental behaviour. As presented in the Methodology (Chapter 2), different classifiers are trained and implemented to predict behaviour. The analysis starts with Logistic Regression and continues with tree-based methods: Decision Tree, Random Forest, and Gradient Boosting. These algorithms are implemented in *scikit-learn*. For each classifier best tuning parameters, called hyperparameters, are fitted. The technique adopted for knowing the optimal hyperparameter is called *random search* (RandomizedSearchCV() in *scikit-learn*) (Benner, 2020).[[2]](#footnote-2) I adopt, in turn, for random search 3 fold cross-validation, a resampling procedure to evaluate model performance. In practice, each possible random combination will train and evaluate for three different folds.

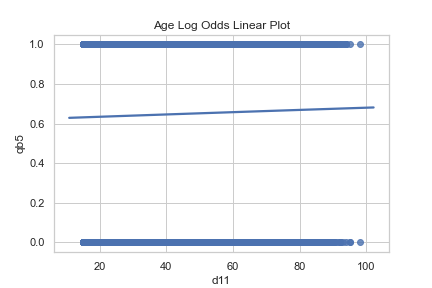
I remember that the independent variables, at an individual level, of this part are:

* Climate change risk perception (qb2)
* CCA’s class (cca)
* Cluster’s class (cluster)
* Political orientation (d1)
* Marital status (d7)
* Stopped full-time education (d8)
* Gender (d10)
* Age (d11)
* Residence (d25)
* Class identity (d63)
* Country

Each independent categorical variable is converted into a dummy, to fit the supervised algorithms. All explanatory variables are categorical, except for risk perception, which is maintened as a metric, and age. After the conversion, the dataset has 65 independent variables.

**Logistic regression**

The first method explored is Logistic regression. I use *LogisticRegression()* function in *scikit-learn* library for the Python programming language. This algorithm supposes that all the assumptions (linearity in the logit for continuous independent variables, the absence of multicollinearity, the absence of multicollinearity, lack of outliers) are satisfied. However, there are some violations of different assumptions.



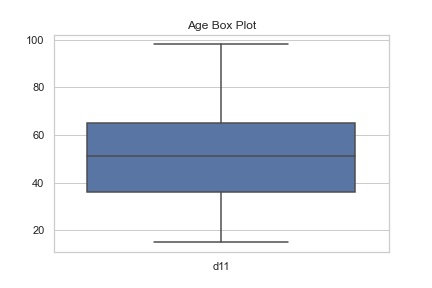
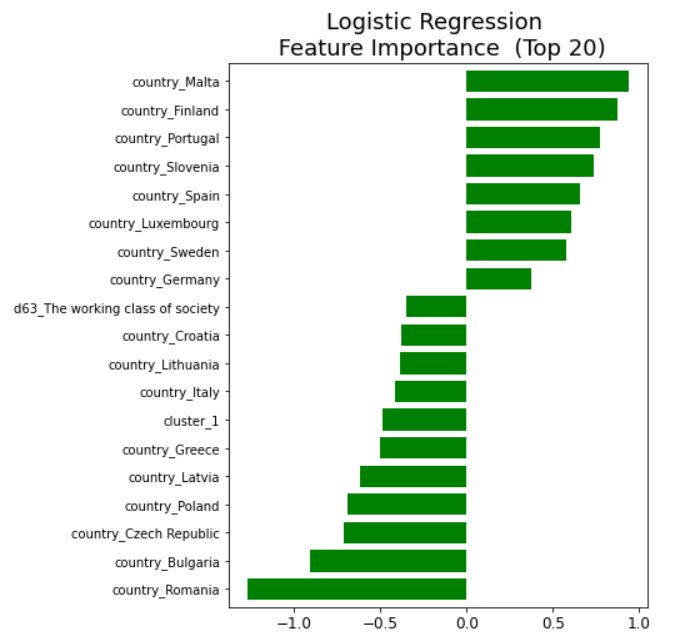
First, I check the assumption of linearity in the logit for continuous independent variables, also in this case only for the continuous variables. In this case, only age is continuous, and it does not verify this assumption, as figure 10 displays (there should be an “s” curve line). Second, I can not check the absence of multicollinearity (independent variables), due to it is not possible to compute correlation with one variable alone. Lastly, lack of outliers, always for continuous variables. Figure 11 shows that there are not outliers in the age variable.

Figure 12: Lack of outliers for Age

Figure 13: Linearity in the Logit for Age

The fitted model, with the best tuning parameters, has 0.67 of accuracy and 0.66 of macro f1-score.[[3]](#footnote-3) The figure displays all feature importance. Different countries are the most important variables in this model, as shown in figure 13. The coefficient plot shows that Romania, Bulgaria, Czech Republic, and Poland have a negative influence on predicting pro-environmental behaviour. it very interesting due to all these first countries are situated in Eastern Europe. Instead, Malta, Finland, Portugal have a positive influence on predicting pro-environmental behaviour. I remember that these coefficients are in log-odds terms.[[4]](#footnote-4) It is not easy to compare these results with the feature importance of the other algorithms. These last describe Gini importance.



**Decision Tree**

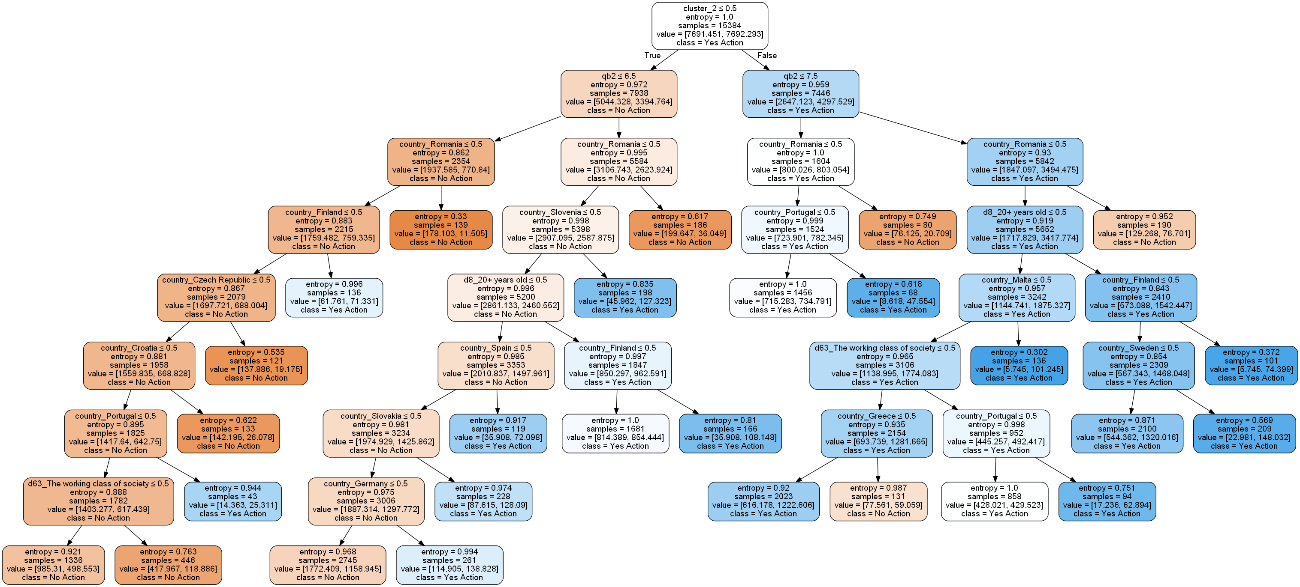


Figure 14: Decision Tree

First of all, the decision tree is implemented with *DecisionTreeClassifier()* function using the *scikit-learn*. Figure 14 displays the tree, the color of the dots represents behaviour: orange for no-action, blue for yes-action. The main advantage is the immediate interpretation analysing the figure. However, this model gets the worst performance if compared with the other ones. Accuracy is 66% and f1 score is 53. Cluster 2, or called extreme green-identity, is the predictor variable used for the primary split, it is the root node. Risk perception is the predictor variable used for the second and the third split. Whether the individual belongs to cluster 2 and he/she has a level of risk perception greater or equal than 7,5 the outcome is yes-action. if the individual does not belong to cluster 2 and he/she has a level of risk perception (less or equal to 6.5), he/she classify in the no-action class.

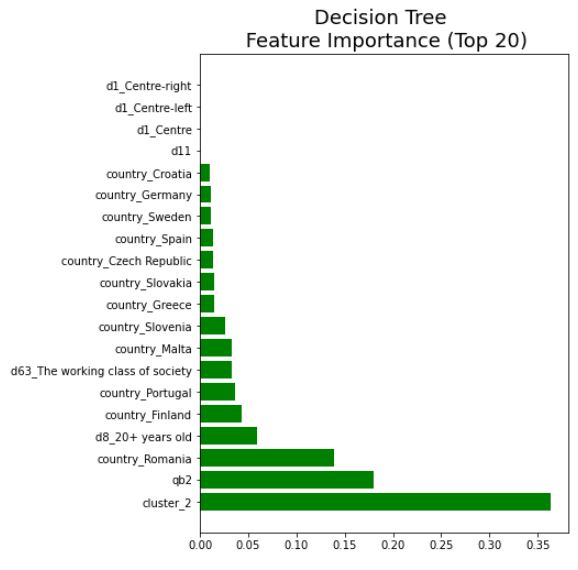


Figure 15: Decision Tree - Feature Importance

Figure 16 shows feature importance. As I have already discussed, the principal variable is cluster 2, the second one is climate change risk perception.

