choyze
Sustainability made by you.
developed by vtmw®

Sustainability: Preferences, Explainability and Efficiency

Patrick Stahl, Yannik Porrmann and Lukas Nolte
2023

Acknowledgements

We would like to thank the German Environmental Agency for the support and the joint exchange as well as for the provision of the collected data.

Abstract

This white paper focusses on the technical steps taken to reduce the number of questions needed to project the attitude towards environmental topics. This was achieved by using two different kind of machine learning algorithms: one binary classification and one multiclass classification algorithm to project the different clusters (or the two groups of cluster in case of the binary classification). Followed by an analysis of the features with the highest impact on the models to identify which features or questions need to be asked to build a model based on a smaller sample of questions. As it turns out throughout the analysis, it is possible to reduce the number of questions significantly in both types of algorithms without losing too much quality in the projections. Furthermore, it turns out that both algorithms laying a lot of weight towards similar questions. Interestingly both algorithms do not lay a huge focus towards sociodemographic factors.

Is it possible to deduce clusters using algorithms?

As a practical approach, we proved that it is possible to deduce the German Environmental Agency (GEA) clusters using a machine learning algorithm in two steps. The first study was a simple decision classifier with two clusters: a sustainable cluster, which includes the clusters open-minded, oriented and consistent, and a not-sustainable cluster, which includes the clusters rejecting, sceptical and undecided.

In addition, we analysed which questions have the highest influence on the classification. In a second study we created a classifier-model for all six clusters and analysed which questions have the biggest impact on the decision.

With this at hand, the question arises if there exists something like a best choice in regard to simplicity, e.g. low number of questions, and explainability which are diametrically opposed.

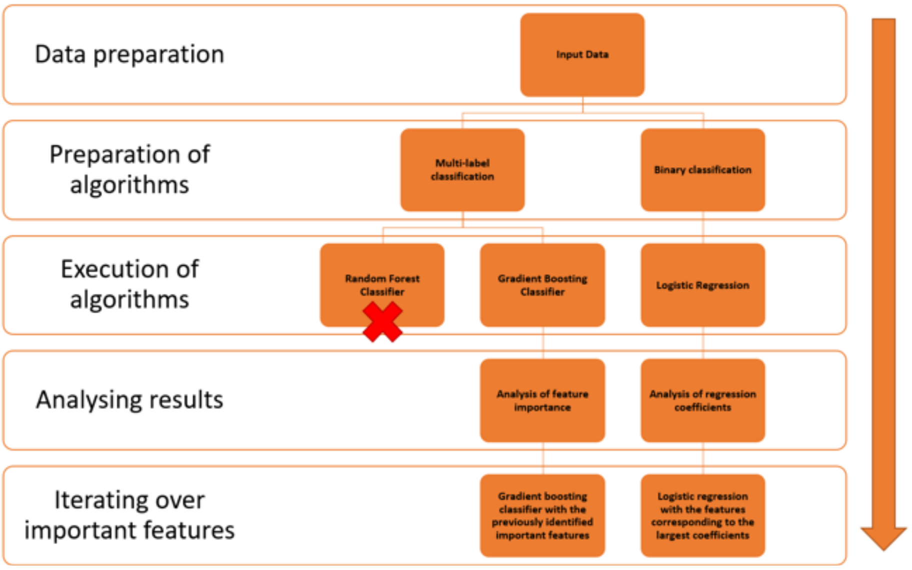

Data preparation
Input Data
Preparation of algorithms
Multi-label classification
Binary classification
Execution of algorithms
Random Forest Classifier
Gradient Boosting Classifier
Logistic Regression
Analysing results
Analysis of feature importance
Analysis of regression coefficients
Iterating over important features
Gradient boosting classifier with the previously identified important features
Logistic regression with the features corresponding to the largest coefficients

Figure 1: Flowchart of our modelling approach

Participants/Material

The data set used for the modelling was collected by the German Environmental Agency, analysed and divided into 6 clusters (the rejecting, the sceptical, the undecided, the open-minded/ready, the oriented, the consistent). The results were published in the context of the “Representative survey on environmental awareness and behaviour in 2020” (Stallmann, 2022).

