**Analysis**

The following section illustrates the different steps undertaken to obtain a prediction model for pro-environmental action. In particular, the first step consists of Exploratory Data Analysis, to investigate climate change attitudes. Then, the best fitting models tested to predict the final price are presented.

**Exploratory Data Analysis: 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 1, 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 1: Single Most Serious Problem per Country

Another interesting example is the difference in the climate change risk perception. As you can see in figure 2, about half of the citizens of Malta and Luxemburg declared that they are extremely worried about the phenomenon studied.



Figure 2: Climate Change Risk Perception per Country



Figure 3: Pro-environemntal 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 or temperature growth regarding natural hazards and the “level of environmentalism in the political arena of a given country” (Echavarren et al., 2019, p. 815) for political variables. 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 method 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 to use the same data of year of the survey) (Yale Center for Environmental Law & Policy, 2020). EPI quantifies numerically environmental health and ecosystem vitality around the world. 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 3 shows the score across European Union (EU). The best score is obtained from Denmark, while the worst from Bulgaria.

Figure 4: 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 4 shows.

According to scholars (Echavarren et al., 2019; van der Linden, 2015) socio-cultural context influces individual attitudes towards climate change concerns. Therefore, the notable diferencess in attitudes across coutries 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 5:The 2019 Climate Change Policy

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

**Analysis: Data Exploratory**

**3.1 Correlation Class Analysis**



Figure 6: CCA

The first step is to perform CCA, a method for partitioning the dataset into classes to apprehend unobserved heterogeneity (Rossoni et al., 2021). For our dataset, the CCA algorithm produces five groups. Figure 5 illustrates the individual modules or classes created as a network. Each node corresponds to one item, while the edges show 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. Group 2 obtains a high positive correlation (0.8) mainly between qb4\_3 and qb4\_5. Group 3 has a slight correlation among all variables, except for qb4\_5. In group 4 we find a strong positive correlation among qb7, qb8, and qb9 and separately between qb4\_3 and qb4\_5. 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 7: 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 meanings among citizens of the same country. We 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, if we look at the composition for each country, we find that over 30% of citizens from Cyprus, Ireland, Malta, Portugal, Slovakia, Spain, and 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, we 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.

**3.2 Partitioning Around Medoids Clustering**

PAM clustering is an unsupervised method that looks for patterns without any knowledge of the classification purpose. 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. We remember that also in this case the questions about climate change (qb4\_3, qb4\_5, qb7, qb8, qb9) are fitted in that algorithm.

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 8: 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 individuals belong to cluster 1, and 10807 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), we can see that 2 clusters resemble two different types of green-identity, which we call “moderate green identity”, “extreme green-identity”.

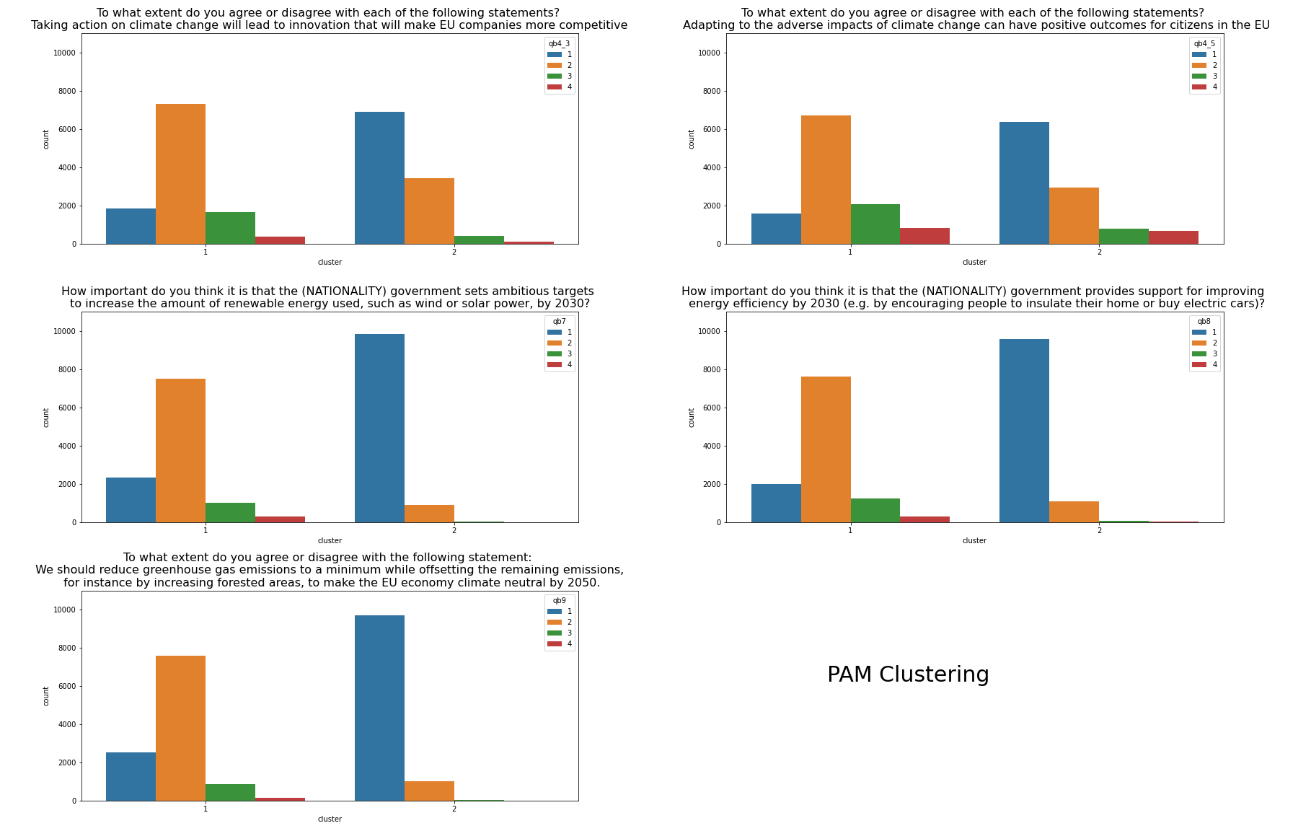


Figure 9: Summary of PAM Clustering

**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.

**Prediction**

As mentioned in the literature review, 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 behaviour of citizens in a dummy outcome.

