Graduate School of Natural Sciences
Master Applied Data Science

Using Multiverse Analysis for Estimating Response Models: Towards an Archive of Informative Features

Master thesis

Lieve Göbbels

Supervisors:
Dr. Kyle M. Lang
Dr. Maarten Cruyff

Final version

Utrecht, June, 2023

Abstract

This study investigates which features are informative in predicting non-response for certain target variables in the context of European behavioral- and social-pattern data and how a multiverse approach can guide the process of identifying these predictors, with the long-term goal of building an archive of essential predictors. This will help researchers to design their studies in such a way that the missing at random mechanism can be assumed safely, ensuring valid use of advanced imputation techniques. Within the context of this study, a consensus on the types of variables that are informative can be discerned. That is, the results suggest the importance of variables related to employment, education level, domicile, and household and partner information. Limitations remain in accounting for the researcher degrees of freedom and the missing data in the observed variables, indicating the relevance of conducting similar, additional analyses to get a more robust collection of essential predictors. Nevertheless, this study provides an initial set of important predictors in the context of social science-related data and shows that multiverse analysis can adequately guide the process of identifying predictors of non-response by enabling flexibility in the construction and deployment of a set of models, rendering it easy to implement in different domains.

Keywords: *Multiverse analysis*, *item non-response*, *identifying predictors*, *binary classification*

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

The past decade, emerging big data practices have had an increasing influence on day-to-day life and and the sciences studying this day-to-day life (Bormida, 2021). This influence is likely to continue to grow over the upcoming years, not just in the societal context, but also in for instance medicine and law (Bormida, 2021). According to some, big data will change the world for the better. However, more and more researchers are warning for the (potential) downsides of big data, such as a lack of ethical and legal standards or the low quality of data leading to incorrect knowledge (Weinhardt, 2020; Inside BigData, 2020; Montvilas, 2022).

According to Peng et al. among others, low-quality data can be caused by missing values (Peng et al., 2022). Missing values – as the term indicates – regards incomplete information in the sense that (meaningful) values are unobserved or hidden and thus not present in a data set (Graham, 2012). A multitude of methods exist to deal with missingness (non-response) in data sets, like deleting incomplete observations (*listwise deletion*) or imputing them with a constant or a predicted value. These methods are however limited in their abilities and can only be validly used under the right circumstances (Peng et al., 2022; Graham, 2012; Van Buuren, 2018). That is, they can lead to problematic inference procedures where invalid conclusions are drawn from the analysis of the data when erroneous assumptions are made about the underlying missingness mechanism (Van Buuren, 2018).

A central assumption for modern missing data imputation techniques regards the type of missing data mechanism underlying the data: missing at random (MAR) (Van Buuren, 2018). That is, the missingness – that is in itself not random – is (sufficiently) accounted for by the observed information of the predictor variables present in the data set (Van Buuren, 2018; Graham, 2012; Peng et al., 2022). So, to satisfy this assumption one should include all variables that correlate with the missingness of the 'target' variable. However, determining which variables are informative regarding missingness of the target variable is currently mostly based on educated guessing, thus lacking valuable guidance.

This study therefore focuses on providing a starting point of such guidance by attempting to determine essential predictors of missingness in a data set regarding European behavioral- and social-pattern data. Ideally, combining this study with other studies, this will result in an archive of essential predictors so that social science researchers – and possibly researchers of other fields – can design their studies in such a way that MAR is (sufficiently) guaranteed and that advanced imputation techniques can be used validly.

More specifically, in this study, a multiverse analysis approach is taken, where several models are used to individually predict the missingness of a single target feature, and doing this for a multitude of targets. That is, for each target, the best model is sought by (automatic) hyper-parameter tuning and cross-validation, after which for the best model the feature importances are

calculated. Moreover, to make the resulting feature importances more generalizable and significant, this process is applied to not one, but three different types of models, namely Support Vector Machines (SVMs), Random Forests (RFs), and Multilayer Perceptrons (MLPs). This allows for comparison of the feature importances – either strengthening or nuancing the importance of the respective features. As such, the central research question of this study is defined as follows:

What are important predictors of non-response in the context of European behavioral- and social-pattern data, and how can this analysis contribute to the creation of an archive of essential predictors?

First, underlying concepts and theory will be discussed in Chapter 2, after which the used data and methods will be described. This is followed by a description and discussion of the main results in Chapter 5 and 6. Lastly, after describing an ethical assessment of the study using the Fundamental Rights and Algorithms Impact Assessment (FRAIA) guidelines in Chapter 7, in Chapter 8, the main gains and limitations of this study will be discussed.