The collection of data took place in an online survey between 01.11.2020 and 08.12.2020. Recruited were thereby via the Infas ad hoc panel 2,115 people. Of these, 43% percent are women, the average age (median) is 57.7 (60) years (range 14 - 92 years), 21.7% went to school for 10 years, 19.4% percent 11 to 13 years, 49.8% have a university degree and 9.1% other. The self-reported net household income is distributed across the income categories as shown follows:

Income categories	Percentage distribution of the sample
< 1000 €	2.7 %
1000 € – 2000 €	14.3 %
2000 € – 3000 €	20.6 %
3000 € – 4000 €	20.7 %
4000 € – 5000 €	16.4 %
5000 € – 6000 €	10.3 %
> 6000 €	11.6 %

Table 1: Net household income of the sample in percent.

Preprocessing

The goal of preprocessing was to create an Analytical Base Table (ABT). The starting point for preprocessing was the data set provided by the German Environmental Agency. This data set consists of ordinal, categorical and numerical data as well as features calculated based on the data set.
During data preparation, the answers were first transformed into numerical values so that they could be statistically evaluated and used for the algorithms. In a second step, the questions and the answer choices were analysed in order to apply the required procedure for the respective data. For questions whose answers are categorical in nature (for example: state, gender, ...), the answers were subsequently converted into a binary representation using the one-hot encoding procedure (F. Tomaschek, 2018). To prevent collinearity, one of the newly generated features is removed in each case. From the resulting data set, the features calculated by GEA were removed with the exception of the clusters, so that only the processed responses as well as the clusters determined by GEA were included in the ABT. This resulted in 336 remaining features.
To avoid dependencies, the next step was to analyse the correlations between the considered features. For this purpose, a train-test split (80 % / 20 %) with fixed seed and stratified random sampling was performed in advance. This aims to perform the correlation analysis based on the training data only, in order to exclude the information of the later test data from the correlation analysis. On the training data, the Pearson correlation coefficient is calculated between each feature. For features that have an absolute coefficient greater than 0.85, only one feature was retained and the rest were removed. In this way, the ABT was reduced by 40 features to 296.
Before applying the algorithms, the ABT data is standardized. For standardization, the mean and variance from the previously created training data set are used and applied to the entire data set (StandardScaler).

General Procedure

First Study

Based on the previously created ABT, a target vector for the binary classification using logistic regression (Nasteski, 2017) was now created using the clusters determined by the GEA. For this purpose, the six clusters are divided into the classes "not ecologically sustainable" (0) = [the rejecting, the sceptical, the undecided] and "ecologically sustainable" (1) = [the open-minded/the ready, the oriented, the consistent].

This separation was based on the evaluation of the six clusters by the GEA in respect to four dimensions ("Environmental attitude", "Climate attitude", "Environmental behaviour", "open to change") on a scale consisting of low, middle, high and very high. This scale was mapped to an ordinal scale from one to four so the average could be calculated. Looking at the average across the four dimensions for each cluster it shows the following:

Cluster	Environmental attitude	Climate attitude	Environmental behaviour	open to change	Average
the rejecting	2	1	1	1	1.25
the sceptical	3	2	2	2	2.25
the undecided	3	3	1	2	2.25
the open-minded	4	4	2	3	3.25
the oriented	4	3	3	2	3.00
the consistent	4	4	3	3	3.50

Table 2: GEA results on the four dimensions on the six clusters (own representation)

Looking at the average values ranking between 1.25 and 3.5 there were two natural steps between 1.25 and 2.25 and one between 2.25 and 3.00. In this case it was decided to go with the separation between the third and fourth cluster to ensure that both classes had enough samples to use the algorithm. Besides this step over the curse of the procedure further steps were performed to handle the imbalanced data set, e.g. usage of the correct metrics (in this case F1-Score).

For the construction of a logistic regression model, a grid search was performed over different constellations of hyperparameters[1] (including regularization). The goal is to optimize the model with respect to the different hyperparameters. The

[1] Hyperparameters describe parameters that are given to the model and are not directly learned by the algorithm.

different constellations are calculated and evaluated with respect to the F1 score (Sammut, 2011).