**Evaluating Classification Models**

Before proceeding with the analysis, it is important to first explain the ways the errors will be measured, since one of the main factors according to which we will choose the best model is the minimization of the error. We consider two measurements of error:

* **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**: combine precision and recall, where the precision is the number of true positives divided by the total number of element labeled as positive, and the recall is the number of true positives divided by the total number of elements that belong to the positive class (Shmueli, 2019). Macro F1-score is computed harmonic mean, as shown in the formula:

The range of the score is 0-1, where 0 is the worst score and 1 is the best score. 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, we decide to use a first compromise: the balance between accuracy and f1-score macro average. 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, we explain how to understand what variables have a fundamental role in our models. We compute feature importance by the different algorithms in scikit-learn. Logistic regression finds a set of coefficients to use in the weighted sum to make a prediction. These coefficients can be used directly as a crude type of 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 feature is used to split a node. Mean Decrease Accuracy measures how much accuracy the model losses by excluding each variable. However, scikit-learn implemented Gini importance, also in our analysis, it is adopted.

**Prediction**

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) We 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.

We 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

Table 1: Metrics comparison

|  |  |  |
| --- | --- | --- |
| **Models** | **Accuracy** | **Macro-f1 score** |
| **Logistic Regression** | 0.67 | 0.65 |
| **Decision Tree** | 0.64 | 0.62 |
| **Random Forest** | 0.70 | 0.67 |
| **Gradient Boosting** | 0.70 | 0.66 |

**Logistic regression**

The first method explored is Logistic regression. This algorithm supposes that all the assumptions (~~independence of errors~~, 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.

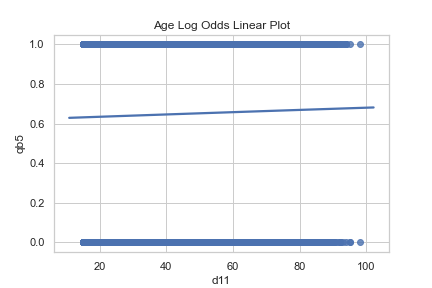
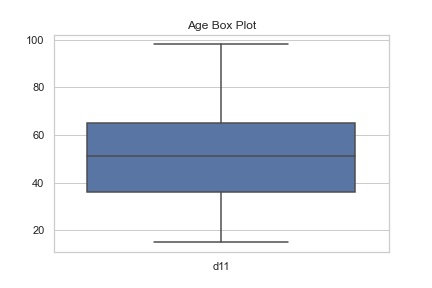
First, we check the assumption of linearity in the logit for continuous independent variables, also in this case only for the continuous variables. In this case, only age is continuous, and it does not verify this assumption, as figure 10 displays (there should be an “s” curve line). Second, we can not 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 11: Linearity in the Logit for Age

Figure 10: Lack of outliers for Age

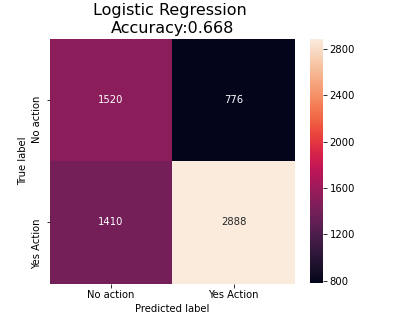
The fitted model, with the best tuning parameters, has 0.67 of accuracy and 0.66 of macro f1-score. As shown by figure 12,