The result is quite different from logistic regression. Extreme green-identity (cluster 2) and risk perception for the first time are the main predictors, as the literature review affirmed. 75% of individuals belonging to cluster 2 have done some eco-friendly actions. For this reason, that cluster is particularly influencing on predicting behaviour. Climate change risk perception is even more interesting to analyse, as figure 14 shows. As referred to in the literature review (chapter 1), the more an individual worries, the more he tends to take place environmental action. This hypothesis is confirmed. If I analyse the percentage of those who take place action within the level of perceived risk, I find that for the classes with a low level there are more than 50% of individuals that declare to not take place action. The more the risk increases, the more the percentage of taking place action increase in turn.

Figure 16: Crosstab between Risk Perception and Behaviour

**Random Forest**

Random Forest is implemented with *RandomForestClassifier()* function using the *scikit-learn*. Compared to the previous models, the accuracy and f1-score greatly improve, there are respectively 70% and 66%. It is the best model.

Once the model is proved to fit the data, the variable importance is explored. Plot 15 describes the 20 most important variables. I can identify, another time, climate change risk perception. It is the second influential variable for predicting behaviour. Instead, the first one is age. Also, cluster 2 and 1 are relatively important, even if definitely less than the first two.

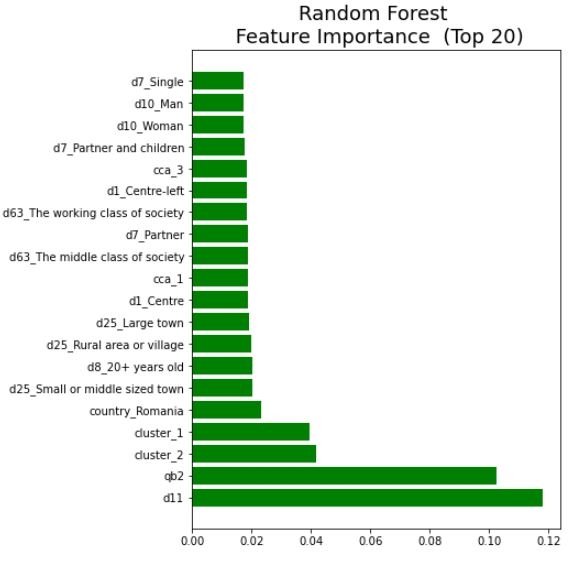


Figure 17: Random Forest - feature importance

As shown in figure 15, analysing the feature importances of the model, I see that age and, again, climate risk perception are the most important variable to classify and predict pro-environmental behaviour. I expect that most young people take place pro-environmental action. However, this relationship is not so clear. If I analyse the percentage of those who take place action within the age group, we find that on average 60% do something. This percentage drops for classes over 80. The relationship between behaviour and risk perception has already been explained in the decision tree model. Other important features are both created clusters, which correspond to the different green-identity. As we have already seen, cluster 1, moderate green-identity, is more likely to do not behave in favor of the environment than cluster 2.

**Gradient Boosting**

Gradient Boosting concludes the first part of the analysis. *XGBClassifier()* function is implemented in using the *scikit-learn*. Gradient Boosting gets slightly worse results than the Random Forest classifier. Accuracy is 69% and f1-score 65%. Again, the variable importance is checked and compared to the previous list. Cluster 1, moderate green identity, is the best predictor. It is followed by Romania and climate change risk perception. However, a new important variable presents in the top 5: high-level if education (d8\_20+years old). 72% of those who have declared to stop study after 20 years take place pro-environmental behaviour. The percentage is considerably higher than the other categories that have studied less or the students.

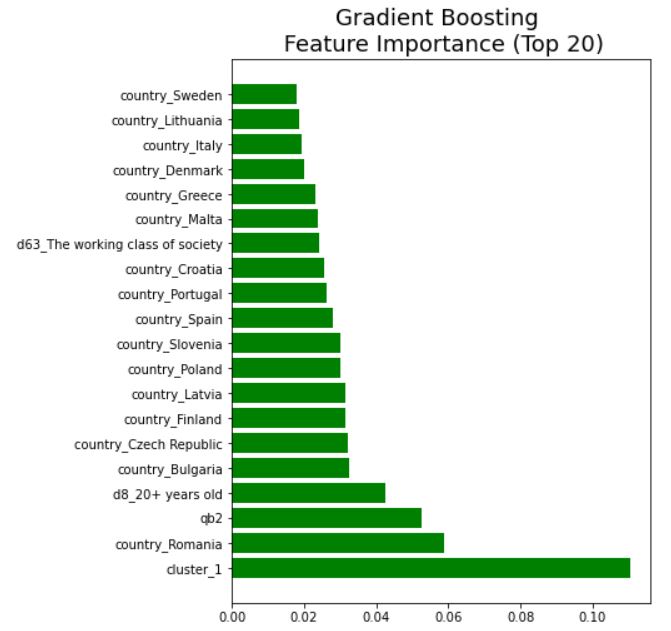


Figure 18: Gradient Boosting- Feature Importance

In conclusion, I see that the main pattern is to identify as important features climate change risk perception and green-identity. Another recurve important variable is country\_Romania. As I have seen Romania has a high percentage of citizens who do not behave eco-friendly. However, our aim is not to identify what is the country more ecological but is to understand what are the individual characteristics that shape the behaviour. I can confirm the hypothesis that climate change risk perception is the most important variable. It is appeared in all models, except for logistic regression. Therefore, we can continue with the next analysis.

**4.3 Prediction methods summary**

It can be convenient, to summarize quickly, the predictive accuracy and macro-f1 of the models trained in table 1.

Table 1: Metrics comparison

|  |  |  |
| --- | --- | --- |
| **Models** | **Accuracy** | **Macro-F1** |
| **Logistic Regression** | 0.67 | 0.65 |
| **Decision Tree** | 0.64 | 0.62 |
| **Random Forest** | 0.70 | 0.67 |
| **Gradient Boosting** | 0.69 | 0.65 |

Most of the features with predictive importance are the same across models. Moreover, these predictors align with the hypotheses presented in section 2.1.

Climate change risk perception is one of the predictors which align in almost all models. Therefore I confirm that a higher individual climate change risk perception positively influences and predicts pro-environmental behaviour.

Green identity, measured with clusters, is another important variable found in almost all models. Therefore also the second hypothesis is confirmed.

However, the third hypothesis is rejected. Cultural schemas, measured with CCA classes, are not relevant in the prediction of behaviour.

The last hypothesis concerns the importance of socio-demographic information. Random Forest, our best model, suggests that the most important variable is age. However, it is not clear the relationship with the behaviour. Not only young people perform pro-environmental behaviour. Also, higher education has a positive effect on pro-environmental behaviour and it is one of the most important variables in the prediction. Gender, income, and political orientation seem to be relevant.

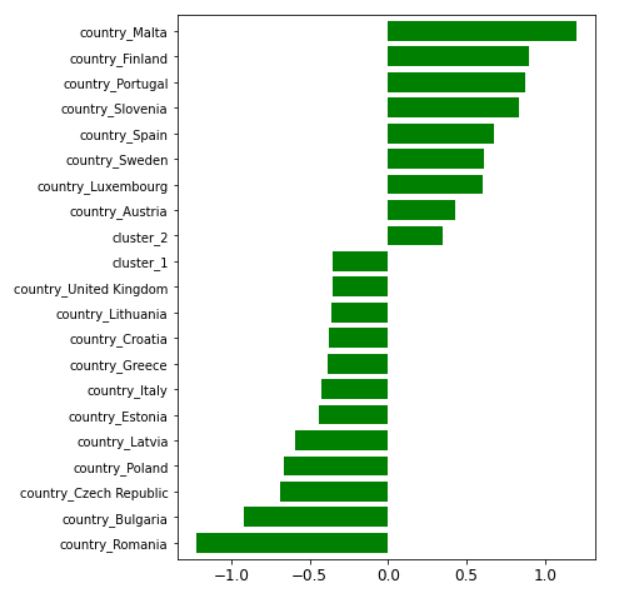
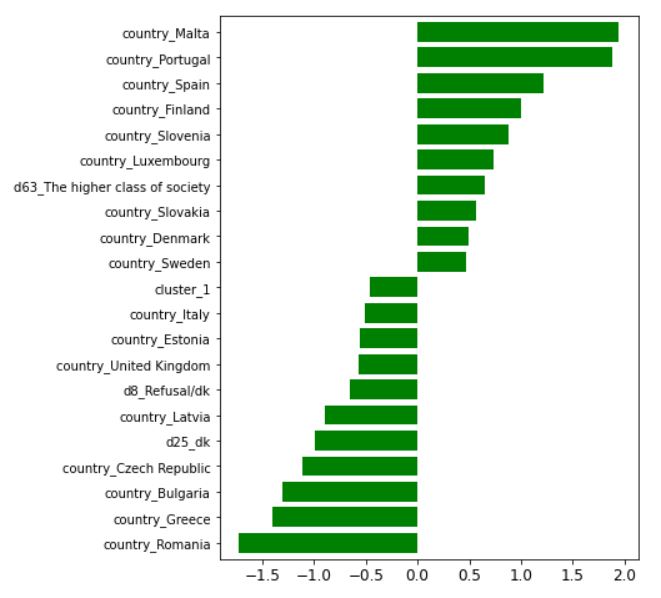
However, I discover that a fundamental role in the prediction is the country of origin. Especially we find Romania as an important variable in all models. It is a negative effect on pro-environmental behaviour.

**4.4 Pro-environemntal behaviour according to risk perception level**

In the second part of the analysis, I divide the subset into two: one with the only observations of those who have a high-risk perception (level greater than or equal to 6) and those who have a low-risk perception (level less than or equal to 5). For convenience, I call the first subset “of those who have a high-risk perception” and the second one “of those who have a low-risk perceptio”. At this point of the analysis, I can confirm the hypothesis suggests from the literature: climate change risk perception is one of the main factors for predicting pro-environmental behaviour. Now, I want to understand what are the most important predictors if we change the level of risk perception. Also in this part, I implemented the same algorithms for both subsets: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Hyperparameters using random search are found.