Chapter 2

Theoretical Framework

2.1 Related work in predicting non-response

Even though item non-response is something that exists in most questionnaire-related studies, there are few studies that attempt to define what variables may be informative regarding such missingness. One study that uses this as main objective is the recent work by Kern, where a longitudinal framework for predicting non-response in panel studies is proposed (Kern et al., 2023). A more deductive approach is taken in Bulut et al. (2020), where the impact of several predefined features, such as types of bullying and grade level, on non-response in a self-reported bullying questionnaire has been analyzed. Similarly, most – of the already scarce – studies that aim to define predictors of missingness take a similar deductive approach, where one or a few variables are chosen to be assessed based on a set of assumptions or theory (Lipps and Monsch, 2022; Alexander, 2017; Lee et al., 2017; Kutschar and Weichbold, 2019). However, similar to the approach taken in this study and in Kern et al. (2023), there are some studies that take a more inductive approach, where first a number of analyses or observations is conducted, patterns are extracted and only then assumptions are made and conclusions are drawn (Elliott et al., 2005; Mignogna et al., 2022; Blumenberg et al., 2018). Most of these studies found, like Elliott et al. (2005), take place in the context of health sciences, which can be informative to the social sciences but at the same time highlights the gap still present in the latter context.

2.2 Core concepts

2.2.1 Missingness mechanisms

In general, three types of missingness or non-response mechanisms are distinguished (Rubin, 1976; James et al., 2013; Graham, 2012). These are *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). These mechanisms determine the type of missing data, which is organized into the same three main categories (James et al., 2013; Van Buuren, 2018; Graham, 2012). That is, the type of missingness of the data specifies the mechanism that generates this missingness. MCAR, which is in the statistical sense an unrealistic assumption, means that the missingness probability is the same for all causes, so the causes of missing data are unrelated to the data used during analysis (i.e. the observed data). This is unrealistic, because in practice, there is always some (group of) variable(s) that explains the missingness pattern as a whole or in part (James et al., 2013; Peng et al., 2022; Graham, 2012). MAR, on the other hand, is a more realistic assumption, which states that the probability of being missing is the same within groups defined by the observed data (James et al., 2013; Van Buuren,

2018; Graham, 2012). This would for example be the case when individuals from the Netherlands would be less likely to completely fill in a questionnaire about their grocery shopping behavior compared to individuals from Belgium, as the missingness relates to the country of the individual and not the grocery shopping behavior – given that the country feature is fully observed. Lastly, the MNAR assumption is used when the missingness is systematically related to events or factors that have not been measured (i.e. the unobserved data) (James et al., 2013; Van Buuren, 2018; Graham, 2012).

In practice, this categorization should however be taken somewhat loosely and is instead best interpreted as a spectrum rather than a hard distinction. That is, because – even though MCAR can be ruled out as soon as there is any association between the measured data and the missingness – MNAR and MAR cannot be distinguished from each other as the supposed predictor data are by definition unavailable, making it impossible to test the imperative assumptions (Little, 1988; Enders, 2022; Goldberg et al., 2021; Graham, 2012). Graham (2012) specifically argues that MNAR and MAR are a continuum and that both mechanisms in the pure sense are merely theoretical concepts that do not exist in practice. That is, a cause of missingness is neither purely MNAR nor purely MAR. By definition, the missingness mechanism is context-dependent, based on the requirements and applications of the data, meaning that the missingness mechanism is characterized by both the data and the applied analysis.

For validly using the MCAR assumption it needs to be shown that the probability of the data being missing is completely unrelated to any other observed variable and to its own missingness, which can be done by testing the group characteristics for equality among the complete cases and the cases with missingness (Little, 1988; Enders, 2022; Goldberg et al., 2021; Graham, 2012; Collins et al., 2001). MAR can be identified by checking the correlations between missing values and complete variables in the data, where (high) correlation indicates a MAR assumption. However, related to what is stated by Graham, exact minimal correlation requirements are context-dependent. With MNAR, observed features do not provide any significant explanation of the missingness pattern, which may be due to the values of the target feature correlating with the missing values (e.g. respondents with an age over 65 do not report their age) or the missing values being caused by an unobserved feature (e.g. a scale wearing out over time).

MNAR can be transformed into MAR by including auxiliary or proxy variables. These are variables that help estimating incomplete data as they relate to the missingness probability of the incomplete variable, but do not take part in the main analysis (Collins et al., 2001). Proxy variables can be measured relatively easily and thereby replace a variable that cannot or difficultly be measured. For example, historical environmental conditions can be explained by the width of the rings in trees. Here, tree ring width is the proxy variable to replace the historical environmental condition variable with.

2.2.2 Imputation methods and their limitations

Since collecting all data, which is often considered the 'golden road,' is practically infeasible and generally unethical, methods have been developed to deal with missing values in data sets. Some methods remove incomplete observations from the data set, such as listwise deletion (complete case analysis) or pairwise deletion (available case analysis). Other methods deal with missing

values by imputing them with estimated values to produce a complete data set.