Figure 12: Logistic Rigression- Confusion Matrix

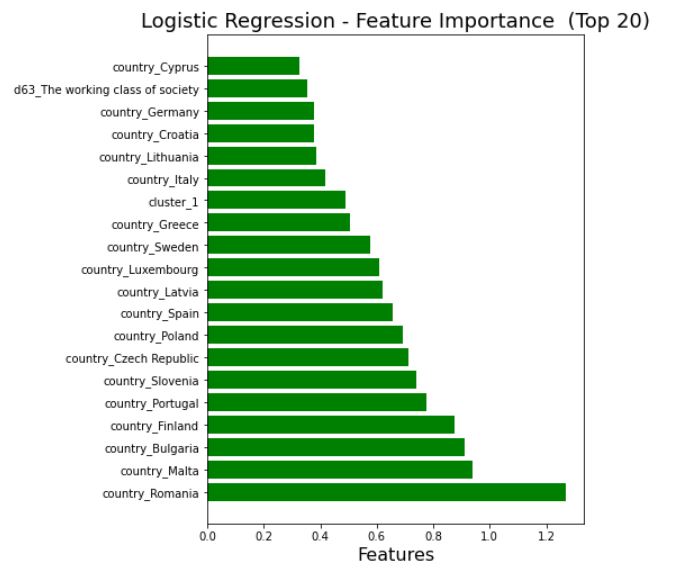


Figure 13: Logistic Regression- Feature Importance

**Decision Tree**

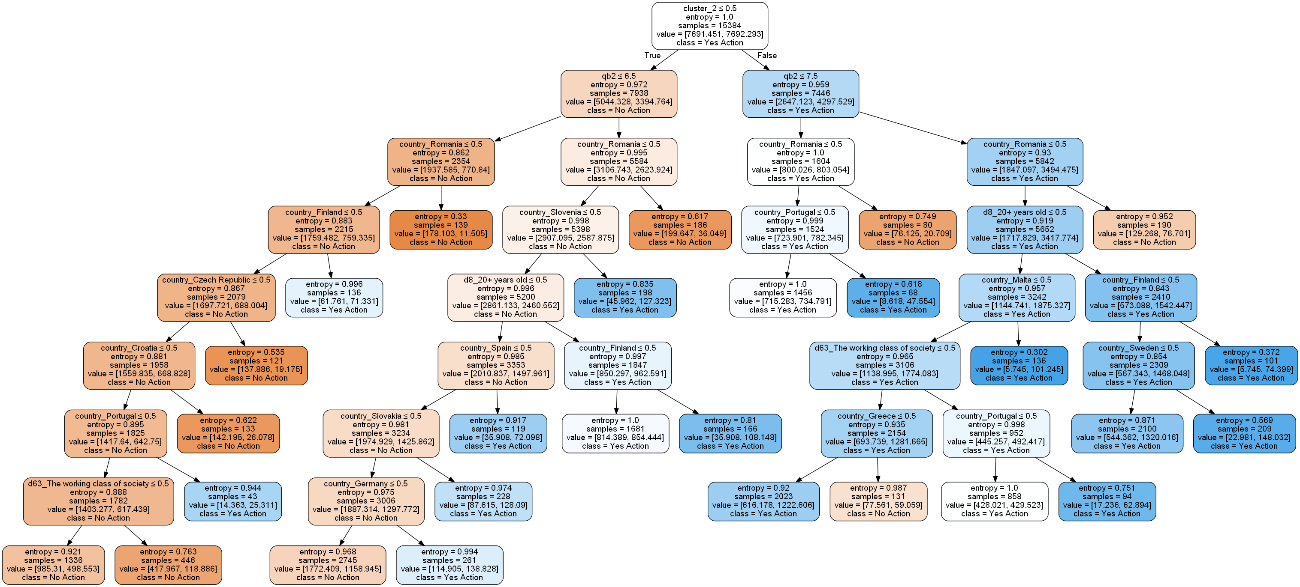


Figure 14: Decision Tree

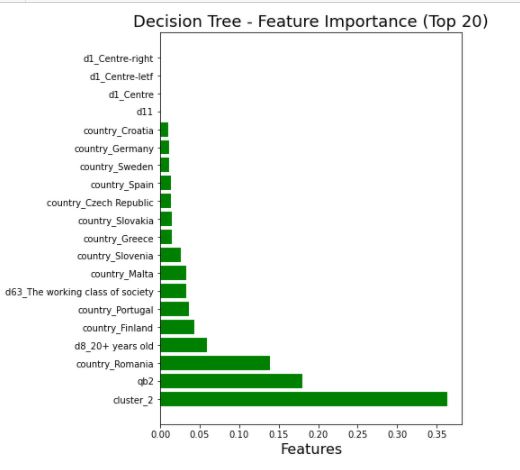


Figure 15: Decision Tree - Feature Importance

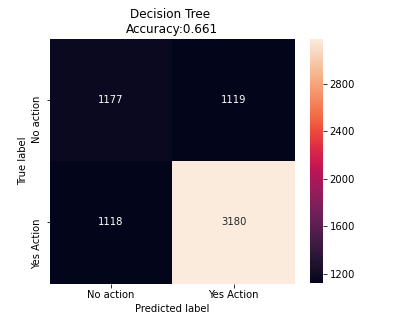
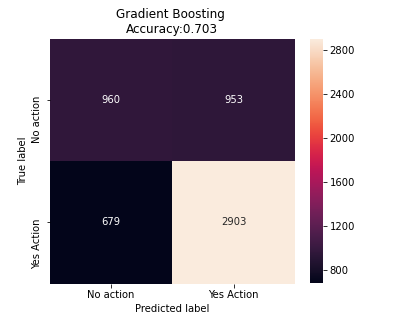
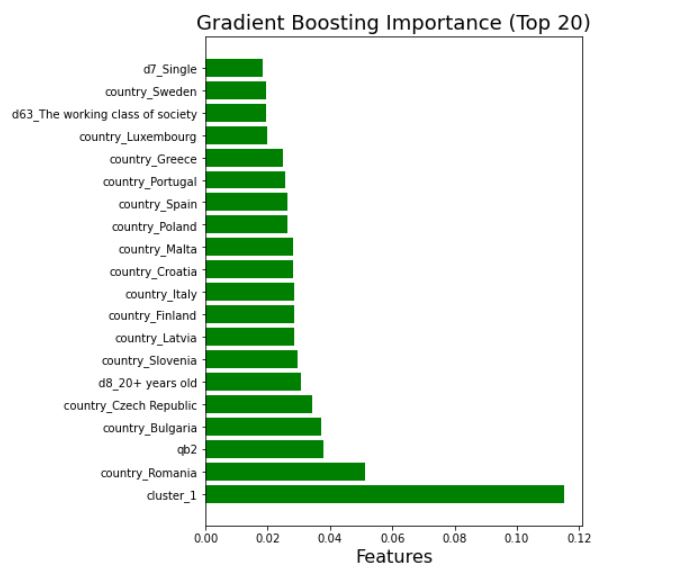


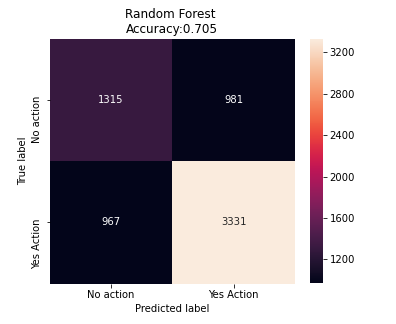
Figure 16: Decision Tree- Confusion Matrix