We remember that the independent variables, at an individual level, of this part are:

* CCA’s class (cca)
* Cluster’s class (cluster)
* Political orientation (d1)
* Marital status (d7)
* Stopped full-time education (d8)
* Gender (d10)
* Age (d11)
* Residence (d25)
* Class identity (d63)
* Country

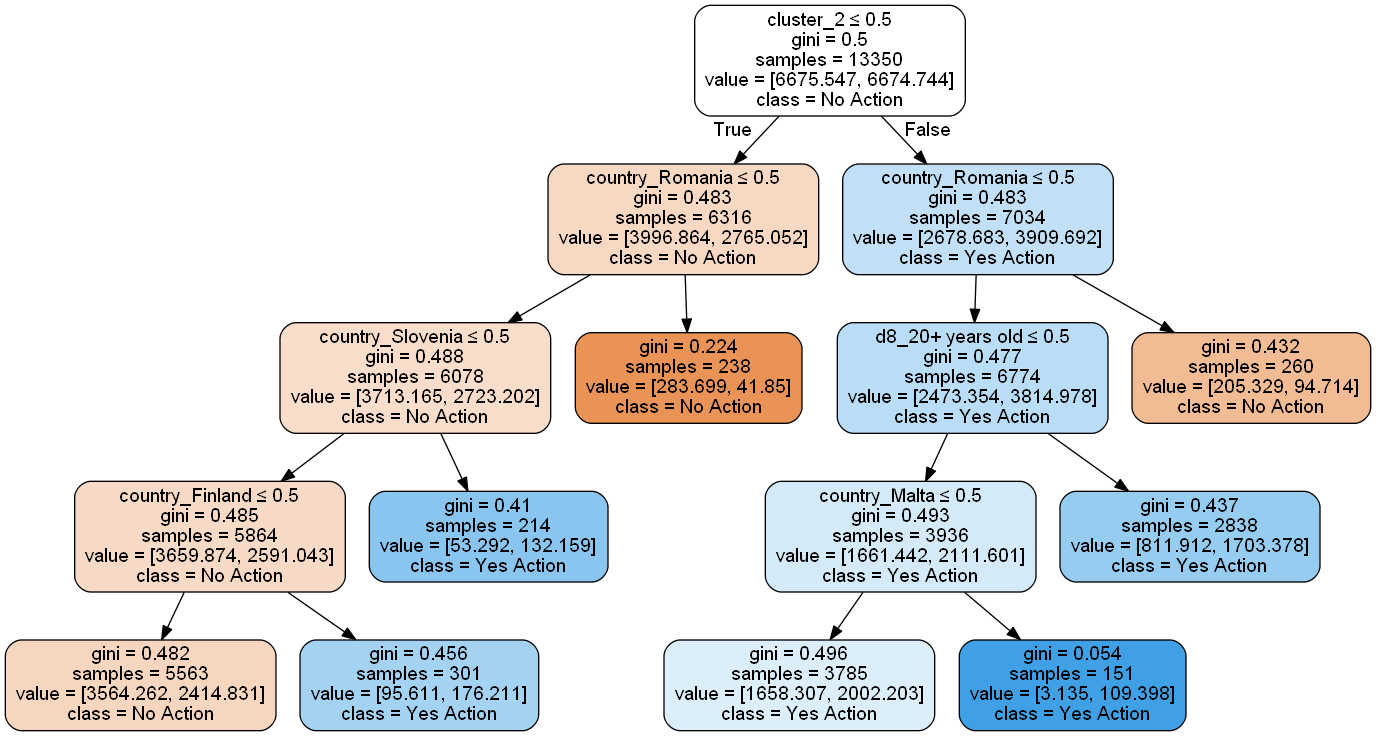
**Logistic Regression**

|  |  |
| --- | --- |
| 1. Subset with High-risk Perception observations | 1. Subset with Low-risk Perception observations |

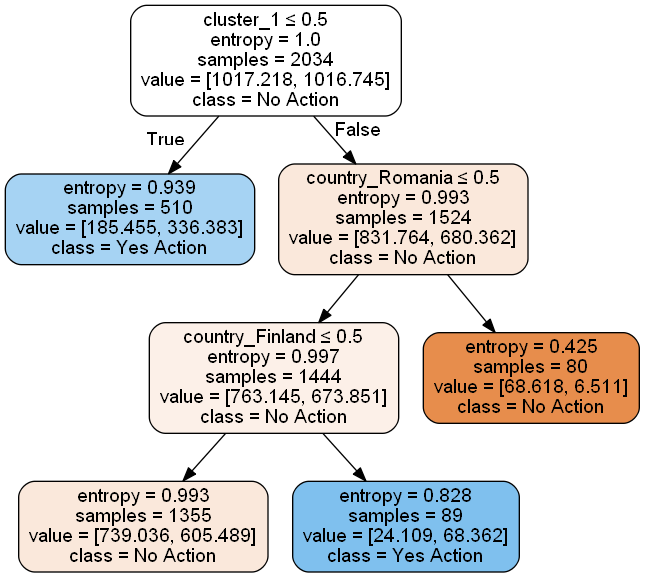
Figure 19: Comparison of variable importance (top 20) - Logistic Regression

Figure 17 displays the first 20 feature importance for predicting pro-environmental behaviour according to the different levels of climate change risk perception. The important variables in the two different subsets are similar. I found always Romania with a negative influence on predicting pro-environmental behaviour. Instead, in both cases, Malta is the first variable with a positive influence.

**Decision Tree**

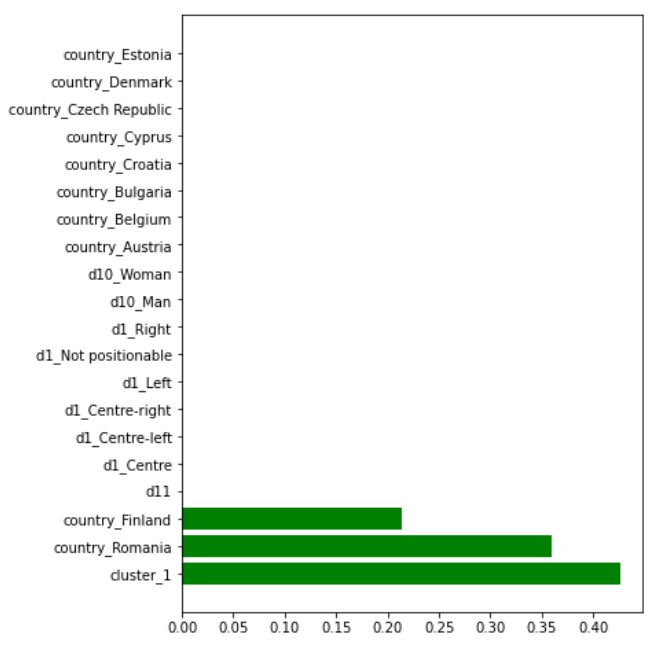
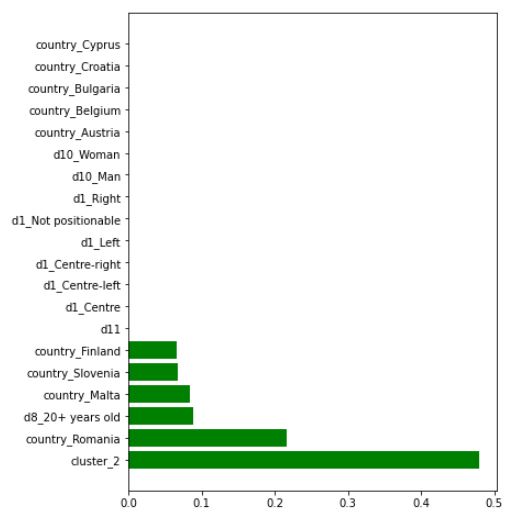
****

1. **Subset with High-risk Perception observations**



1. **Subset with High-risk Perception observations**

Figure 18 shows the two different decision trees. The first one, with the observations of those who have a high-risk perception, is more branched. For both trees, the root is the green-identity. On one side in the first tree I have cluster 2, the extreme green-identity, on the other side we have cluster 1, the moderate green-identity. I have another confirmation of the hypothesis formulated. A strong green-identity, mixed with a high-risk perception, influences positively pro-environmental behaviour. the opposite happens in the second tree: and moderate green-identity mixed with a low-risk perception influences anti-environmental behaviour.