There are several different types of imputation, which can be roughly distinguished into two main categories: single and multiple imputation. Single imputation regards replacing the missing value once. Multiple imputation on the contrary regards imputing the missing data with different values, often in a principled way comparing different plausible values, thus performing the imputation for a data set multiple times (Dong and Peng, 2013). Some well-known single imputation methods are mean or mode imputation, (stochastic) regression imputation, Last or Baseline Observation Carried Forward (LOCF/BOCF), and the indicator method. A popular multiple imputation technique is Multiple Imputation by Chained Equations (MICE) which is also known as Fully Conditional Specification (FCS). Moreover, there are also principled non-imputation methods like Full Information Maximum Likelihood (FIML) which originated from the field of structural equation modeling (Dong and Peng, 2013).

With listwise deletion, observations with one or more missing values are removed from the data set. This procedure is unbiased under MCAR, but biased under MAR and MNAR and contains the risk that too large a part of the data is removed, especially when dealing with high-dimensional data (Van Buuren, 2018). With pairwise deletion, summary statistics of the observed data are used in the analysis, which provides consistent estimates of the mean, correlations and covariances under MCAR. However, this procedure also suffers from bias under MAR and MNAR and additionally requires strong assumptions regarding the distribution of and relationships within the data (Van Buuren, 2018). Mean or mode imputation is severely limited and should in principle not be used, as it is not only biased under the MAR and MNAR mechanisms but can also be biased under MCAR, which can lead to significant distortions of the distribution and underestimation of the variance (Van Buuren, 2018). Regression imputation, which replaces missing values by a prediction of the value, on the other hand, provides unbiased estimates of the weights (not the mean) under MAR and MNAR in certain specific cases, but suffers from underestimation of the prediction error, uncertainty and variability of imputations (Van Buuren, 2018). Stochastic regression imputation, by adding noise to the prediction process, mitigates most of the shortcomings of regular regression imputation, but incorrectly treats the imputed data as real data (Van Buuren, 2018). LOCF and BOCF, which impute the last seen or baseline value, respectively, can be biased under all three mechanisms and have the additional limitation that they are only suitable for specific cases of longitudinal data (Van Buuren, 2018).

The MICE procedure works by replacing missing values with initial imputations in the first iteration, after which these imputed values are used to (re-)estimate the values for the missing elements in subsequent iterations (Van Buuren, 2018). This is done by a regression function. So, after the first round of imputations, in each subsequent iteration, each feature is consecutively assigned being the target and its imputed values are removed again once per iteration, and are estimated by the remaining complete variables. The previous imputation for the current target is then updated by the newly estimated value. This continues for each feature for a predefined number of iterations or until other predefined stopping criteria are met (Van Buuren, 2018).

In regression-based imputation methods like single or multiple stochastic regression, auxiliary variables are included in the imputation model, therewith providing the imputation algorithm with additional information on the missingness. This helps estimating the imputed values with more precision and less bias (Van Buuren, 2018). With likelihood-based methods like Full Information

Maximum Likelihood (FIML), proxy variables are included through the covariance matrix in the model estimation, for which several methods exist. Collins et al. (2001) describe multiple general requirements for adding proxy variables to likelihood-based models: 1) the auxiliary variables should be directly correlated with the measured predictor variables, 2) they should also be directly correlated with the error of the outcome variable, and 3) the proxy variables should also be correlated with each other.

So, to ascertain state-of-the-art imputation methods can be used in a valid and unbiased way, it is vital to ensure the data are following the missing at random mechanism sufficiently.

2.2.3 Binary classification

Binary classification is a task that belongs to the field of supervised machine learning and attempts to categorize observations into one of two possible categories. Supervised machine learning describes a class of methods where the data used to train the model has observations and corresponding outcomes (i.e. the category to which a particular observation belongs to). There are several commonly used methods for binary classification that are suitable for different domains and data characteristics within the area of binary classification.

Naive Bayes is often considered the most basic binary classification model one could implement because it assumes the presence of one feature being uncorrelated to the presence of the other features (Witten et al., 2016). It makes classifications based on the log-prior (i.e. the (log) probability of picking a particular target value when random sampling) and log-likelihood of an observed value (Witten et al., 2016). So, $prediction = \log_{prior} + \sum_{i=1}^{n} \log_{likelihood}(X_i)$. If the predicted value is around 0, the model is uncertain, while a value with a higher magnitude represents one of two classes (e.g. 10 for fruits and -10 for vegetables). Its main advantages are that it is a relatively fast method, it works well with categorical input variables, and is generally better than other methods while using less training data – given that the features are independent (Witten et al., 2016). However, this feature independence assumption rarely holds in practice, rendering the method unsuitable for most real-world data sets (Witten et al., 2016). Moreover, it suffers from the zero frequency problem, which can be defined as the inability to make predictions about categories that are not present during training (Gupta, 2020).