**Random Forest**

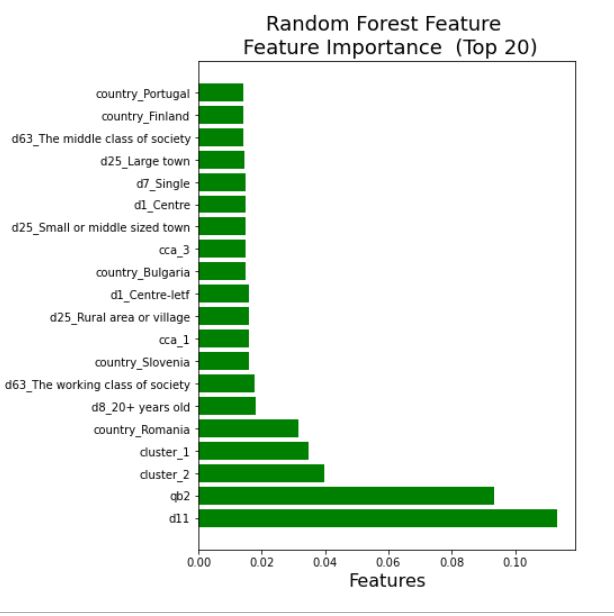
**Gradient Boosting**



The Random Forest algorithm has yielded the best performance when compared to other classifiers, as shown in the table. Random Forest gets a better result both in accuracy (0.70) and in macro f1-score average (0.67).



The confusion matrix, figure, displays the quality of the output of the classifier. It shows the counts from predicted and actual values. In this case, the count is done with the test set, showing if the predictions math the reality and therefore are accurately classify or not (Battiti & Brunato, 2014). We can see that 1315 observations, for the no-action class, and 3331 for the yes-action class, are correctly classified. Instead, there are for each class about 900 cases of misclassification.



As shown in figure x, analysing the feature importances of the model, we see that age and climate risk perception are the most important variable to classify and predict pro-environmental behaviour. The interpretation of feature importance is not simple and immediate. We expect that most young people take place pro-environmental action. However, this relationship is not so clear. If we 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. Instead, climate change risk perception is more interesting to analyse. 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 we analyse the percentage of those who take place action within the level of perceived risk, we 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, as figure x shows. Other important features are the created clusters, which correspond to the different green-identity. Cluster 2 is composed of 70% of individuals who take place action, while cluster 1 by 55%. Therefore if an individual belongs to cluster 2 it is more likely that he/she does something for the environment. Being of Romania is a discriminant to predict behaviour, since 66% of them declare do no do anything, as we have already explained. Lastly, a higher level of educated and the working class are other two important variables, even if much less than age and risk perception. 72% of those who have a high level of education, or better have stopped to study after 20 years, has to take place pro-environmental action. The average of other classes of this variable is 60%. The opposite case happens with working-class where 40% of them does not anything, instead of other social class where the proportion is lower.

In the second part of the analysis we 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, we call the first subset “with high-risk Perception observations” and the second one “with low-risk perception observations”. At this point of the analysis we can confirm the hypothesis suggest from the literature: climate risk perception is one of the main factors for predicting pro-environmental behaviour. Now, we want to understand what are the most important predictors if we change the level of risk perception. Also in this part, we 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)
* Maritial status (d7)
* Stopped full-time education (d8)
* Gender (d10)
* Age (d11)
* Residence (d25)
* Class identity (d63)
* Country

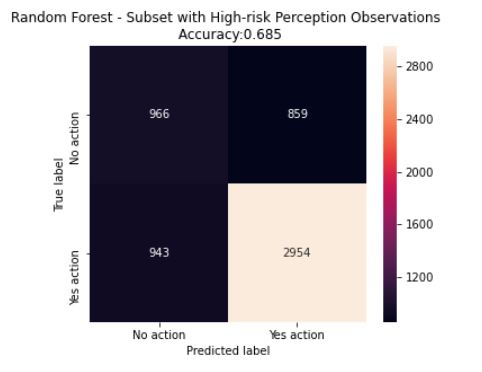
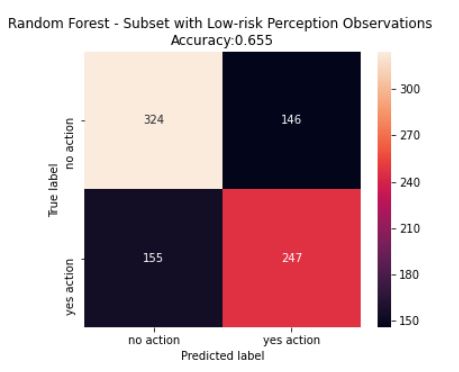
Table 2:Metrics comparison- Subsert with High-Risk Perception observations

|  |  |  |
| --- | --- | --- |
| **Models** | **Accuracy** | **Macro-f1 score** |
| **Logistic Regression** | 0.65 | 0.63 |
| **Decision Tree** | 0.63 | 0.60 |
| **Random Forest** | 0.69 | 0.64 |
| **Gradient Boosting** | 0.67 | 0.64 |