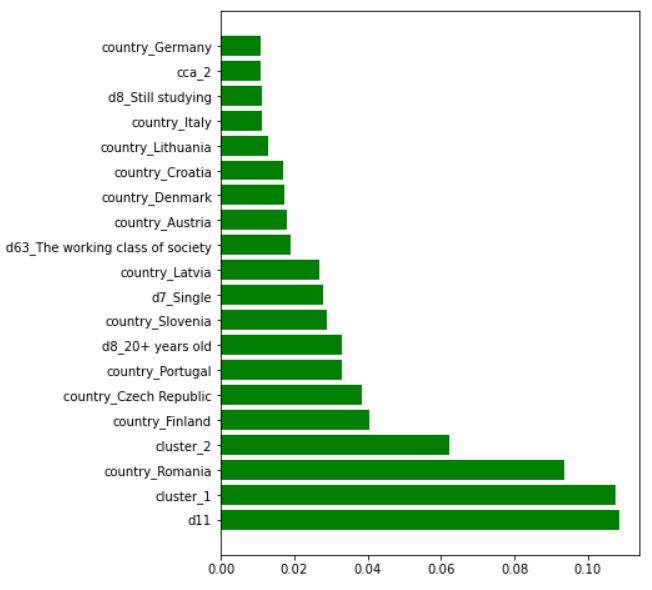
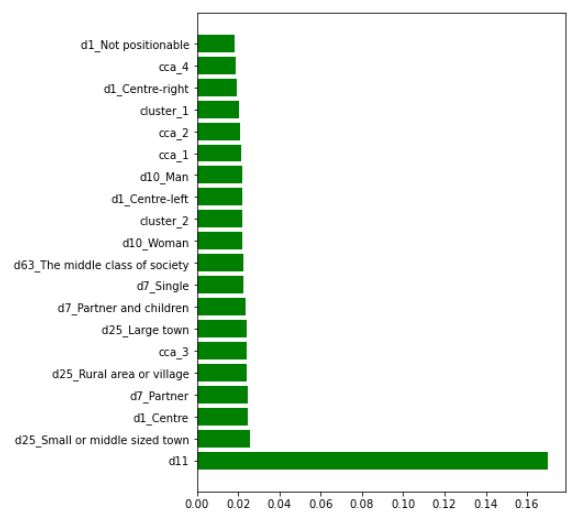
|  |  |
| --- | --- |
| 1. Subset with High-risk Perception observations | 1. Subset with Low-risk Perception observations |

Figure 20: Comparison of feature importance- Decision Tree

The comparison of feature importance in figure 19, is extremely different from the previous model. As I have just explained, the important variables for the respective subsets are cluster 2 for high-risk perception observations and cluster 1 for low-risk perception. Country Romania is the second important variable for both models. In the first one, Romania has the 63% of citizens who do not perform pro-environmental behaviour. In the second one, the percentage goes to 85%. Thus, Romanian, with a low-risk perception, is very likely not to perform a pro-environment behavior. The third variable important for the first subset is the high level of education. In this subset, there are 75% of high-educated individuals perform environmental-friendly actions. In the second subset, in the third place, I have Finland. In this case, Finland has a positive influence in predicting pro-environmental behaviour, due to 70% of citizens who do not worry, at the same time, they behave in favor of nature.

**Random Forest**

Both Random forest’s models classify and predict well the outputs. 0.70 accuracy and 0.65 macro-f1 for the first subset with high-risk perception, 0.65 and again 0.65 for the second one with low-risk perception.



|  |  |
| --- | --- |
| 1. Subset with High-risk Perception observations | 1. Subset with Low-risk Perception observations |

Figure 21: Comparison of variables importance- Random Forest

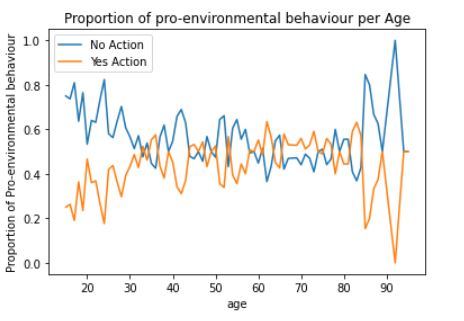


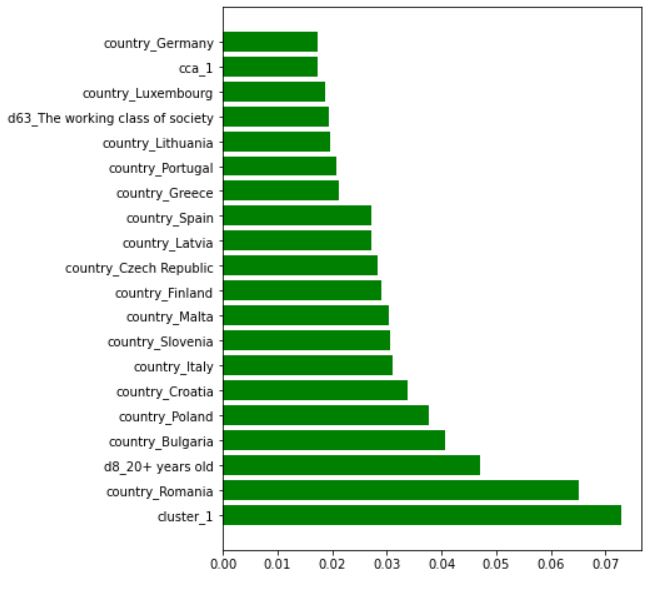
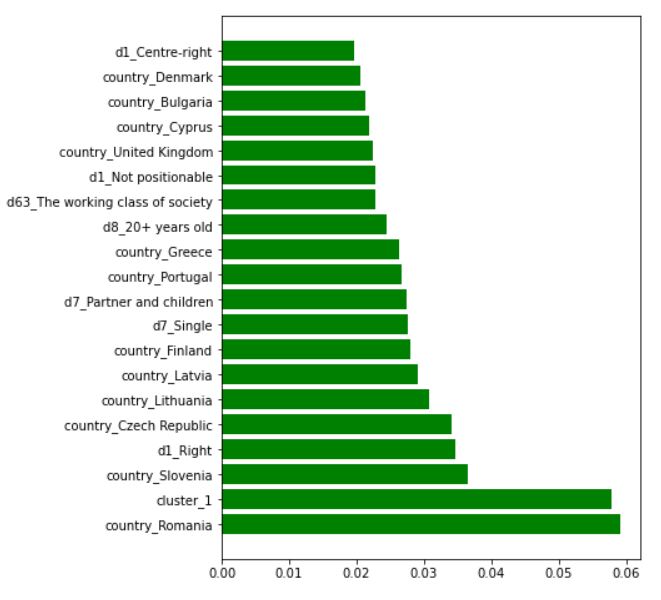
Figure 22: trend of age in the Subset with low-risk perception observations

Figure 20 shows the importance of variables in both subsets. In both cases, age is the most important variable for predicting behaviour. The variation of age in the subset with high-risk perception is similar to the complete model. On average 60% of each age group perform eco-friendly action. Another time, this percentage drops from 80 years old. More interesting is the trend in the second subset, as figure 21 shows. On average 70% of younger have a negative behaviour in favor of the environment. This percentage drops with the increase of years, and the trend is reversed. This trend is opposite to the sociological theories explained in Literature Review (Chapter 1).

Then, the first subset has many important variables but in a much lesser way, such as the small or middle-sized town, 68% of this class in this subset perform environmental-friendly action. In the second subset, I have more relevant variables, or rather with a higher Gini Importance score. For example, cluster 1 is the second most important variable, where 60% do perform environmental actions. Romania is again in the top three. Then, there is cluster 2, where 62% do perform environmental actions even if they are not worried. In this case, behavior seems to not change with the same risk perception between two different types of green-identity.

**Gradient Boosting**

The first model of Gradient Boosting gets 0.67 of accuracy and 0.64 of macro-f1, and the second one gets respectively 0.64 and 0.63. The model with high-risk perception observations gets better results when compared with others. The accuracy is slightly less than a few percentage points than Random Forest’s accuracy but in turn, macro-f1 is higher. Since the first subset is particularly unbalanced, it is better considered macro-f1 as evaluating parameter.



|  |  |
| --- | --- |
| 1. Subset with High-risk Perception observations | 1. Subset with Low-risk Perception observations |

Figure 23: Comparison of varibale importance- Gradient Boosting

Figure 22 shows the comparison of feature importance of our last models, using Gradient Boosting classifier. In both models, I find cluster 1, moderate green-identity, and Romania. In the first subset, there is also the high-education level as an important variable. In the second one, I find two new variables: Slovenia and right (political orientation). Slovenia has a positive influence on pro-environmental behavior, due to 70% of Slovaks in this subset perform eco-friendly actions. For the first time, political orientation compare. If I compute the proportion in this subset 50% of radicals take place pro-environmental behaviour, while the other orientations get a smaller percentage.

**4.5 Summary**

Tables 2 and 3 summarize the performance of all models.

Table 2: Metrics comparison-Subset of those who have a high-risk perception

|  |  |  |
| --- | --- | --- |
| **Models** | **Accuracy** | **Macro-F1** |
| **Logistic Regression** | 0.65 | 0.63 |
| **Decision Tree** | 0.63 | 0.60 |
| **Random Forest** | 0.70 | 0.60 |
| **Gradient Boosting** | 0.67 | 0.64 |

Table 3: Metrics comparison-Subset of those who have a low-risk perception

|  |  |  |
| --- | --- | --- |
| **Models** | **Accuracy** | **Macro-f1** |
| **Logistic Regression** | 0.63 | 0.63 |
| **Decision Tree** | 0.60 | 0.58 |
| **Random Forest** | 0.65 | 0.65 |
| **Gradient Boosting** | 0.65 | 0.64 |

In both cases, Random Forest has yielded the best performance when compared to other classifiers, as shown in tables 2 and 3. However, Gradient Boosting has the best macro-f1 score in the subset of those who worry about climate change. this subset is particularly unbalanced, therefore macro-f1 is preferred to evaluate the model. With this comparison, I do not find any particular differences between the two opposite models. I expect some differences in the prediction of behaviour dividing the subset according to risk perception level.

REFERENCES

Battiti, R., & Brunato, M. (2014). *The LION way: Machine learning plus intelligent optimization* (1. ed). CreateSpace.

Benner, J. (2020, agosto 6). Cross-Validation and Hyperparameter Tuning: How to Optimise your Machine Learning Model. *Towards Data Science*. https://towardsdatascience.com/cross-validation-and-hyperparameter-tuning-how-to-optimise-your-machine-learning-model-13f005af9d7d

Burck, J. (2018). *CCPI, climate change performance index background and methodology*.