In the context of this calculation, a cross-validation (k = 10) is also performed. Thus, for each new model calculation, the training data set is split into k parts and trained on k-1 parts and checked against the remaining part in each case. The model is thereby evaluated based on the average error of the k trainings.

After selecting the model with the lowest average error in the CV, the question arose as to which features had the greatest impact. The coefficients of the logistic model were analysed and compared with respect to their absolute size. This can be justified due to the structure of the logistic function, since the output is calculated as follows:

$$g(z) = \frac{1}{1 + e^{(-z)}}$$

with

$$z(x) = \sum_{i=1}^{n} c_i \cdot x_i + b$$

where $c_1, \dots, c_n$ correspond to the coefficients of the logistic regression, x_1, ..., x_n to the values of the data point for the features 1 to n with b as intercept. Since the data was standardized before calculating the model, the coefficients have comparable magnitudes.

Based on the values determined in this way, the features were ranked in descending order and stored (see Figure 2).

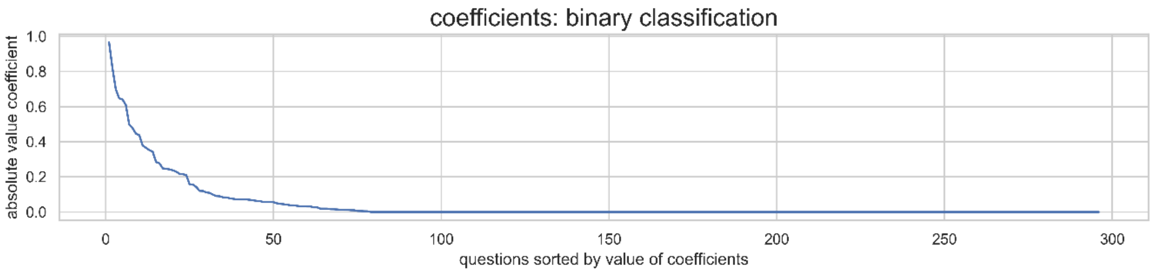


Figure 2: Ranking of the questions according to their corresponding regression coefficients.

In the last step, we iterated over the features with the largest coefficients, so that initially a model was formed that consisted only of the feature with the largest absolute coefficient, followed by a model with the two features of the two largest coefficients up to a model with the 20 largest coefficients. The models created in this way were compared to identify at what number of features sufficient model accuracy is achieved[2].

Second Study

In order to be able to assign more accurate environmental behavior to potential customers, this step was based on the German Environmental Agency's clustering (the rejecting, the skeptical, the undecided, the open-minded/ready, the oriented, the consistent) as the target variable; it is thus a multiclass classification. Again, the main objective is to reduce the number of questions to be asked while maintaining as much explanatory power of the model as possible.

The independent variables also result here from the survey collected by the GEA, which was preprocessed as described in the "Preprocessing" section. For the later evaluation of the created model, the data set was first divided into training and test sets (80 % / 20 %) using stratified random sampling. Two types of models were tested against each other: a random forest classifier (Khaled Fawagreh, 2014) and a gradient boosting classifier (Natekin Alexey, 2013). In both cases, the training data set was scaled using a standard scaler. The optimization of the hyperparameters of both models was done using grid search with an integrated cross-validation (k=10). With the help of the test data set, the created models were evaluated. The gradient boosting classifier achieves the best results (see paragraph: "Technical results" within subparagraph: "Second Study"). To reduce the number of questions, the model was used to rank the features based on the feature importance (see Figure 3).

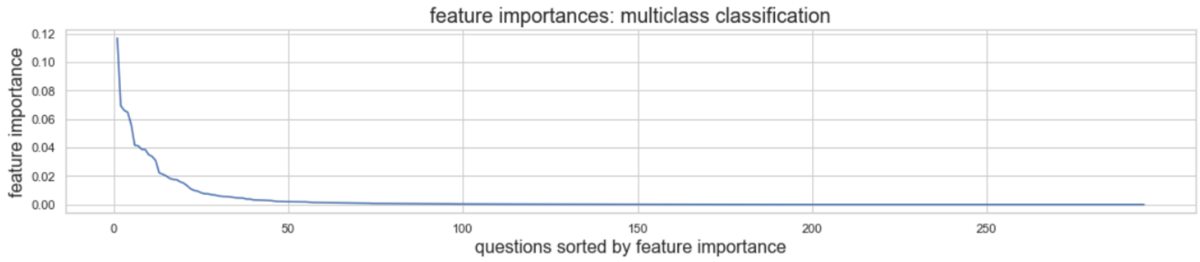


Figure 3: Ranking of questions according to feature importance.