Table 3: Metrics comparison-Subset with low-risk perception observations

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

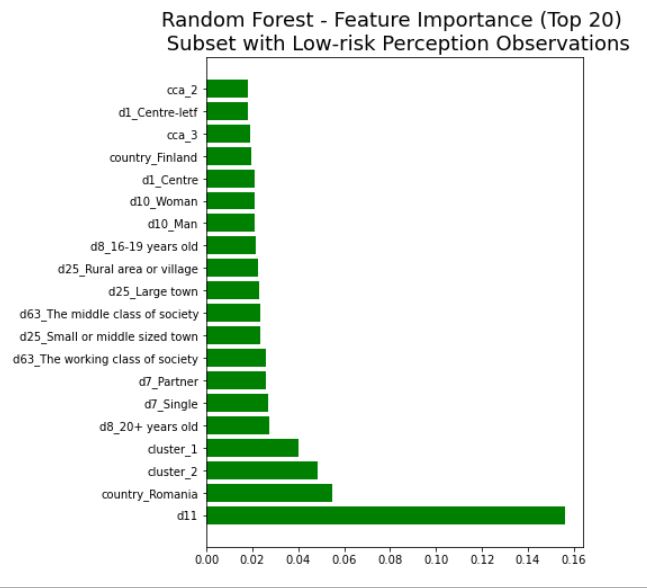
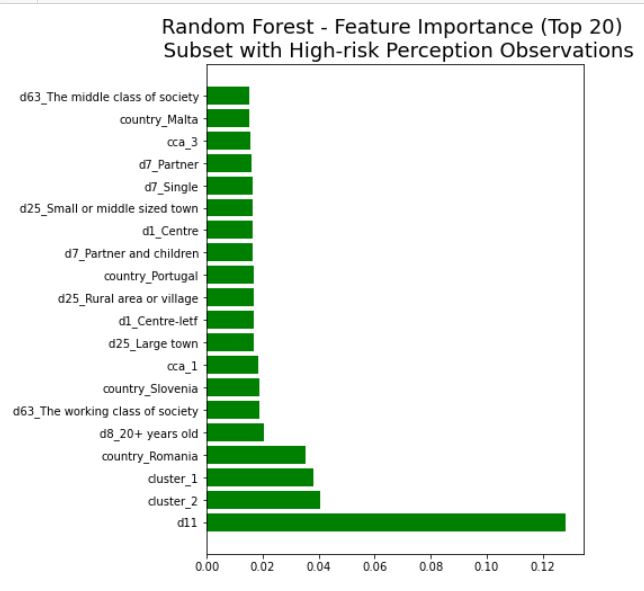
In both cases, Random Forest has yielded the best performance when compared to other classifiers, as shown in table 2 and 3.



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

Figure 17: Comparison of confusion matrices- Random Forest

Figure 17 displays the comparison of confusion matrices of the two different models. Both models classify and predict well the outputs. Subset with high-risk perception model mistakes no action class, 966 cases are classified correctly, while 859 are misclassified as yes action.



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

Figure 18: Comparison of variables importance- Random Forest

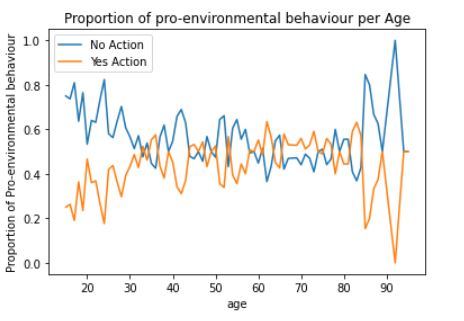


Figure 19: Subset with low-risk perception

Figure 18 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 does eco-friendly something. Another time, this percentage drops from 80 years old. More interesting is the trend in the second subset, as figure 19 shows. On average 70% of younger has a negative behaviour in favor of the environment. This percentage drop with the increase of years.

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., 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]}

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

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

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)}

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': [100, 300, 500, 800, 1200],