Echavarren, J. M., Balžekienė, A., & Telešienė, A. (2019). Multilevel analysis of climate change risk perception in Europe: Natural hazards, political contexts and mediating individual effects. *Safety Science*, *120*, 813–823. https://doi.org/10.1016/j.ssci.2019.08.024

Krajhanzl, J. (2010). *Environmental and Pro-environmental Behavior*. *School and Health*, 251–274.

Lee, T. M., Markowitz, E. M., Howe, P. D., Ko, C.-Y., & Leiserowitz, A. A. (2015). Predictors of public climate change awareness and risk perception around the world. *Nature Climate Change*, *5*(11), 1014–1020. https://doi.org/10.1038/nclimate2728

Lengyel, A., & Botta‐Dukát, Z. (2019). Silhouette width using generalized mean—A flexible method for assessing clustering efficiency. *Ecology and Evolution*, *9*(23), 13231–13243. https://doi.org/10.1002/ece3.5774

Rossoni, L., Gonçalves, C. P., da Silva, M. P., & Gonçalves, A. F. (2020). *Mapping Organizational Culture Schemes Based on Correlational Class Analysis* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/sf2v4

Rossoni, L., Gonçalves, C. P., Silva, M. P. da, & Gonçalves, A. F. (2021). Mapping Organizational Culture Schemas Based on Correlational Class Analysis: A Tutorial. *Revista de Administração Contemporânea*, *25*(1), e200096. https://doi.org/10.1590/1982-7849rac2021200096

Shmueli, B. (2019, luglio 3). Multi-Class Metrics Made Simple, Part II: the F1-score. *Towards Data Science*. https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

van der Linden, S. (2015). The social-psychological determinants of climate change risk perceptions: Towards a comprehensive model. *Journal of Environmental Psychology*, *41*, 112–124. https://doi.org/10.1016/j.jenvp.2014.11.012

Xie, B., Brewer, M. B., Hayes, B. K., McDonald, R. I., & Newell, B. R. (2019). Predicting climate change risk perception and willingness to act. *Journal of Environmental Psychology*, *65*, 101331. https://doi.org/10.1016/j.jenvp.2019.101331

Yale Center for Environmental Law & Policy. (2020). *Environmental Performance Index 2020*. https://epi.yale.edu/

**APPENDIX D. Correlation matrix for each CCA’s group.**

Group 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **qb4\_3** | **qb4\_5** | **qb7** | **qb8** | **qb9** |
| **qb4\_3** | 1 | 1 | 1 | 1 | 1 |
| **qb4\_5** | 1 | 1 | 1 | 1 | 1 |
| **qb7** | 1 | 1 | 1 | 1 | 1 |
| **qb8** | 1 | 1 | 1 | 1 | 1 |
| **qb9** | 1 | 1 | 1 | 1 | 1 |

Group2

|  | **qb4\_3** | **qb4\_5** | **qb7** | **qb8** | **qb9** |
| --- | --- | --- | --- | --- | --- |
| **qb4\_3** | 1 | 0.8209258 | 0.34825902 | 0.1660830 | 0.22869773 |
| **qb4\_5** | 0.8209258 | 1 | 0.48758501 | 0.3021902 | 0.13889897 |
| **qb7** | 0.34825902 | 0.48758501 | 1 | 0.3923910 | -0.01232148 |
| **qb8** | 0.1660830 | 0.3021902 | 0.3923910 | 1 | -0.14657797 |
| **qb9** | 0.22869773 | 0.13889897 | -0.01232148 | -0.14657797 | 1 |

Group 3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **qb4\_3** | **qb4\_5** | **qb7** | **qb8** | **qb9** |
| **qb4\_3** | 1 | 0.07469395 | 0.4400537 | 0.5437509 | 0.5195602 |
| **qb4\_5** | 0.07469395 | 1 | -0.1859863 | -0.1112473 | -0.1150310 |
| **qb7** | 0.44005368 | -0.18598635 | 1 | 0.4965637 | 0.4023176 |
| **qb8** | 0.54375089 | -0.11124727 | 0.4965637 | 1 | 0.5111735 |
| **qb9** | 0.51956019 | -0.11503102 | 0.4023176 | 0.5111735 | 1 |

Group 4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **qb4\_3** | **qb4\_5** | **qb7** | **qb8** | **qb9** |
| **qb4\_3** | 1 | 0.9001736 | -0.2393692 | -0.2281035 | -0.2411191 |
| **qb4\_5** | 0.9001736 | 1 | -0.1607610 | -0.1477551 | -0.1715444 |
| **qb7** | -0.2393692 | -0.1607610 | 1 | 0.8869407 | 0.8548438 |
| **qb8** | -0.2281035 | -0.1477551 | 0.8869407 | 1 | 0.8476377 |
| **qb9** | -0.2411191 | -0.1715444 | 0.8548438 | 0.8476377 | 1 |

Group 5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **qb4\_3** | **qb4\_5** | **qb7** | **qb8** | **qb9** |
| **qb4\_3** | 1 | -0.1514635 | -0.05905457 | -0.03462609 | -0.1298436 |
| **qb4\_5** | -0.15146353 | 1 | 0.75879238 | 0.73159084 | 0.8669082 |
| **qb7** | -0.05905457 | 0.7587924 | 1 | 0.54913637 | 0.6663558 |
| **qb8** | -0.03462609 | 0.7315908 | 0.54913637 | 1 | 0.6540521 |
| **qb9** | -0.12984356 | 0.8669082 | 0.66635581 | 0.65405206 | 1 |

APPENDIX E. Grid of parameters and Hyperparameters for complete model

***Logistic Regression***

{'solver': ['newton-cg', 'lbfgs', 'liblinear'],

'penalty': ['l2'],

'C': [100, 10, 1.0, 0.1, 0.01]}

Complete model

LogisticRegression(C=0.1, class\_weight={0: 1.43631197, 1: 0.76700502}, 'solver': 'lbfgs', 'penalty': 'l2')

Model of those who has a high-risk perception

LogisticRegression(C=0.1, class\_weight={0: 1.56837406, 1: 0.73400044}, 'solver': 'lbfgs', 'penalty': 'l2')

Model with Low Risk Perception Observations

LogisticRegression(C=10, class\_weight={0: 0.92119565, 1: 1.09354839},solver='newton-cg', 'penalty': 'l2')

***Decision Tree***

{'criterion': ['gini', 'entropy'],

'max\_depth': range(1, 10),

'min\_samples\_split': range(0, 10),

'min\_samples\_leaf': range(1, 5)}

Complete model

DecisionTreeClassifier(ccp\_alpha=0.000993163342228487,class\_weight={0: 1.43631197, 1: 0.76700502},criterion='entropy', max\_depth=8, min\_samples\_split=4,random\_state=123)

DecisionTreeClassifier(class\_weight={0: 1.56739773, 1: 0.73421448}, max\_depth=7,min\_samples\_leaf=3, min\_samples\_split=6, ccp\_alpha=0.0022213271836104)

DecisionTreeClassifier(ccp\_alpha=0.009396010830688362,class\_weight={0: 0.92727273, 1: 1.08510638},criterion='entropy', max\_depth=5, min\_samples\_leaf=2,min\_samples\_split=8, random\_state=50)

***Random Forest***

{'n\_estimators': [50, 120, 190, 260, 330, 400],

'max\_depth': [5, 8, 15, 25, 30, None],

'min\_samples\_split': [2, 5, 10, 15, 100],

'min\_samples\_leaf': [1, 2, 5, 10],

'max\_features': ['auto', 'sqrt', 'log2'],

'bootstrap': [True, False]}

RandomForestClassifier(bootstrap=False,class\_weight={0: 1.43631197, 1: 0.76700502}, max\_depth=30, max\_features='log2', max\_leaf\_nodes=49,min\_samples\_leaf=10, min\_samples\_split=10, n\_estimators=136)

RandomForestClassifier(class\_weight={0: 1.56739773, 1: 0.73421448},max\_depth=25, min\_samples\_split=5, n\_estimators=150)

RandomForestClassifier(class\_weight={0: 0.9272727, 1: 1.08510638}, max\_depth=25,min\_samples\_leaf=5, n\_estimators=500)

***Gradient Boosting***

{'max\_depth': [3, 4, 5],

'learning\_rate': [0.1, 0.25, 0.5, 0.75, 1],

'n\_estimators': [50, 100, 150],

'gamma': [0.5, 1, 1.5, 2],

'min\_child\_weight': [1, 5, 10]}

XGBClassifier(n\_estimators= 100, class\_weight={0: 1.43631197, 1: 0.76700502}, gamma=0.5, max\_depth=4, min\_child\_weight=10, scale\_pos\_weight=0.75, learning\_rate=0.1)

XGBClassifier(n\_estimators= 50, class\_weight={ 1.56726931, 0.73424266}, gamma=1, max\_depth=5, min\_child\_weight=5, scale\_pos\_weight=0.5, learning\_rate=0.25)

XGBClassifier(n\_estimators= 50, class\_weight={ 0: 0.92727273 , 1:1.08510638}, gamma=0.5, max\_depth=5, min\_child\_weight=1, scale\_pos\_weight=0.75, learning\_rate=0.1)

