**4.2 Prediction Behavior on 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*, using the following functions in order: LogisticRegression() , *DecisionTreeClassifier(), RandomForestClassifier(), XGBClassifier().*

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).[[1]](#footnote-1) 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 maintained as a metric, and age. After the conversion, the dataset has 65 independent variables.

**4.2 Prediction Behavior In the Entire Dataset**

It can be convenient, to summarize quickly, the predictive accuracy and macro-f1 of the models trained in table 1. The accuracy and macro-f1 are slightly better in the Random Forest’s model, which are respectively 0.70 and 0.67.

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 |

|  |  |
| --- | --- |
|  |  |
| Figura 1: Comparison of variables importance |  |

Logistic regression model has 0.67 of accuracy and 0.66 of macro f1-score. 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, and Portugal have a positive influence on predicting pro-environmental behaviour. [[2]](#footnote-2)

Decision tree model gets the accuracy of 0.66 and macro-f1 is 0.53. Cluster 2, or called extreme green-identity, is the predictor variable used for the primary split. 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 classification's outcome is most likely *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 is classified in the *no-action* class.

The accuracy and the f1-score of Random Forest model greatly improve, there are respectively 0.70 and 0.66. It is the best model. As shown in figure 17, analysing the feature importances of the model, I see that age and, once 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 favour of the environment than cluster 2.

Gradient Boosting concludes the first part of the analysis. Gradient Boosting gets slightly worse results than the Random Forest classifier. The accuracy is 0.6 and macro-f1 0.6. 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.

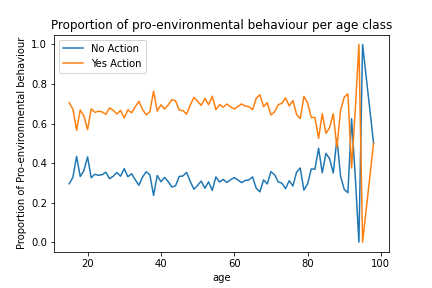
To sup up, most of the features with predictive importance are the same across models. Moreover, these predictors align with the hypotheses presented in section 2.1.

As referred to in the literature review (Chapter 1), the more an individual worries, the more he tends to take place environmental action. This statement is confirmed in the data. If I analyse the percentage of those who perform actions within the level of perceived risk, I find that for the classes with a low level there are more than 50% of individuals declare to not perform pro-environmental behaviour. The more the risk increases, the more the percentage of taking place action increase in turn. 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.

Figura 2: Crosstab between Risk Perception and Behaviour

Green identity, measured with clusters, is another important variable found in almost all models. 75% of individuals belonging to cluster 2 (extreme green-identity) have taken some eco-friendly actions. While 55% of individuals belonging to the cluster 1 (moderate green-identity) perform pro-environmental behaviour. Extreme green-identity has a positive influence on predicting behaviour. Therefore, also the second hypothesis is confirmed.

However, the third hypothesis is rejected. Cultural schemas, measured with CCA classes, are not particularly 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. 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 perform pro-environmental actions within the age class, I find that not only youngsters behave pro-environmentally, as sown in figure.

Also, higher education has a positive effect on pro-environmental behaviour and it is one of the most important variables in the Decision Tree and Gradient Boosting model. 72% of those who have declared to stop study after 20 years perform pro-environmental behaviour. The percentage is considerably higher than the other categories that have studied less or the students.

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-environmental 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 will call the first subset “of those who have a high-risk perception” and the second one “of those who have a low-risk perception”. At this point of the analysis, I can confirm the hypothesis suggests from the literature review: climate change risk perception is one of the main factors for predicting pro-environmental behaviour. Now, I want to understand what the most important predictors are 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.[[3]](#footnote-3)

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

Tables 2 and 3 summarize the performance of all models. In both cases, Random Forest has yielded the best performance when compared to other classifiers. 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.

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 |

|  |  |
| --- | --- |
| Subset of those who have a high-risk perception. | Subset of those who have a low-risk perception. |
| Logistic Regression | |
|  |  |
| Decision Tree | |
|  |  |

|  |  |
| --- | --- |
| Random Forest | |
|  |  |
| Gradient Boosting | |
|  |  |

Figure x displays the comparison among models of the first 20 feature importance for predicting pro-environmental behaviour according to the different levels of climate change risk perception.

The accuracies of the two Logistic regression models are 0.65 for the subset of those who have a high-risk perception, and 0.63 for the subset of those who have a low-risk perception. The important variables in the two different models are similar. For both models, Romania has a negative influence on predicting pro-environmental behaviour, while Malta has a positive influence.

The accuracies of the Decision Tree models are 0.63 and 0.60. These results are the worst when compared with the other models. For both trees, the root is the green-identity. On one side in the first tree the root is occupied by cluster 2, the extreme green-identity, on the other side the root is employed by 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.

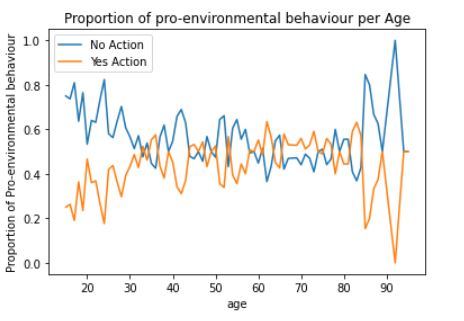
Random Forest yields the best accuracies in both models, 0.70 and 0.65. However, it does not happen in macro-f1 in the model of those who have a high-risk perception. In both cases, age is the most important variable for predicting behaviour. The trend of age in the subset with high-risk perception is similar to the complete model. On average 60% of individuals 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 x shows. On average 70% of younger have 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.

Figure 1: trend of age in the Subset with low-risk perception observations.

Accuracies of Gradient Boosting models yield 0.67 and 0.65. The model of those who have a high-risk perception macro-f1 is 0.64, which is the best result. In both models, there are two main important variables: 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. Then, for the first time, political orientation turns out to have an important role. If I compute the proportion in this subset 50% of radicals take place pro-environmental behaviour, while the other orientations get a smaller percentage.

To sum up, in this second part of the analysis I do not discover any significant differences in the comparison between the two opposite models. The important variables remain roughly the same. The high level of education is an important variable that often appears in the feature important plot for those who worry about climate change. In this subset, 75% of high-educated citizens perform pro-environmental behaviour. Therefore, high- education mixed with high-risk perception may positively influence to behave pro-environmentally.

The opposite case is the subset of those who do not worry about climate change. In this case, Romania and cluster 1, moderate green-identity, are the variables found most often in models. 85% of Romanians who do not worry about climate change do not perform any actions. Therefore, this variable has a negative influence on predicting pro-environmental behaviour. Also, moderate green-identity could have a negative influence on predicting pro-environmental behaviour, albeit in a less direct way. 58% of those who do not worry and with a moderate green-identity do not behave pro-environmentally.

1. See appendix for the grids of parameters and the hyperparamenters for each algorithm. [↑](#footnote-ref-1)
2. I remember that these coefficients are in log-odds terms the 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 ratio 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-2)
3. See appendix for more details. [↑](#footnote-ref-3)