Based on this feature ranking, the gradient boosting classifier was recalculated, optimized and evaluated[3] by successively adding features. In the first step, the model with the feature with the highest feature importance was calculated and

[2] To evaluate the individual modelling steps, the confusion matrix, the accuracy, the precision, the recall and the F1 value were calculated and evaluated in each case.

[3] The same way as described in footnote 2.

optimized. The additional features were added similar to the binary classification to determine at what number of features a sufficient evaluation is achieved.

Technical results

First Study

The model determined using the procedure described in paragraph: “General Procedure” within subparagraph: “First Study”, trained on 296 features, yields an accuracy of 94.09% and an F1 score of 0.95 for cluster 1 ("ecologically sustainable") on the test data.

As part of the analysis of the size of the coefficients (see Figure 2), the three features with the largest impact were identified from these 296 features, which are shown in Table.

Question	Answer
I donate money to environmental or conservation groups.	1: yes, applies 2: no, does not apply 8: I cannot say
I would be willing to switch to a green power plan.	1: yes, definitely 2: rather yes 3: rather no 4: no, definitely not 8: I cannot say
I would be willing to live on less living space.	1: yes, definitely 2: rather yes 3: rather no 4: no, definitely not 8: I cannot say

Table 3: Top 3 corresponding questions: binary classification

In the comparison of the accuracy (with successive addition of features according to the size of the coefficients) it can be seen that with a model with four features an accuracy of over 80 % is reached for the first time and with nine features an accuracy of over 90 % is reached for the first time (see Figure 4).

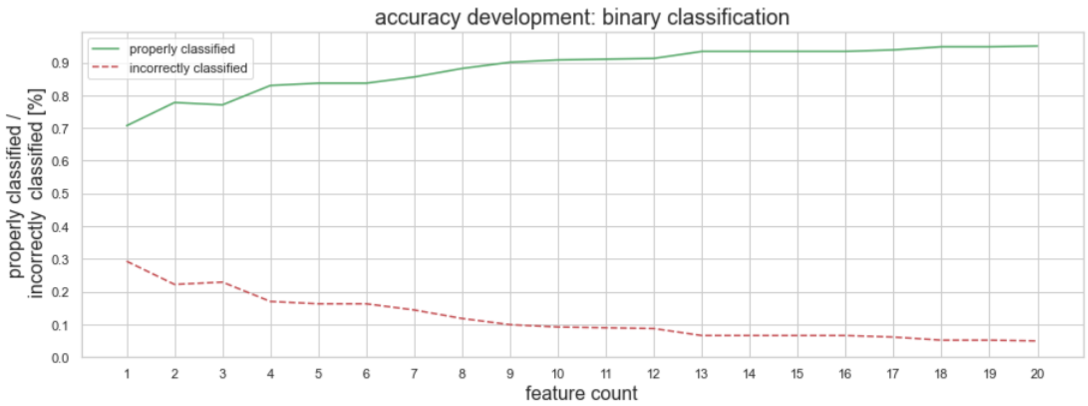
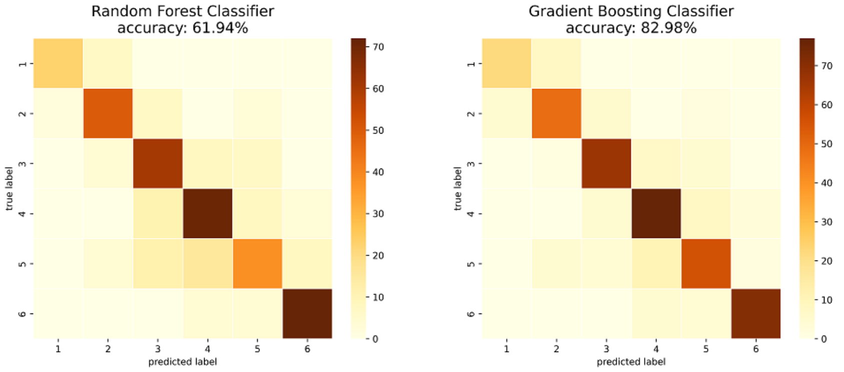


Figure 4: Accuracy over the course of the models with the respective number of features.