***Appendix F. Comparison of Feature importance***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Logistic Regressio (log-odds)** | **Logistic Regression (log-ratio)** | **Decision Tree** | **Random Forest** | **Gradient Boosting** |
| cca\_1 | -0,209858 | 1.094700 | 0,000000 | 0,019062 | 0,012829 |
| cca\_2 | 0,738323 | 0,874998 | 0,000000 | 0,016596 | 0,007485 |
| cca\_3 | 0,090480 | 0,985721 | 0,000000 | 0,018206 | 0,008831 |
| cca\_4 | -0,693412 | 0,817540 | 0,000000 | 0,015214 | 0,006943 |
| cca\_5 | -0,317492 | 0,882812 | 0,000000 | 0,012256 | 0,004835 |
| cluster\_1 | -0,379284 | 0,613437 | 0,000000 | 0,041727 | 0,110380 |
| cluster\_2 | -0,207715 | 1.110870 | 0,363349 | 0,039542 | 0,000000 |
| country\_Austria | 0,376858 | 1.385805 | 0,000000 | 0,007761 | 0,015435 |
| country\_Belgium | -0,084771 | 0,946520 | 0,000000 | 0,008014 | 0,006064 |
| country\_Bulgaria | -0,110913 | 0,402731 | 0,000000 | 0,011924 | 0,032368 |
| country\_Croatia | 0,233174 | 0,684351 | 0,010332 | 0,007805 | 0,025696 |
| country\_Cyprus | 0,012143 | 0,720662 | 0,000000 | 0,004430 | 0,008818 |
| country\_Czech Republic | -0,063769 | 0,490561 | 0,013328 | 0,009673 | 0,032311 |
| country\_Denmark | -0,385178 | 1.262602 | 0,000000 | 0,006220 | 0,020020 |
| country\_Estonia | -0,417317 | 0,770702 | 0,000000 | 0,005026 | 0,012538 |
| country\_Finland | 0,033848 | 2.399747 | 0,043743 | 0,013164 | 0,031635 |
| country\_France | -0,133533 | 1.012217 | 0,000000 | 0,007010 | 0,000000 |
| country\_Germany | -0,260454 | 1.457698 | 0,011016 | 0,008913 | 0,012628 |
| country\_Greece | -0,909487 | 0,604072 | 0,014445 | 0,010817 | 0,023253 |
| country\_Hungary | -0,353149 | 1.066808 | 0,000000 | 0,006774 | 0,006763 |
| country\_Ireland | 0,105144 | 1.017430 | 0,000000 | 0,007397 | 0,008521 |
| country\_Italy | -0,327586 | 0,658812 | 0,000000 | 0,008272 | 0,019470 |
| country\_Latvia | -0,078007 | 0,537925 | 0,000000 | 0,008844 | 0,031471 |
| country\_Lithuania | -0,712205 | 0,680330 | 0,000000 | 0,007319 | 0,018772 |
| country\_Luxembourg | 0,326281 | 1.837222 | 0,000000 | 0,004957 | 0,011951 |
| country\_Malta | 0,001655 | 2.552832 | 0,032584 | 0,011038 | 0,023969 |
| country\_Netherlands | -0,087574 | 0,956347 | 0,000000 | 0,005708 | 0,000000 |
| country\_Poland | -0,129992 | 0,499868 | 0,000000 | 0,009214 | 0,030119 |
| country\_Portugal | -0,119521 | 2.172106 | 0,035880 | 0,011508 | 0,026177 |
| country\_Romania | 0,172588 | 0,281159 | 0,138376 | 0,022591 | 0,058919 |
| country\_Slovakia | 0,064671 | 1.368817 | 0,014299 | 0,008303 | 0,010795 |
| country\_Slovenia | 0,006813 | 2.092424 | 0,025906 | 0,016035 | 0,029955 |
| country\_Spain | -0,305527 | 1.929948 | 0,013195 | 0,011806 | 0,028085 |
| country\_Sweden | -0,054963 | 1.776633 | 0,011141 | 0,007111 | 0,017834 |
| country\_United Kingdom | -0,504062 | 0,727973 | 0,000000 | 0,007358 | 0,014201 |
| d10\_Man | 0,017280 | 0,736735 | 0,000000 | 0,017423 | 0,012519 |
| d10\_Woman | -0,098238 | 0,924958 | 0,000000 | 0,017610 | 0,000000 |
| d11 | -0,124643 | 1.001656 | 0,000000 | 0,117401 | 0,008397 |
| d1\_Centre | -0,052656 | 0,895017 | 0,000000 | 0,019336 | 0,006432 |
| d1\_Centre-letf | -0,488677 | 1.034428 | 0,000000 | 0,018209 | 0,012128 |
| d1\_Centre-right | 0,189972 | 0,887345 | 0,000000 | 0,014253 | 0,006708 |
| d1\_Left | -0,201456 | 1.006836 | 0,000000 | 0,010170 | 0,009450 |
| d1\_Not positionable | -0,080494 | 0,938221 | 0,000000 | 0,014547 | 0,005224 |
| d1\_Right | 0,657493 | 0,878102 | 0,000000 | 0,009807 | 0,006163 |
| d25\_Large town | -0,090355 | 0,911024 | 0,000000 | 0,018818 | 0,008761 |
| d25\_Rural area or village | 0,009865 | 0,913607 | 0,000000 | 0,019882 | 0,007375 |
| d25\_Small or middle sized town | 0,608254 | 0,810700 | 0,000000 | 0,020090 | 0,008478 |
| d25\_dk | -0,014382 | 1.009914 | 0,000000 | 0,000000 | 0,000000 |
| d63\_Refusal/Other | -0,256266 | 0,887628 | 0,000000 | 0,004359 | 0,000000 |
| d63\_The higher class of society | -0,056869 | 1.127653 | 0,000000 | 0,001199 | 0,000000 |
| d63\_The lower middle class of society | 0,574720 | 0,880021 | 0,000000 | 0,014478 | 0,004229 |
| d63\_The middle class of society | -0,093186 | 0,906433 | 0,000000 | 0,019116 | 0,007185 |
| d63\_The upper middle class of society | 0,775697 | 1.214980 | 0,000000 | 0,009316 | 0,010250 |
| d63\_The working class of society | 0,875363 | 0,702473 | 0,033201 | 0,018282 | 0,024077 |
| d7\_Partner | -0,127810 | 0,922660 | 0,000000 | 0,019254 | 0,010784 |
| d7\_Partner and children | 0,194728 | 1.074891 | 0,000000 | 0,017980 | 0,006382 |
| d7\_Refusal/Other | 0,120138 | 0,918723 | 0,000000 | 0,000656 | 0,005651 |
| d7\_Single | -0,044634 | 0,816348 | 0,000000 | 0,018027 | 0,014263 |
| d7\_Single with children | -0,119203 | 0,916151 | 0,000000 | 0,006349 | 0,000000 |
| d8\_16-19 years old | 0,072220 | 0,944718 | 0,000000 | 0,016040 | 0,004562 |
| d8\_20+ years old | -1.268834 | 1.209216 | 0,059630 | 0,020213 | 0,042662 |
| d8\_Refusal/dk | -0,620036 | 0,773936 | 0,000000 | 0,002582 | 0,009517 |
| d8\_Still studying | -0,202914 | 0,948706 | 0,000000 | 0,005767 | 0,000000 |
| d8\_Up to 15 years old | 0,937203 | 0,812438 | 0,000000 | 0,009676 | 0,007241 |
| qb2 | 0,313947 | 1.188376 | 0,179575 | 0,101564 | 0,052489 |