Another relatively easy to implement method is logistic regression. Herein, a sigmoid function is used to map the predicted values from the linear equation into a range between 0 and 1. Using the objective of minimizing the cost function, so the prediction error, logistic regression attempts to find the model that best separates the observations into the two classes. It is in general more efficient to train and provides outputs that are easier to interpret compared to some of the other methods (Witten et al., 2016). Moreover, it does not make assumptions about class distributions across the feature space and has generally good performance on simple, (nearly) linearly separable data sets, implying low risk of overfitting. This risk, however, increases when dealing with high-dimensional data sets, but can to some extent be mitigated by adding some form of regularization (Witten et al., 2016). Another shortcoming is the tight space in which logistic regression provides good classifications. That is, the classes must be close to being linearly separable, but neither exactly linearly separable, nor too far from it (Witten et al., 2016).

The Support Vector Machine (SVM) method is designed for handling relatively small data

sets. It works especially well on small and complex data sets (Witten et al., 2016). This is due to its objective to find the hyperplane that bests separates the observations into the two classes in an N-dimensional space, where N equals the number of features. To handle non-linearity in the data, a kernel is used to transform the data into this higher dimensional feature space. A well-known kernel is the *radial basis function*, which is based on the similarity between two points in terms of Euclidean distance. Besides being able to handle complexity in small data sets relatively well, another advantage of SVMs is that they require low memory compared to other approaches, due to using merely a portion of the training data (Witten et al., 2016). The main disadvantages are however the long training period, therewith being impractical for large data sets, its inability to accurately handle overlapping classes, and poor performance when the number of features is larger than the training sample size (Witten et al., 2016).

K-Nearest Neighbors (K-NN) is yet another method. It works by proximity and majority (or plurality) voting, which means that a data point is assigned the class label of the most frequently occurring class label of its neighbors. The main difference between this method and the previously described ones is that it is non-parametric, meaning that it does not assume a probability model regarding the outcome. Moreover, there is no training step involved and it is memory-based, so it immediately adapts when being presented with new data. Compared to the other methods, which require two or more hyperparameters to be set, with K-NN only one hyperparameter needs to be predefined, after which the rest aligns automatically (Witten et al., 2016). However, it is relatively slow, can suffer from the curse of dimensionality, requires homogeneous features, so with the same scale, is sensitive to outliers and does not perform well when data are imbalanced.

The Random Forest (RF) ensemble method is considered a superior method for binary classification. It combines multiple decision trees to determine the final classification (Witten et al., 2016). A single decision tree classifier works by maximizing the split-quality of subsetting the data, measured for example by Gini impurity (Witten et al., 2016). In other words, a decision tree works like a flow chart where simple yes-no questions (e.g. "is it sunny?") are answered to determine the final classification (Witten et al., 2016). In a Random Forest, these individual outcomes are then combined and the most frequently occurring class is identified as the final result (Witten et al., 2016). To ensure low correlation between decision trees and therewith minimize bias, a random subset of features – sampled by bootstrapping – is used for the individual trees, which is also known as *feature randomness* (Witten et al., 2016). Random Forest is robust to noise and outliers, has generally good performance and is suitable for large data sets due to the possibility to parallelize the training process (Witten et al., 2016). However, hyperparameter tuning should be performed with caution as the number and depth of trees are highly influential on the risk of overfitting and the duration of the training process. Its main disadvantages are that it is more black-box than for example a single decision tree and it requires relatively high memory usage.

The final method discussed here is the Multi-Layer Perceptron (MLP), which is a type of Neural Network. An MLP consists of a fully-connected input and output layer with one or more hidden layers in between. Input data is propagated forward through the layers by calculating the dot product of the input and the weights between the layers and then utilizing some activation function, most commonly the Rectified Linear Unit (ReLU) or a sigmoid function. Then, at the output layer, the resulting values will be either used for further improving the network during training or for making a final decision during testing. Its main limitations are that the extent

to which the independent feature is affected by the dependent (target) feature is unknown, that computations are not interpretable and time consuming, and that the functioning of the model depends on the quality of training (Witten et al., 2016; Akkaya, 2019). However, it has several benefits that can outweigh these limitations: it can be applied to complex, non-linear problems, it works well with large data due to the ability of quickly making predictions once the model is trained, and is robust to smaller data sets (Witten et al., 2016; Akkaya, 2019).