'max\_depth': [5, 8, 15, 25, 30],

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

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

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. Feature importance***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Logistic Regression | Decision Tree | Random Forest | Gradient Boosting |
| cca\_1 | -0,209858 | 0,005738 | 0,015648 | 0,017856 |
| cca\_2 | 0,738323 | 0,038062 | 0,014081 | 0,006258 |
| cca\_3 | 0,090480 | 0,000000 | 0,015214 | 0,007451 |
| cca\_4 | -0,693412 | 0,014549 | 0,014080 | 0,008575 |
| cca\_5 | -0,317492 | 0,000000 | 0,013144 | 0,005238 |
| cluster\_1 | -0,379284 | 0,000000 | 0,038189 | 0,115195 |
| cluster\_2 | -0,207715 | 0,104149 | 0,037514 | 0,000000 |
| country\_Austria | 0,376858 | 0,024019 | 0,009704 | 0,015863 |
| country\_Belgium | -0,084771 | 0,000000 | 0,010664 | 0,012057 |
| country\_Bulgaria | -0,110913 | 0,025748 | 0,015110 | 0,037226 |
| country\_Croatia | 0,233174 | 0,013107 | 0,010962 | 0,028162 |
| country\_Cyprus | 0,012143 | 0,019097 | 0,006560 | 0,010000 |
| country\_Czech Republic | -0,063769 | 0,020431 | 0,010890 | 0,034470 |
| country\_Denmark | -0,385178 | 0,000000 | 0,007748 | 0,013575 |
| country\_Estonia | -0,417317 | 0,000000 | 0,008352 | 0,014401 |
| country\_Finland | 0,033848 | 0,062730 | 0,014386 | 0,028387 |
| country\_France | -0,133533 | 0,013062 | 0,009995 | 0,005899 |
| country\_Germany | -0,260454 | 0,000000 | 0,012087 | 0,018161 |
| country\_Greece | -0,909487 | 0,014254 | 0,013101 | 0,025033 |
| country\_Hungary | -0,353149 | 0,000000 | 0,008771 | 0,010412 |
| country\_Ireland | 0,105144 | 0,008863 | 0,009948 | 0,000000 |
| country\_Italy | -0,327586 | 0,000000 | 0,011417 | 0,028379 |
| country\_Latvia | -0,078007 | 0,027252 | 0,011419 | 0,028502 |
| country\_Lithuania | -0,712205 | 0,000000 | 0,010228 | 0,013672 |
| country\_Luxembourg | 0,326281 | 0,000000 | 0,006469 | 0,019941 |
| country\_Malta | 0,001655 | 0,000000 | 0,011621 | 0,028014 |
| country\_Netherlands | -0,087574 | 0,000000 | 0,007622 | 0,007409 |
| country\_Poland | -0,129992 | 0,000000 | 0,013878 | 0,026524 |
| country\_Portugal | -0,119521 | 0,013340 | 0,014138 | 0,025619 |
| country\_Romania | 0,172588 | 0,075794 | 0,032519 | 0,051159 |
| country\_Slovakia | 0,064671 | 0,008874 | 0,010777 | 0,012134 |
| country\_Slovenia | 0,006813 | 0,039291 | 0,016329 | 0,029475 |
| country\_Spain | -0,305527 | 0,000000 | 0,012394 | 0,026427 |
| country\_Sweden | -0,054963 | 0,000000 | 0,010422 | 0,019353 |
| country\_United Kingdom | -0,504062 | 0,000000 | 0,009529 | 0,016439 |
| d10\_Man | 0,017280 | 0,000000 | 0,013190 | 0,011037 |
| d10\_Woman | -0,098238 | 0,023733 | 0,013209 | 0,000000 |
| d11 | -0,124643 | 0,198840 | 0,112385 | 0,006403 |
| d1\_Centre | -0,052656 | 0,012924 | 0,014548 | 0,004256 |
| d1\_Centre-letf | -0,488677 | 0,019199 | 0,015798 | 0,010665 |
| d1\_Centre-right | 0,189972 | 0,000000 | 0,013754 | 0,008571 |
| d1\_Left | -0,201456 | 0,012649 | 0,011952 | 0,011913 |
| d1\_Not positionable | -0,080494 | 0,015773 | 0,014026 | 0,003961 |
| d1\_Right | 0,657493 | 0,000000 | 0,011426 | 0,005466 |
| d25\_Large town | -0,090355 | 0,011508 | 0,014306 | 0,006583 |
| d25\_Rural area or village | 0,009865 | 0,029812 | 0,015756 | 0,005991 |
| d25\_Small or middle sized town | 0,608254 | 0,021807 | 0,014634 | 0,008433 |
| d25\_dk | -0,014382 | 0,000000 | 0.000018 | 0,000000 |
| d63\_Refusal/Other | -0,256266 | 0,000000 | 0,006437 | 0,000000 |
| d63\_The higher class of society | -0,056869 | 0,006224 | 0,002424 | 0,000000 |
| d63\_The lower middle class of society | 0,574720 | 0,020113 | 0,012840 | 0,005650 |
| d63\_The middle class of society | -0,093186 | 0,000000 | 0,014094 | 0,006356 |
| d63\_The upper middle class of society | 0,775697 | 0,000000 | 0,010179 | 0,017296 |
| d63\_The working class of society | 0,875363 | 0,026876 | 0,017572 | 0,019518 |
| d7\_Partner | -0,127810 | 0,000000 | 0,014846 | 0,003724 |
| d7\_Partner and children | 0,194728 | 0,006674 | 0,014226 | 0,008050 |
| d7\_Refusal/Other | 0,120138 | 0,000000 | 0,001502 | 0,000000 |
| d7\_Single | -0,044634 | 0,023816 | 0,014587 | 0,018327 |
| d7\_Single with children | -0,119203 | 0,009735 | 0,008850 | 0,000000 |
| d8\_16-19 years old | 0,072220 | 0,000000 | 0,012616 | 0,006520 |
| d8\_20+ years old | -1.268834 | 0,031956 | 0,018603 | 0,030646 |
| d8\_Refusal/dk | -0,620036 | 0,000000 | 0,004136 | 0,007697 |
| d8\_Still studying | -0,202914 | 0,000000 | 0,005876 | 0,000000 |
| d8\_Up to 15 years old | 0,937203 | 0,000000 | 0,009439 | 0,007628 |
| qb2 | -0,209858 | 0,005738 | 0,015648 | 0,017856 |