Second Study

Based on the optimization procedure described in paragraph: "General Procedure" within subparagraph: "Second Study" and using all features, it is shown that for the best estimator of the respective model class, the gradient boosting classifier with an accuracy of 82.98 % leads to a better generalization to new data (in this case the test data set) compared to the random forest classifier and thus to a better classification result (see Figure 5).



Harmonic means of F1-scores

Gradient Boosting Classifier: 0.8002
Random Forest Classifier: 0.7347

Figure 5: The best estimators per model type and the harmonic means of their F1-scores.

The confusion matrix for the random forest classifier shows that significantly more "false positive/false negative" classifications occur than with the gradient boosting classifier. Accordingly, only 61.94 % of the test data are correctly classified. In contrast, the gradient boosting classifier correctly classifies 82.98 % of all test data. When calculating the models by successively adding features with the greatest influence in each case, it becomes apparent for the respective confusion matrix that the characteristic diagonal shape of such a matrix can already be recognized when using the first three features, i.e., those with the highest feature importance (see Figure 6).

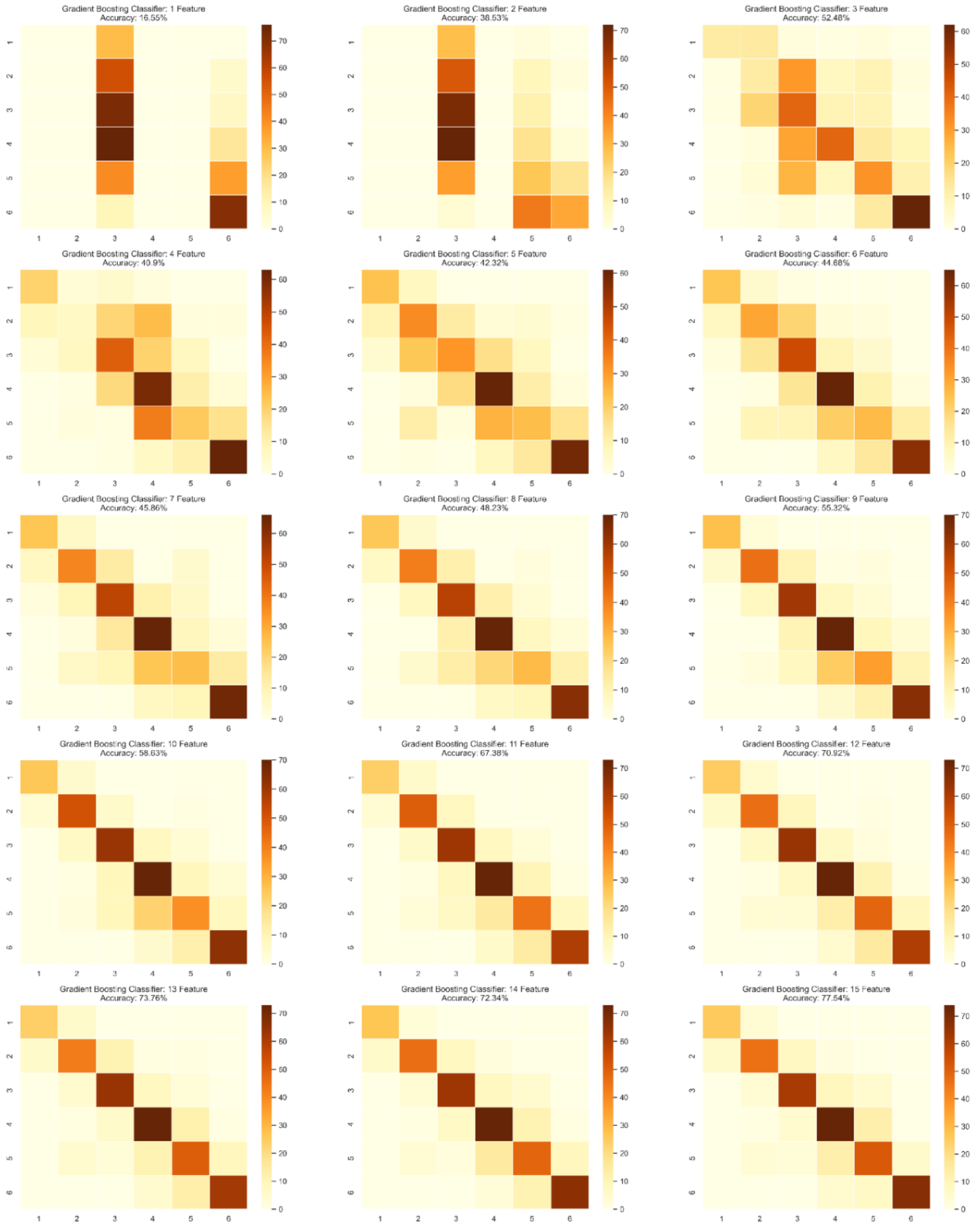


Figure 6: Change of confusion matrix during remodelling and evaluation under successive addition of the first 15 features with greatest influence.

Based on the "Representative Survey on Environmental Awareness and Behavior in 2020," these are the following three questions:

Question	**Answer**