***HIGH RISK PERCEPTION***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Logistic Regressio (log-odds)** | **Logistic Regression (log-ratio)** | **Decision Tree** | **Random Forest** | **Gradient Boosting** |
| cca\_1 | 0,871945 | 1.160816 | 0,000000 | 0,021365 | 0,017443 |
| cca\_2 | 0,674228 | 0,877755849 | 0,000000 | 0,020861 | 0,008659 |
| cca\_3 | 0,052552 | 1.101637 | 0,000000 | 0,023717 | 0,008094 |
| cca\_4 | -0,011284 | 0,894653232 | 0,000000 | 0,018692 | 0,008399 |
| cca\_5 | -0,111319 | 0,992147437 | 0,000000 | 0,016508 | 0,006375 |
| cluster\_1 | 0,302820 | 0,70380212 | 0,000000 | 0,020606 | 0,072865 |
| cluster\_2 | -0,385593 | 1.415650 | 0,479606 | 0,022067 | 0,000000 |
| country\_Austria | -0,687568 | 1.527121 | 0,000000 | 0,007862 | 0,012066 |
| country\_Belgium | -0,191327 | 0,895993039 | 0,000000 | 0,010056 | 0,010948 |
| country\_Bulgaria | 0,131567 | 0,398998689 | 0,000000 | 0,010004 | 0,040672 |
| country\_Croatia | -0,444270 | 0,686748555 | 0,000000 | 0,009013 | 0,033704 |
| country\_Cyprus | -0,081006 | 0,722403396 | 0,000000 | 0,005639 | 0,006562 |
| country\_Czech Republic | -0,133467 | 0,502797324 | 0,000000 | 0,007923 | 0,028346 |
| country\_Denmark | 0,608179 | 1.138685 | 0,000000 | 0,006895 | 0,013247 |
| country\_Estonia | -0,375787 | 0,641291948 | 0,000000 | 0,005906 | 0,007428 |
| country\_Finland | -0,141899 | 2.449681 | 0,065555 | 0,009097 | 0,029050 |
| country\_France | -0,075103 | 1.094098 | 0,000000 | 0,009301 | 0,012242 |
| country\_Germany | -0,424495 | 1.353670 | 0,000000 | 0,009952 | 0,017283 |
| country\_Greece | 0,129874 | 0,680047108 | 0,000000 | 0,010254 | 0,021152 |
| country\_Hungary | -0,008101 | 1.096890 | 0,000000 | 0,009453 | 0,009158 |
| country\_Ireland | 0,187259 | 0,922188179 | 0,000000 | 0,009359 | 0,008543 |
| country\_Italy | -0,325172 | 0,654100034 | 0,000000 | 0,008348 | 0,031129 |
| country\_Latvia | -0,918797 | 0,552486363 | 0,000000 | 0,007263 | 0,027204 |
| country\_Lithuania | 0,895958 | 0,697040363 | 0,000000 | 0,008151 | 0,019618 |
| country\_Luxembourg | -0,109823 | 1.828238 | 0,000000 | 0,005526 | 0,018782 |
| country\_Malta | -0,010549 | 3.321828 | 0,083321 | 0,007927 | 0,030413 |
| country\_Netherlands | 0,049980 | 0,891017971 | 0,000000 | 0,007820 | 0,007267 |
| country\_Poland | 0,115581 | 0,514181038 | 0,000000 | 0,009137 | 0,037728 |
| country\_Portugal | 0,199826 | 2.391557 | 0,000000 | 0,009300 | 0,020809 |
| country\_Romania | 0,002008 | 0,294012064 | 0,215886 | 0,016822 | 0,065110 |
| country\_Slovakia | 0,092479 | 1.405489 | 0,000000 | 0,008383 | 0,011394 |
| country\_Slovenia | -0,049147 | 2.298981 | 0,067584 | 0,010354 | 0,030685 |
| country\_Spain | -0,119250 | 1.962518 | 0,000000 | 0,009518 | 0,027077 |
| country\_Sweden | 0,423384 | 1.837082 | 0,000000 | 0,007497 | 0,015841 |
| country\_United Kingdom | 0,089930 | 0,699261643 | 0,000000 | 0,008697 | 0,015792 |
| d10\_Man | -0,069786 | 0,887585663 | 0,000000 | 0,021851 | 0,011721 |
| d10\_Woman | 0,218749 | 1.122525 | 0,000000 | 0,021773 | 0,000000 |
| d11 | -0,007884 | 1.002010 | 0,000000 | 0,169988 | 0,009328 |
| d1\_Centre | -0,130387 | 0,989506082 | 0,000000 | 0,024838 | 0,007971 |
| d1\_Centre-letf | 0,832466 | 1.140614 | 0,000000 | 0,022215 | 0,013901 |
| d1\_Centre-right | -1.224134 | 0,867709065 | 0,000000 | 0,019467 | 0,010425 |
| d1\_Left | 1.200515 | 1.221190 | 0,000000 | 0,013801 | 0,008603 |
| d1\_Not positionable | -0,081722 | 0,95204112 | 0,000000 | 0,018297 | 0,009256 |
| d1\_Right | 0,340386 | 0,875056135 | 0,000000 | 0,012818 | 0,007727 |
| d25\_Large town | -0,016638 | 1.053957 | 0,000000 | 0,023648 | 0,007752 |
| d25\_Rural area or village | 0,096797 | 0,991932114 | 0,000000 | 0,024329 | 0,008468 |
| d25\_Small or middle sized town | -0,056988 | 0,927648257 | 0,000000 | 0,025911 | 0,006678 |
| d25\_dk | -0,351258 | 1.027350 | 0,000000 | 0,000109 | 0,000000 |
| d63\_Refusal/Other | 0,026982 | 0,919288034 | 0,000000 | 0,005680 | 0,007998 |
| d63\_The higher class of society | -0,665180 | 1.205939 | 0,000000 | 0,002002 | 0,000000 |
| d63\_The lower middle class of society | -0,279845 | 0,944605474 | 0,000000 | 0,017826 | 0,009212 |
| d63\_The middle class of society | 0,125863 | 0,950759703 | 0,000000 | 0,022926 | 0,007731 |
| d63\_The upper middle class of society | -0,593327 | 1.323864 | 0,000000 | 0,009712 | 0,014237 |
| d63\_The working class of society | -0,360912 | 0,755900959 | 0,000000 | 0,018019 | 0,019305 |
| d7\_Partner | 0,347589 | 0,983499725 | 0,000000 | 0,024984 | 0,005069 |
| d7\_Partner and children | -0,357730 | 1.134127 | 0,000000 | 0,023721 | 0,008515 |
| d7\_Refusal/Other | 0,280555 | 0,93326374 | 0,000000 | 0,001092 | 0,000000 |
| d7\_Single | -0,084156 | 0,921527674 | 0,000000 | 0,022939 | 0,010200 |
| d7\_Single with children | 0,037895 | 1.038623 | 0,000000 | 0,009661 | 0,009651 |
| d8\_16-19 years old | -0,050494 | 0,932592956 | 0,000000 | 0,016212 | 0,008454 |
| d8\_20+ years old | 0,603353 | 1.244519 | 0,088049 | 0,015743 | 0,047157 |
| d8\_Refusal/dk | -0,069067 | 1.051250 | 0,000000 | 0,003583 | 0,006065 |
| d8\_Still studying | 0,149123 | 0,988778962 | 0,000000 | 0,006516 | 0,006548 |
| d8\_Up to 15 years old | -0,115391 | 0,825862643 | 0,000000 | 0,011068 | 0,006944 |

***Low risk perception***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Logistic Regressio (log-odds)** | **Logistic Regression (log-ratio)** | **Decision Tree** | **Random Forest** | **Gradient Boosting** |
| cca\_1 | 0,017753 | 1.006476 | 0,000000 | 0,006200 | 0,007691 |
| cca\_2 | 0,167401 | 1.129289 | 0,000000 | 0,011028 | 0,015933 |
| cca\_3 | 0,398833 | 1.017911 | 0,000000 | 0,006102 | 0,006277 |
| cca\_4 | -0,405617 | 1.027453 | 0,000000 | 0,005584 | 0,012052 |
| cca\_5 | -0,561250 | 0,841240 | 0,000000 | 0,007494 | 0,014207 |
| cluster\_1 | -0,256186 | 0,671103 | 0,426276 | 0,107565 | 0,057859 |
| cluster\_2 | 0,735099 | 1.490084 | 0,000000 | 0,062236 | 0,000000 |
| country\_Austria | -0,398833 | 0,979002 | 0,000000 | 0,017888 | 0,000000 |
| country\_Belgium | 0,080522 | 1.023511 | 0,000000 | 0,007876 | 0,000000 |
| country\_Bulgaria | 0,052831 | 0,272859 | 0,000000 | 0,009870 | 0,021334 |
| country\_Croatia | 1.883838 | 0,731877 | 0,000000 | 0,017007 | 0,017186 |
| country\_Cyprus | 0,387567 | 1.162884 | 0,000000 | 0,004371 | 0,021824 |
| country\_Czech Republic | -0,164886 | 0,329505 | 0,000000 | 0,038408 | 0,034103 |
| country\_Denmark | -0,452506 | 1.650174 | 0,000000 | 0,017159 | 0,020530 |
| country\_Estonia | -0,552504 | 0,575507 | 0,000000 | 0,004880 | 0,013605 |
| country\_Finland | 0,096397 | 2.732727 | 0,213650 | 0,040581 | 0,027986 |
| country\_France | -0,210631 | 0,636032 | 0,000000 | 0,007886 | 0,012712 |
| country\_Germany | -0,389685 | 1.455285 | 0,000000 | 0,010843 | 0,019017 |
| country\_Greece | -0,108819 | 0,248032 | 0,000000 | 0,004275 | 0,026235 |
| country\_Hungary | 0,469477 | 0,810073 | 0,000000 | 0,008655 | 0,000000 |
| country\_Ireland | 0,408612 | 0.810073 | 0,000000 | 0,003904 | 0,010234 |
| country\_Italy | 1.005300 | 0,603315 | 0,000000 | 0,011286 | 0,016200 |
| country\_Latvia | -0,021222 | 0,409695 | 0,000000 | 0,027013 | 0,029190 |
| country\_Lithuania | -1.727335 | 0,773998 | 0,000000 | 0,012872 | 0,030785 |
| country\_Luxembourg | -1.298802 | 2.085688 | 0,000000 | 0,007529 | 0,014170 |
| country\_Malta | 0,010044 | 7.249611 | 0,000000 | 0,000000 | 0,000000 |
| country\_Netherlands | 0,883120 | 0,791134 | 0,000000 | 0,008268 | 0,012930 |
| country\_Poland | 1.980948 | 0,677270 | 0,000000 | 0,003947 | 0,000000 |
| country\_Portugal | 0,120858 | 6.578703 | 0,000000 | 0,032986 | 0,026622 |
| country\_Romania | -0,004143 | 0,177757 | 0,360074 | 0,093431 | 0,059134 |
| country\_Slovakia | -1.110164 | 1.771408 | 0,000000 | 0,009930 | 0,014453 |
| country\_Slovenia | 0,023239 | 2.418433 | 0,000000 | 0,028926 | 0,036506 |
| country\_Spain | 0,104157 | 3.374126 | 0,000000 | 0,005975 | 0,010503 |
| country\_Sweden | -1.394199 | 1.599158 | 0,000000 | 0,007808 | 0,010157 |
| country\_United Kingdom | 0,500881 | 0,570496 | 0,000000 | 0,006896 | 0,022480 |
| d10\_Man | -0,015008 | 0,908103 | 0,000000 | 0,006237 | 0,015099 |
| d10\_Woman | -0,033445 | 1.101196 | 0,000000 | 0,005818 | 0,000000 |
| d11 | 0,121588 | 1.010094 | 0,000000 | 0,108500 | 0,018443 |
| d1\_Centre | -1.058548 | 1.128465 | 0,000000 | 0,007556 | 0,012482 |
| d1\_Centre-letf | 0,027082 | 0,995865 | 0,000000 | 0,004108 | 0,011041 |
| d1\_Centre-right | -0,049085 | 0,896893 | 0,000000 | 0,006651 | 0,019590 |
| d1\_Left | -0,002212 | 1.054252 | 0,000000 | 0,007751 | 0,000000 |
| d1\_Not positionable | 0,670363 | 1.109775 | 0,000000 | 0,004947 | 0,022822 |
| d1\_Right | 0,219321 | 0,847991 | 0,000000 | 0,005855 | 0,034645 |
| d25\_Large town | 0,375201 | 1.620177 | 0,000000 | 0,009755 | 0,010110 |
| d25\_Rural area or village | -0,505315 | 1.504727 | 0,000000 | 0,004085 | 0,011965 |
| d25\_Small or middle sized town | 0,162582 | 1.182228 | 0,000000 | 0,010400 | 0,011851 |
| d25\_dk | -0,096398 | 0,346959 | 0,000000 | 0,000000 | 0,000000 |
| d63\_Refusal/Other | 0,077097 | 0,952101 | 0,000000 | 0,004815 | 0,016932 |
| d63\_The higher class of society | -0,312142 | 1.954947 | 0,000000 | 0,004068 | 0,005841 |
| d63\_The lower middle class of society | 0,006455 | 0,985104 | 0,000000 | 0,008583 | 0,009595 |
| d63\_The middle class of society | 0,250209 | 0,967108 | 0,000000 | 0,006989 | 0,012307 |
| d63\_The upper middle class of society | 0,482535 | 0,866171 | 0,000000 | 0,005473 | 0,014925 |
| d63\_The working class of society | 0,122655 | 0,651060 | 0,000000 | 0,019158 | 0,022833 |
| d7\_Partner | -0,172879 | 0,997790 | 0,000000 | 0,007793 | 0,011527 |
| d7\_Partner and children | -0,429153 | 1.080147 | 0,000000 | 0,007664 | 0,027435 |
| d7\_Refusal/Other | 0,571775 | 1.284294 | 0,000000 | 0,000113 | 0,000000 |
| d7\_Single | -0,892342 | 0,666566 | 0,000000 | 0,027964 | 0,027686 |
| d7\_Single with children | -0,143673 | 1.083853 | 0,000000 | 0,004817 | 0,011125 |
| d8\_16-19 years old | -0,625210 | 1.176544 | 0,000000 | 0,006233 | 0,009601 |
| d8\_20+ years old | 1.216136 | 1.245231 | 0,000000 | 0,032899 | 0,024464 |
| d8\_Refusal/dk | 0,150903 | 0,535149 | 0,000000 | 0,008131 | 0,009971 |
| d8\_Still studying | -0,234288 | 1.473391 | 0,000000 | 0,011274 | 0,005794 |
| d8\_Up to 15 years old | -0,144260 | 0,865663 | 0,000000 | 0,005683 | 0,000000 |