***HIGH RISK PERCEPTION***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Logistic Regression | Decision Tree | Random Forest | Gradient Boosting |
| cca\_1 | 0,871945 | 0,000000 | 0,020947 | 0,013126 |
| cca\_2 | 0,674228 | 0,000000 | 0,018335 | 0,009548 |
| cca\_3 | 0,052552 | 0,000000 | 0,021029 | 0,008298 |
| cca\_4 | -0,011284 | 0,000000 | 0,017162 | 0,008021 |
| cca\_5 | -0,111319 | 0,000000 | 0,016452 | 0,006832 |
| cluster\_1 | 0,302820 | 0,000000 | 0,030087 | 0,068876 |
| cluster\_2 | -0,385593 | 0,479606 | 0,028733 | 0,000000 |
| country\_Austria | -0,687568 | 0,000000 | 0,008939 | 0,017476 |
| country\_Belgium | -0,191327 | 0,000000 | 0,010312 | 0,012147 |
| country\_Bulgaria | 0,131567 | 0,000000 | 0,012754 | 0,035578 |
| country\_Croatia | -0,444270 | 0,000000 | 0,008983 | 0,022194 |
| country\_Cyprus | -0,081006 | 0,000000 | 0,005536 | 0,006703 |
| country\_Czech Republic | -0,133467 | 0,000000 | 0,010234 | 0,025127 |
| country\_Denmark | 0,608179 | 0,000000 | 0,006956 | 0,013925 |
| country\_Estonia | -0,375787 | 0,000000 | 0,007560 | 0,013816 |
| country\_Finland | -0,141899 | 0,065555 | 0,014074 | 0,036589 |
| country\_France | -0,075103 | 0,000000 | 0,009546 | 0,014224 |
| country\_Germany | -0,424495 | 0,000000 | 0,011372 | 0,017192 |
| country\_Greece | 0,129874 | 0,000000 | 0,012464 | 0,024240 |
| country\_Hungary | -0,008101 | 0,000000 | 0,009469 | 0,010452 |
| country\_Ireland | 0,187259 | 0,000000 | 0,008982 | 0,008954 |
| country\_Italy | -0,325172 | 0,000000 | 0,008788 | 0,018022 |
| country\_Latvia | -0,918797 | 0,000000 | 0,009537 | 0,022870 |
| country\_Lithuania | 0,895958 | 0,000000 | 0,009052 | 0,017361 |
| country\_Luxembourg | -0,109823 | 0,000000 | 0,007420 | 0,021888 |
| country\_Malta | -0,010549 | 0,083321 | 0,010992 | 0,042689 |
| country\_Netherlands | 0,049980 | 0,000000 | 0,006794 | 0,007619 |
| country\_Poland | 0,115581 | 0,000000 | 0,011808 | 0,029407 |
| country\_Portugal | 0,199826 | 0,000000 | 0,012355 | 0,029406 |
| country\_Romania | 0,002008 | 0,215886 | 0,025811 | 0,068904 |
| country\_Slovakia | 0,092479 | 0,000000 | 0,009060 | 0,015634 |
| country\_Slovenia | -0,049147 | 0,067584 | 0,012413 | 0,043463 |
| country\_Spain | -0,119250 | 0,000000 | 0,012564 | 0,030713 |
| country\_Sweden | 0,423384 | 0,000000 | 0,010024 | 0,022826 |
| country\_United Kingdom | 0,089930 | 0,000000 | 0,009475 | 0,012452 |
| d10\_Man | -0,069786 | 0,000000 | 0,019574 | 0,011217 |
| d10\_Woman | 0,218749 | 0,000000 | 0,019266 | 0,000000 |
| d11 | -0,007884 | 0,000000 | 0,141934 | 0,008264 |
| d1\_Centre | -0,130387 | 0,000000 | 0,020382 | 0,006683 |
| d1\_Centre-letf | 0,832466 | 0,000000 | 0,019408 | 0,014058 |
| d1\_Centre-right | -1.224134 | 0,000000 | 0,017829 | 0,008780 |
| d1\_Left | 1.200515 | 0,000000 | 0,013578 | 0,013073 |
| d1\_Not positionable | -0,081722 | 0,000000 | 0,017756 | 0,005463 |
| d1\_Right | 0,340386 | 0,000000 | 0,012323 | 0,008050 |
| d25\_Large town | -0,016638 | 0,000000 | 0,021275 | 0,008623 |
| d25\_Rural area or village | 0,096797 | 0,000000 | 0,021830 | 0,008283 |
| d25\_Small or middle sized town | -0,056988 | 0,000000 | 0,022412 | 0,007487 |
| d25\_dk | -0,351258 | 0,000000 | 0.000087 | 0,000000 |
| d63\_Refusal/Other | 0,026982 | 0,000000 | 0,006362 | 0,000000 |
| d63\_The higher class of society | -0,665180 | 0,000000 | 0,002262 | 0,007354 |
| d63\_The lower middle class of society | -0,279845 | 0,000000 | 0,017399 | 0,009677 |
| d63\_The middle class of society | 0,125863 | 0,000000 | 0,020621 | 0,006617 |
| d63\_The upper middle class of society | -0,593327 | 0,000000 | 0,011355 | 0,017373 |
| d63\_The working class of society | -0,360912 | 0,000000 | 0,019323 | 0,013882 |
| d7\_Partner | 0,347589 | 0,000000 | 0,021043 | 0,006860 |
| d7\_Partner and children | -0,357730 | 0,000000 | 0,020260 | 0,007528 |
| d7\_Refusal/Other | 0,280555 | 0,000000 | 0,001424 | 0,000000 |
| d7\_Single | -0,084156 | 0,000000 | 0,019984 | 0,008559 |
| d7\_Single with children | 0,037895 | 0,000000 | 0,009800 | 0,003185 |
| d8\_16-19 years old | -0,050494 | 0,000000 | 0,016461 | 0,008512 |
| d8\_20+ years old | 0,603353 | 0,088049 | 0,017852 | 0,030327 |
| d8\_Refusal/dk | -0,069067 | 0,000000 | 0,003787 | 0,005921 |
| d8\_Still studying | 0,149123 | 0,000000 | 0,007144 | 0,009490 |
| d8\_Up to 15 years old | -0,115391 | 0,000000 | 0,011254 | 0,008159 |
| qb2 | 0,871945 | 0,000000 | 0,020947 | 0,013126 |