I donate money to environmental or conservation groups.	1: yes, applies 2: no, does not apply 8: I cannot say
I am actively involved in environmental protection and nature conservation.	1: yes, applies 2: no, does not apply 8: I cannot say
For the sake of the environment, we should all be willing to cut back on our current standard of living.	1: yes, definitely 2: rather yes 3: rather no 4: no, definitely not 8: I cannot say

Table 1: Top 3 corresponding questions: multiclass classification

The comparison of the features based on the highest regression coefficients (binary classification) with those of the highest feature importance (multiclass classification) shows that in both cases the question about donations for environmental protection/nature conservation has the greatest influence on the classification.
The addition of further features leads to a reduction of “false positive/false negative” classifications and serves to improve the accuracy (see Figure 4 and Figure 7). Using only 15 features with the highest feature importance, an accuracy of almost 80 % can be achieved.

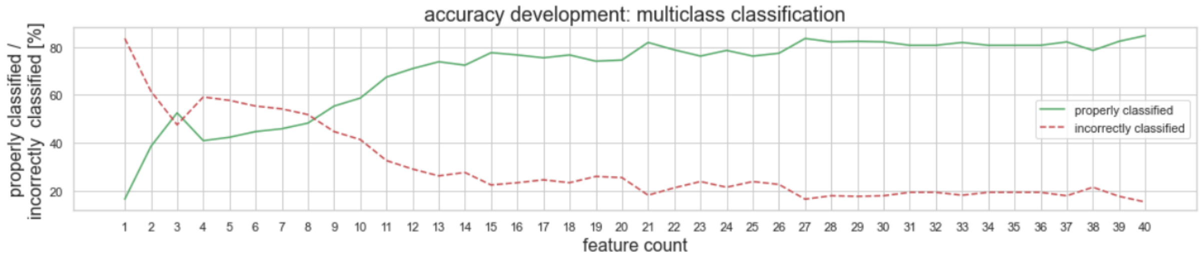


Figure 1: An accuracy of 77.54 % can already be achieved with 15 features used for classification.

Overall, both approaches show that the number of features can be significantly reduced. The best estimators of the binary and multiclass classification lead to similar results.

Looking at the questions with the highest relative impact on the two different models it shows that using the 25 features with the highest relative impact on the models, roughly 80% of the explainability at both models is included. (multiclass 83.4%, binary 78.6%).