These standard methods are often adapted using different techniques, mitigating or nuancing some of the limitations of a particular approach. For example, bagging can be applied to ensure methods designed for small data sets can deal with larger sets as well (Breiman, 1999, 1996). For instance, standard Support Vector Machines can be transformed to an ensemble by bagging (with bootstrapping, i.e. drawn with replacement) or pasting (without bootstrapping). This creates an ensemble of classifiers for subsets of data, therewith reducing the variance of the estimator (Scikit-Learn Developers, 2023b; Breiman, 1999, 1996). Alternatively, boosting uses the idea of combining a collection of weak learners (e.g. decision trees), so classifiers that merely slightly correlates with the true classification, into one strong learner after a defined number of learning iterations (Zhang and Ma, 2012; Zhou, 2012). Here, the advantage is in using weights to direct later learners to the mistakes of previous learners, therewith minimizing bias and variance, similar to the aim of the bagging approach (Zhang and Ma, 2012; Zhou, 2012). The main difference between bagging and boosting approaches is that the former runs the classifiers in parallel, while the latter requires sequentiality in the learning process to ensure the classifiers learn from each other (Zhang and Ma, 2012; Zhou, 2012).

Table 2.1: Classification evaluation type ratios

Name	Ratio	Also known as
True positive rate	$\frac{TP}{TP+FN}$	sensitivity, recall, hit rate
True negative rate	$\frac{TN}{TN+FP}$	specificity, selectivity
Positive predictive value	$\frac{TP}{TP+FP}$	precision
Negative predictive value	$\frac{TN}{TN+FN}$	–
False negative rate	$\frac{FN}{FN+TP}$	miss rate
False positive rate	$\frac{FP}{FP+TN}$	fall-out
False discovery rate	$\frac{FP}{FP+TP}$	–
False omission rate	$\frac{FN}{FN+TN}$	–
Threat score	$\frac{TP}{TP+FN+FP}$	critical success index (CSI)
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	–
Balanced accuracy	$\frac{\frac{TP}{TP+FN}+\frac{TN}{TN+FP}}{2}$	–
F_1-score	$\frac{2 \cdot TP}{2 \cdot TP+FP+FN}$	harmonic mean of precision and sensitivity

sources: (Ting, 2011; Tharwat, 2021)

Evaluating model performance

The classification performance of these binary classifiers can be evaluated in different ways. Most of these metrics are based on some ratio of values in the so-called confusion matrix, which contains the number of truly positive (TP), truly negative (TN), falsely positive (FP, type I errors) and falsely negative (FN, type II errors) predicted outcomes given the observations. Here, positive and negative refer to the two categories, which are usually assigned 1 and 0, respectively.

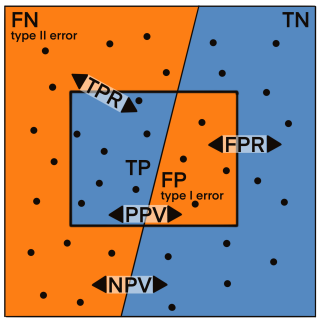

Figure 2.1: Illustration of the different outcome types and ratios in a confusion matrix

The basic ratios that are indicated in Figure 2.1 are true positive rate (or sensitivity or recall), positive predictive value (precision), negative predictive value, and false positive rate. The specific calculations of these and other ratios can be found in Table 2.1. Depending on the domain, there may be a preference for certain ratios, as the goals and type of data used usually differ across domains. For example, in medicine sensitivity and specificity are common, while in informatics precision and recall are preferred (Wikipedia, 2023). A metric that is not commonly used, yet possibly one of the best metrics for assessing a classifier's overall performance is the Matthews correlation coefficient (MCC) (Chicco and Jurman, 2020, 2023). This metric is one of few that provides an unbiased representation of the performance, due to taking into account the performance on all four parts of the confusion matrix:

$$\mathrm{MCC} = \frac{\mathrm{TN} \cdot \mathrm{TP} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TN} + \mathrm{FN}) \cdot (\mathrm{FP} + \mathrm{TP}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{FN} + \mathrm{TP})}} \tag{2.1}$$

While the F1 score can also be a suitable metric, especially when more emphasis is put on the class labeled as positive, MCC may be overall more desirable, because it allows for a more balanced assessment of the classifiers, which is especially desirable if the cost of having low precision and recall is unknown (Chicco and Jurman, 2020, 2023).

Cross-validation and methods to avoid overfitting

Overfitting is an important cause of mediocre or bad performance of supervised models on unseen data. It regards the idea of a model learning the seen (training) data too closely, leading to high performance in the training phase but low performance when testing the model on new (test) data due to the learning of general patterns being compromised by also learning the noise present in the seen data (Witten et al., 2016).

There are multiple ways to minimize (the risk of) overfitting. One method is dimensionality reduction or feature selection, which is described in more detail in Section 2.2.3. Another, relatively easy to implement method is *early stopping* (or *pruning* in tree-like models), which – as the name indicates – regards prematurely ending the training process before overfitting starts to occur (Witten et al., 2016). Early stopping is a procedure that is often already present as one of the hyperparameters in the model. In addition, regularization hyperparameters are often present in machine learning models, which constrain the parameter estimates, therewith penalizing redundant complexity in a model (Witten et al., 2016).