***Low risk perception***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Logistic Regression | Decision Tree | Random Forest | Gradient Boosting |
| cca\_1 | 0,01775 | 0,00000 | 0,01704 | 0,01292 |
| cca\_2 | 0,16740 | 0,00000 | 0,01716 | 0,01419 |
| cca\_3 | 0,39883 | 0,00000 | 0,01959 | 0,00593 |
| cca\_4 | -0,40562 | 0,00000 | 0,01515 | 0,01599 |
| cca\_5 | -0,56125 | 0,00000 | 0,01337 | 0,01817 |
| cluster\_1 | -0,25619 | 0,42628 | 0,04535 | 0,03892 |
| cluster\_2 | 0,73510 | 0,00000 | 0,04278 | 0,00000 |
| country\_Austria | -0,39883 | 0,00000 | 0,01456 | 0,01775 |
| country\_Belgium | 0,08052 | 0,00000 | 0,00480 | 0,00979 |
| country\_Bulgaria | 0,05283 | 0,00000 | 0,00136 | 0,02487 |
| country\_Croatia | 1.883838 | 0,00000 | 0,00968 | 0,01124 |
| country\_Cyprus | 0,38757 | 0,00000 | 0,00110 | 0,00000 |
| country\_Czech Republic | -0,16489 | 0,00000 | 0,01535 | 0,03989 |
| country\_Denmark | -0,45251 | 0,00000 | 0,00594 | 0,02324 |
| country\_Estonia | -0,55250 | 0,00000 | 0,00706 | 0,01615 |
| country\_Finland | 0,09640 | 0,21365 | 0,01777 | 0,04182 |
| country\_France | -0,21063 | 0,00000 | 0,01007 | 0,01611 |
| country\_Germany | -0,38968 | 0,00000 | 0,00943 | 0,02066 |
| country\_Greece | -0,10882 | 0,00000 | 0,00288 | 0,02585 |
| country\_Hungary | 0,46948 | 0,00000 | 0,00248 | 0,01611 |
| country\_Ireland | 0,40861 | 0,00000 | 0,00496 | 0,00000 |
| country\_Italy | 1.005300 | 0,00000 | 0,00675 | 0,00955 |
| country\_Latvia | -0,02122 | 0,00000 | 0,01593 | 0,03708 |
| country\_Lithuania | -1.727335 | 0,00000 | 0,00582 | 0,01218 |
| country\_Luxembourg | -1.298802 | 0,00000 | 0,00473 | 0,01452 |
| country\_Malta | 0,01004 | 0,00000 | 0,00000 | 0,00000 |
| country\_Netherlands | 0,88312 | 0,00000 | 0,00530 | 0,00000 |
| country\_Poland | 1.980948 | 0,00000 | 0,00283 | 0,01251 |
| country\_Portugal | 0,12086 | 0,00000 | 0,00883 | 0,03292 |
| country\_Romania | -0,00414 | 0,36007 | 0,05411 | 0,06531 |
| country\_Slovakia | -1.110164 | 0,00000 | 0,00580 | 0,02225 |
| country\_Slovenia | 0,02324 | 0,00000 | 0,01817 | 0,02991 |
| country\_Spain | 0,10416 | 0,00000 | 0,00333 | 0,01882 |
| country\_Sweden | -1.394199 | 0,00000 | 0,01053 | 0,01827 |
| country\_United Kingdom | 0,50088 | 0,00000 | 0,01070 | 0,02027 |
| d10\_Man | -0,01501 | 0,00000 | 0,02146 | 0,00920 |
| d10\_Woman | -0,03344 | 0,00000 | 0,01999 | 0,00000 |
| d11 | 0,12159 | 0,00000 | 0,15624 | 0,01348 |
| d1\_Centre | -1.058548 | 0,00000 | 0,02166 | 0,00852 |
| d1\_Centre-letf | 0,02708 | 0,00000 | 0,01832 | 0,00967 |
| d1\_Centre-right | -0,04908 | 0,00000 | 0,01554 | 0,01413 |
| d1\_Left | -0,00221 | 0,00000 | 0,00838 | 0,02550 |
| d1\_Not positionable | 0,67036 | 0,00000 | 0,01108 | 0,01230 |
| d1\_Right | 0,21932 | 0,00000 | 0,00763 | 0,01501 |
| d25\_Large town | 0,37520 | 0,00000 | 0,02218 | 0,01813 |
| d25\_Rural area or village | -0,50532 | 0,00000 | 0,02254 | 0,01548 |
| d25\_Small or middle sized town | 0,16258 | 0,00000 | 0,02420 | 0,01584 |
| d25\_dk | -0,09640 | 0,00000 | 0,00000 | 0,00000 |
| d63\_Refusal/Other | 0,07710 | 0,00000 | 0,00167 | 0,00000 |
| d63\_The higher class of society | -0,31214 | 0,00000 | 0,00025 | 0,01566 |
| d63\_The lower middle class of society | 0,00645 | 0,00000 | 0,01274 | 0,01279 |
| d63\_The middle class of society | 0,25021 | 0,00000 | 0,02487 | 0,01669 |
| d63\_The upper middle class of society | 0,48254 | 0,00000 | 0,00965 | 0,00939 |
| d63\_The working class of society | 0,12266 | 0,00000 | 0,02778 | 0,01926 |
| d7\_Partner | -0,17288 | 0,00000 | 0,02548 | 0,00921 |
| d7\_Partner and children | -0,42915 | 0,00000 | 0,01697 | 0,00763 |
| d7\_Refusal/Other | 0,57177 | 0,00000 | 0,00000 | 0,00000 |
| d7\_Single | -0,89234 | 0,00000 | 0,02581 | 0,02300 |
| d7\_Single with children | -0,14367 | 0,00000 | 0,00460 | 0,01877 |
| d8\_16-19 years old | -0,62521 | 0,00000 | 0,02087 | 0,00774 |
| d8\_20+ years old | 1.216136 | 0,00000 | 0,02832 | 0,01323 |
| d8\_Refusal/dk | 0,15090 | 0,00000 | 0,00554 | 0,01377 |
| d8\_Still studying | -0,23429 | 0,00000 | 0,00501 | 0,00688 |
| d8\_Up to 15 years old | -0,14426 | 0,00000 | 0,01049 | 0,00556 |
| qb2 | 0,01775 | 0,00000 | 0,01704 | 0,01292 |

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)