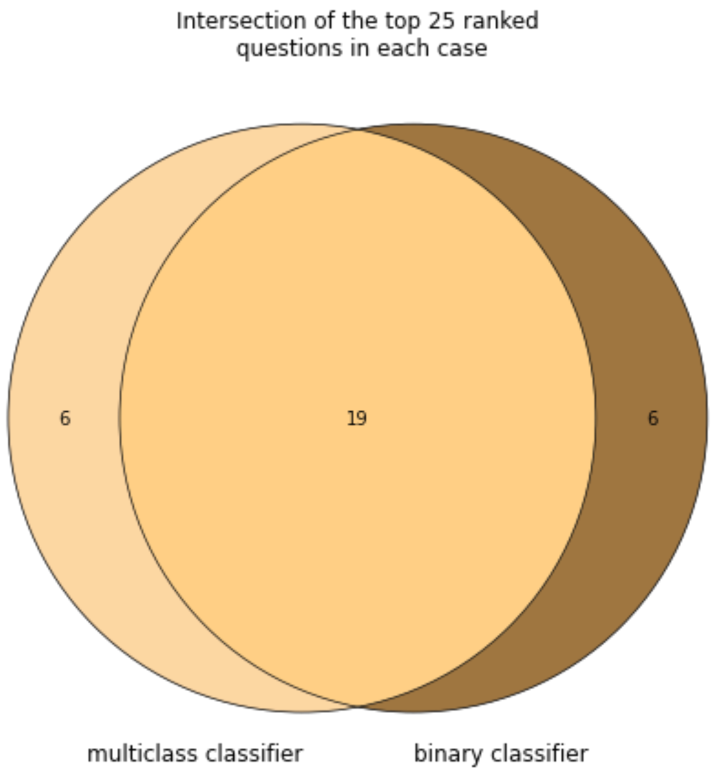


Figure 8: The intersection of the top 25 questions per classifier is 76 %.

Comparing the top 25 questions of the respective best estimator (binary / multiclass) shows that these coincide by 76 % (see Figure 8). These 19 coinciding questions account for 74.6 % of the feature importance in the multiclass classification and for 68.0 % in the binary classification.

Most of the additional questions needed are used to distinct between the "subclusters" of "ecologically sustainable" and "not ecologically sustainable", i.e., it is far more difficult to classify the finer levels of differences between different points of view regarding sustainability in contrast to the question if someone is sustainable or not.

Discussion

Given the importance of low barriers for decision making in sustainable investments outlined in this paper it is noteworthy to realize that with only four questions it is possible to classify a person into one of the clusters “ecologically sustainable” or “not ecologically sustainable” with an accuracy of over 80% as outlined in the first study. This directly relates to the problem, that sustainability is not well understood, since a very low number of questions is enough to classify correctly – it was just not known before which questions should be asked. Interestingly enough, the most important question is one, that directly bridges the value-action-gap since it directly includes action.

As mentioned in the abstract, age, gender and level of education are only conditionally suitable for explaining sustainability preferences, since they do have an impact on explainability, but are not part of the most important features. This is an important finding for classification since it allows algorithms to be precise without being prejudiced.

The findings regarding the overlap of features for both cases and the need to differentiate between the finer clusters clearly defines a research assignment as the differences between the subclusters need to be analysed precisely to understand if it is useful to use more questions for classification and therefore increasing the barrier or to keep the barrier low and therefore loosing explainability.

Bibliography

F. Tomaschek, e. a. (2018). Strategies for addressing collinearity in multivariate linguistic data. Journal of Phonetics, Volume 71, 249-267.

Khaled Fawagreh, M. M. (2014). Random forests: from early developments to recent advancements. Systems Science & Control Engineering, 602-609.

Nasteski, V. (2017). An overview of the supervised machine learning methods. HORIZONS.B. 4. DOI:10.20544/HORIZONS.B.04.1.17.P05., 51-62.

Natekin Alexey, K. A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics.

Sammut, C. W. (2011). F 1-Measure. In C. W. Sammut, Encyclopedia of Machine Learning (p. 397). Boston, MA: Springer.

Stallmann, M. (2022). Repräsentativumfrage zum Umweltbewusstsein und Umweltverhalten im Jahr 2020. Umweltbundesamt. Retrieved from https://www.umweltbundesamt.de/en/publikationen/repraesentativumfrage-umweltbewusstsein-0