A data-splitting strategy to estimate the tuning parameters is taken by cross-validation, where the data are split into some collection of subsets, depending on the type of cross-validation (Witten et al., 2016). The most general form is K-fold cross-validation, where the data is split into k folds. The model is then trained on $k-1$ folds in each iteration, which totals k iterations (see Figure 2.2). That is, each fold is used for testing (validation) exactly once within the cross-validation process (Witten et al., 2016). Since the data are different in each iteration, the estimates of the validation error are more robust, providing information on how to tune the models away from overfitting (Witten et al., 2016). Before applying K-fold cross-validation, usually a portion of the data is temporarily removed to function as a leave-out test set to ensure part of the data remains completely unseen. Regarding the data used during the cross-validation, when k equals the number of observations, it is called leave-one-out cross-validation. Moreover, when k is equal to two it is named a train-validation-test split, where 'test' refers to the left-out test set (Witten et al., 2016). The latter is the simplest version, but comes with the limitation that it is relatively prone to high variance in cases where the data set is small (Witten et al., 2016). Conversely, leave-one-out cross-validation is most suitable for small data sets as it retains most of the training data. For example, if a data set consists of only twelve observations, the leave-one-out approach retains eleven observations for training, while three-fold cross-validation keeps only eight. K-fold cross-validation, however, is preferred over the leave-one-out approach as a large data set will have sufficient training samples when there are fewer folds than the number of observations. Moreover, without adjustments to avoid training the model N times (where N equals the number of observations), leave-one-out cross-validation becomes infeasible for large data sets in terms of the time it takes to run the procedure (Witten et al., 2016). For example, for a data set with 100 000 observations and a model that takes about two seconds to be trained using the leave-one-out split, it would take 200 000 seconds, which is just under 56 hours.

Using a similar idea, the risk of overfitting can also be reduced by turning the model into an ensemble so that additional randomness is added to the model, therewith mixing up the data and averaging the outcomes, leading to more stability in the results (Witten et al., 2016). Moreover, it is common practice to apply several of these methods simultaneously to further decrease the risk that the model fits the observed data too closely (Witten et al., 2016).

Figure 2.2: Example of a K-fold cross-validation procedure with five folds

The importance of scaling and encoding

Many supervised machine learning models are influenced by the variations in magnitude and range of the values in the data (Fitkov-Norris et al., 2012; Singh and Singh, 2020; Lanigan et al., 2023). Having targets with largely differing ranges of values can therefore corrupt the classification process due to larger values being assigned higher importance by the classifier, which introduces bias towards the variables with larger values in the model (Fitkov-Norris et al., 2012; Singh and Singh, 2020; Lanigan et al., 2023). Moreover, encoding is often required as most models are unable to deal with class data, or – when these classes are converted to numbers – the numbers are misinterpreted as

ordinal values instead of categorical ones (Fitkov-Norris et al., 2012). Therefore, scaling or encoding the data is a vital step that needs to be considered during the data preprocessing stage. In short, feature scaling refers to the transformation of features so that all the values in the data have a similar range and magnitude (Fitkov-Norris et al., 2012; Singh and Singh, 2020). Normalization is a scaling technique where the values are shifted in such a way that they lie in a fixed range of 0 and 1, and is also known as min-max scaling.

Another possibility is to standardize the data (Fitkov-Norris et al., 2012; Singh and Singh, 2020). This is also referred to as z-score normalization and rescales the data to a Gaussian distribution, so with a mean of 0 and a variance of 1. Moreover, there are several adaptations of the standard z-score normalization, which have been developed to better mitigate unwanted influence of outliers in the data (Brownlee, 2020). Of these two types of scaling, there is no definite superior technique: their suitability depends on the context and application of the machine learning model and the data. More specifically, z-score normalization is usually applied when the focus is on comparison of similarities based on distance. Normalization, on the other hand, is often preferred in the context of computer vision for normalizing pixel intensities and when using neural networks, as these algorithms typically require zero-to-one scaled data (Brownlee, 2020).

Scaling is only applicable to non-categorical data (Brownlee, 2020). However, when dealing with questionnaire data for example, features are often categorical even though they may be represented with a number. This number is nonetheless meaningless, as a category represented with number 10 is not twice as much as the category represented with number 5. So, to mitigate any unwanted influences when using such data for training machine learning models, a different approach should be taken. A well-known procedure is one-hot encoding, which unfolds each feature into (almost) as many features as there are categories, assigning a 1 to the observations where the specific category is present and 0 to all others. In other words, the features are encoded into dummy variables.