***Summary composition dataset 1***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Obs.** | **Mean** | **Std. Dev.** | **Min** | **Max** |
| **Qb5** | 19072 |  |  |  |  |
| *Yes* | *12988* |  |  |  |  |
| *No* | *6084* |  |  |  |  |
| **Qb2** | 19072 | 8.53 | 1.35 | 6 | 10 |
| **Qb4\_3** | 19072 | 1.69 | 0.68 | 1 | 4 |
| **Qb4\_5** | 19072 | 1.87 | 0.87 | 1 | 4 |
| **Qb7** | 19072 | 1.46 | 0.60 | 1 | 4 |
| **Qb8** | 19072 | 1.50 | 0.63 | 1 | 4 |
| **Qb9** | 19072 | 1.45 | 0.58 | 1 | 4 |
| **D1** | 19072 |  |  |  |  |
| *Left* | *1617* |  |  |  |  |
| *Centre-left* | *5210* |  |  |  |  |
| *Centre* | *7968* |  |  |  |  |
| *Centre-right* | *2972* |  |  |  |  |
| *Right* | *1383* |  |  |  |  |
| *Not positionable* | *2867* |  |  |  |  |
| **D7** | 19072 |  |  |  |  |
| *Partner* | *6704* |  |  |  |  |
| *Partner and children* | *6204* |  |  |  |  |
| *Single* | *5116* |  |  |  |  |
| *Single with children* | *967* |  |  |  |  |
| *Refusal/Other* | *81* |  |  |  |  |
| **D8** | 19072 |  |  |  |  |
| *Still studying* | *1242* |  |  |  |  |
| *Up to 15 years old* | *2268* |  |  |  |  |
| *16-19 years old* | *7988* |  |  |  |  |
| *20+ years old* | *7323* |  |  |  |  |
| *Refusal/dk* | *251* |  |  |  |  |
| **D10** | 19072 |  |  |  |  |
| *Man* | *8968* |  |  |  |  |
| *Woman* | *10104* |  |  |  |  |
| **D11** | 19072 | 50.32 | 17.80 | 15 | 98 |
| **D25** | 19072 |  |  |  |  |
| *Rural area or village* | *6075* |  |  |  |  |
| *Small or middle sized town* | *7422* |  |  |  |  |
| *Large town* | *5572* |  |  |  |  |
| *Dk* | *3* |  |  |  |  |
| **D63** | 19072 |  |  |  |  |
| *The higher class of society* | *134* |  |  |  |  |
| *The lower middle class of society* | *3007* |  |  |  |  |
| *The middle class of society* | *9520* |  |  |  |  |
| *The upper middle class of society* | *1400* |  |  |  |  |
| *The working class of society* | *4561* |  |  |  |  |
| *Refusal/Other* | *450* |  |  |  |  |
| **CCA** | *19072* |  |  |  |  |
| *1* | *5144* |  |  |  |  |
| *2* | *3697* |  |  |  |  |
| *3* | *5296* |  |  |  |  |
| *4* | *2823* |  |  |  |  |
| *5* | *2112* |  |  |  |  |
| **CLUSTER** | 19072 |  |  |  |  |
| *1* | *8982* |  |  |  |  |
| *2* | *10090* |  |  |  |  |

***Summary composition dataset 2***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Obs.** | **Mean** | **Std. Dev.** | **Min** | **Max** |
| **Qb5** | 2906 |  |  |  |  |
| *Yes* | *1339* |  |  |  |  |
| *No* | *1567* |  |  |  |  |
| **Qb2** | 2906 | 4.06 | 1.31 | 1 | 5 |
| **Qb4\_3** | 2906 | 2.08 | 0.82 | 1 | 4 |
| **Qb4\_5** | 2906 | 2.15 | 0.88 | 1 | 4 |
| **Qb7** | 2906 | 1.94 | 0.83 | 1 | 4 |
| **Qb8** | 2906 | 1.95 | 0.83 | 1 | 4 |
| **Qb9** | 2906 | 1.84 | 0.78 | 1 | 4 |
| **D1** | 2906 |  |  |  |  |
| *Left* | *236* |  |  |  |  |
| *Centre-left* | *692* |  |  |  |  |
| *Centre* | *899* |  |  |  |  |
| *Centre-right* | *498* |  |  |  |  |
| *Right* | *220* |  |  |  |  |
| *Not positionable* | *361* |  |  |  |  |
| **D7** | 2906 |  |  |  |  |
| *Partner* | *1087* |  |  |  |  |
| *Partner and children* | *796* |  |  |  |  |
| *Single* | *859* |  |  |  |  |
| *Single with children* | *153* |  |  |  |  |
| *Refusal/Other* | *11* |  |  |  |  |
| **D8** | 2906 |  |  |  |  |
| *Still studying* | *163* |  |  |  |  |
| *Up to 15 years old* | *330* |  |  |  |  |
| *16-19 years old* | *1370* |  |  |  |  |
| *20+ years old* | *975* |  |  |  |  |
| *Refusal/dk* | *68* |  |  |  |  |
| **D10** | 2906 |  |  |  |  |
| *Man* | *1559* |  |  |  |  |
| *Woman* | *1347* |  |  |  |  |
| **D11** | 2906 | 51.75 | 18.32 | 15 | 95 |
| **D25** | 2906 |  |  |  |  |
| *Rural area or village* | *993* |  |  |  |  |
| *Small or middle sized town* | *1088* |  |  |  |  |
| *Large town* | *824* |  |  |  |  |
| *Dk* | *1* |  |  |  |  |
| **D63** | 2906 |  |  |  |  |
| *The higher class of society* | *20* |  |  |  |  |
| *The lower middle class of society* | *449* |  |  |  |  |
| *The middle class of society* | *1422* |  |  |  |  |
| *The upper middle class of society* | *230* |  |  |  |  |
| *The working class of society* | *715* |  |  |  |  |
| *Refusal/Other* | *70* |  |  |  |  |
| **CCA** | 2906 |  |  |  |  |
| *1* | *560* |  |  |  |  |
| *2* | *718* |  |  |  |  |
| *3* | *810* |  |  |  |  |
| *4* | *448* |  |  |  |  |
| *5* | *370* |  |  |  |  |
| **CLUSTER** | 2906 |  |  |  |  |
| *1* | *2189* |  |  |  |  |
| *2* | *717* |  |  |  |  |

1. See appendix D for the correlation of each class separately. [↑](#footnote-ref-1)
2. See appendix for the grids of parameters and the hyperparamenters for each algorithm. [↑](#footnote-ref-2)
3. I try also a simple model with all the variables and the model with the adding of interactions (gender and education, social class and education, risk perception and education). The interactions do not present statistically significant, therefore the simple model is chosen. [↑](#footnote-ref-3)
4. T I remember that these coefficients are in log-odds terms he interpretation of log-odds terms is not so easy. For this reason, we convert the log-odds term into odds ratio, which means the probability of an event occurring. When the odds ra tio is greater than 1, it shows a positive relationship. If an odds ratio less than 1 means a negative relationship. See appendix for more details about log-odd table. [↑](#footnote-ref-4)