However, this becomes infeasible when dealing with large data sets containing many features with many categories due to the inherent sparsity of the matrix that is created with one-hot encoding, eventually leading to the curse of dimensionality problem. As such, an alternative is to target or mean encode the categorical features (Fitkov-Norris et al., 2012). That is, each category of a feature is replaced by a quantification of the effect it may have on the target variable (Fitkov-Norris et al., 2012). For a binary classifier, the standard approach is to calculate what is also known as the posterior probability in Bayesian statistics:

$$\mathrm{P}(y = 1 \mid x = c_i) = \frac{\mathrm{count}_{y=1,\ x=c_i}}{\mathrm{count}_{x=c_i}} \tag{2.2}$$

where y represents the target variable, x represent the categorical feature, and c_i indicates the i-th category. As shown in Equation 2.2, the probability of the target being equal to 1 given category i equals the number of times the target is 1 given this category, divided by the total number of times the category occurs in the observations for the respective feature.

This however introduces the problem of *target leakage*, meaning that the model is being presented information it should actually predict (Micci-Barreca, 2001; Fitkov-Norris et al., 2012). So, information of the target is now present in the observed feature(s). Nevertheless, target leakage can relatively easily be mitigated by adding prior smoothing to the standard encoding procedure

(Micci-Barreca, 2001). This is generally done by using the mean of all categories, instead of the mean from one category only:

$$\text{encoding} = \alpha \cdot \mathrm{P}(y = 1 \mid x = c_i) + (1 - \alpha) \cdot \mathrm{P}(y = 1) \tag{2.3}$$

where α is the smoothing factor, often defined by the following sigmoid function:

$$\alpha = \frac{1}{1+\exp\left(-\frac{n-k}{f}\right)} \tag{2.4}$$

where n is the number of observations, k determines the proportion of the data for which the estimate based on the sample is completely trusted, and f controls the transition rate between prior and posterior probability (Micci-Barreca, 2001).

So, when sufficiently mitigating target leakage, target encoding is a suitable alternative to the classical one-hot encoding which becomes infeasible with large data sets, as it keeps information on the predictability of the target given a particular category. Target encoding can also be extended to the multiclass case, as is described by Micci-Barecca (2001), but since this study focuses on binary classification only, this will not be discussed in detail.

Multicollinearity and feature selection

Another commonly influential factor that should be considered when training classification models is multicollinearity. Multicollinearity refers to situations where two or more explanatory variables closely relate to each other instead of being independent, the latter of which is the standard assumption. These close relations are also referred to as correlation or collinearity. The 'multi' part of the concept refers to the multitude of explanatory variables being collinear. For binary classification models, a serious consequence of multicollinearity is that it significantly decreases interpretability of the model (Masís, 2021; Lundberg and Lee, 2017). Even though this is not necessarily problematic for the classifier's prediction capability, for example extracting its feature importances becomes an unreliable procedure (Masís, 2021; Lundberg and Lee, 2017). That is, because the importance of collinear features either gets divided across these features, for instance when using a decision tree, or becomes zero when using a permutation approach due to the equivalent split (Masís, 2021). In any case, the calculated feature importance will deviate significantly from the actual importance.

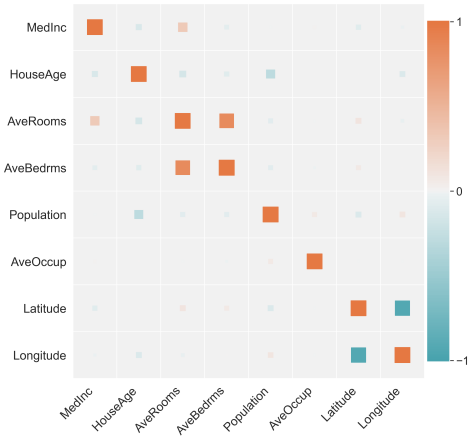

Figure 2.3: Example of a correlation heat map to visualize (multi)collinearity

So, to avoid problems with model interpretability in a later stage of the classification process, it is vital to apply an appropriate form of feature selection, where as much unique information as possible is retained without 'duplicating' it. Usually, this is done after cleaning and scaling the data, but before training a classifier (Witten et al., 2016). These types of feature selection are also referred to as *filter* methods because of their independence to the machine learning model.

However before utilizing any feature selection, determining whether (multi)collinearity is present in the data can be beneficial. Multicollinearity can be detected by plotting a (Pearson) correlation matrix, where high positive and high negative values are considered collinear, while values around zero are considered independent. Usually, for better readability, these correlation matrices are visualized in the form of a correlation heat map, an example of which is shown in Figure 2.3. Here, larger squares indicate a higher correlation value, while the colors indicate not only the size but also the direction of this correlation: large, positive correlations are dark red, while large, negative correlations are dark blue. Another option, which is frequently used in regression analysis, is to calculate the Variance Inflation Factor (VIF) or tolerance, which is the inverse of VIF (James et al., 2013). Tolerance is defined as follows: $1 - R_i^2$, where R_i^2 represents the unadjusted determination coefficient when regressing the i-th independent variable on all other independent variables (Faraway, 2002). Besides not being directly relevant for classification tasks, calculating VIF on high-dimensional data becomes problematic due to the long time it takes to iteratively calculate all factors on the data (Faraway, 2002). Hence, more standard approaches like the aforementioned correlation matrix are often considered more suitable in the context of classification.

One method is to apply some clustering algorithm to extract similar features from the data and use a single feature per cluster for training and testing the model. One clustering approach that can be used to perform dimensionality reduction is hierarchical clustering with Spearman rank-order correlations (Scikit-Learn Developers, 2023a). This type of correlation represents the dependence between the rankings of two variables, or in other words, how well the relationship between two variables can be described using a monotonic function, regardless of the type – linear or non-linear – of relationship (Myers and Well, 2003). Hierarchical clustering focuses on constructing a dendrogram, which is a tree-like structure explaining the relationships between the features in the data, as can be seen in Figure 2.4. In this example, a threshold of 1.1 is considered, which results in three clusters of features, indicated by the three different, colored rectangles. After calculating the Spearman correlation and constructing the dendrogram, a threshold should be defined to determine the clusters used for feature selection (Scikit-Learn Developers, 2023a). This threshold is generally determined by a manual, visual inspection of the dendrogram, as herein the number of clusters can be easily distinguished. If a small number of features is desired, a higher threshold value is more suitable and vice versa. After determining the threshold, for each resulting cluster, one feature is selected and added to the final data set.

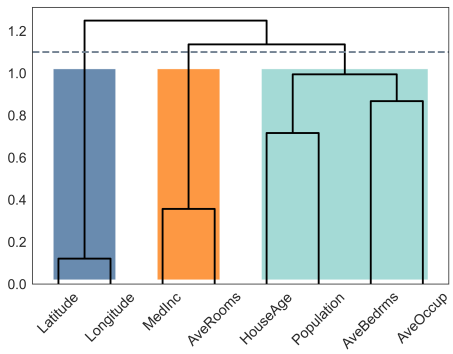

Figure 2.4: Example of a dendrogram to visualize (hierarchical) clusters of similar features in the data

Besides strengthening interpretability and reducing the risk of overfitting, applying feature selection or dimensionality reduction is also beneficial for improving computational complexity and training time (Witten et al., 2016). For example, the training time complexity of a Random Forest is $O(n \cdot \log(n) \cdot d \cdot k)$ with n the number of observations, d the dimensionality (number of features), and k the number of decision trees (Witten et al., 2016). Then, ceteris paribus, a twice as large dimensionality (d) would lead to a doubling of the training time complexity. While

this may not seem significant for smaller data sets, such increases in time complexity become problematic when dealing with large data sets (Witten et al., 2016).

Automatic hyperparameter optimization

Machine learning models usually contain a myriad of tunable parameters, often referred to as hyperparameters, which determine a model's characteristics and therewith influence performance and complexity. For instance, an essential hyperparameter in Random Forests is the maximum allowed depth of a single tree within the ensemble. Since the number of possible values for each hyperparameter in a model is generally immeasurable, it is infeasible to manually try and find the best set of values for the hyperparameters. This is due to the machine learning model, when considering it as a function, being *black-box* because the algebraic form is unknown (Lipton, 2016). As such, automatic hyperparameter optimization algorithms have been developed.

Two types of methods are Bayesian Optimization (BO) and Swarm Optimization (SO). The former is a constrained, sequential optimization approach that solves the problem of finding the optimal set of parameters by building a probability model of the objective function to select and evaluate the hyperparameter values in the true objective function (Mockus, 1989). Swarm optimization algorithms, on the other hand, are based on the collective behavior of decentralized systems in nature, applied in an artificial context. For example, Particle Swarm Optimization (PSO) simulates social behavior of birds in a flock. Tani and Veelken (2022) compare PSO with BO in a machine learning context and find that both perform well, but that PSO is more suitable for larger data sets, while BO is superior for smaller data sets (Tani and Veelken, 2022). Moreover, PSO does not have any significant computational overhead, while this is the case for BO, rendering the swarm algorithm even more suitable for large data sets.

Another SO algorithm is the Artificial Bee Colony (ABC) algorithm, which is based on the foraging behavior of honey bees. ABC is a swarm- or population-based meta-heuristic algorithm developed by Karaboga and Basturk (2007). It has essentially four components: 1) the food sources (representing solutions), 2) employed bees (representing agents searching randomly), 3) onlooker bees (representing the selection of being the best solution with greater probability), and 4) scout bees (representing replacement of abandoned food sources, i.e. local minima). The algorithm uses an initial candidate solution that is then iteratively optimized by going through the different phases, namely initialization, employed bee exploration, onlooking, and scouting. In general, ABC is more effective compared to BO and PSO (Li et al., 2012; Karaboga and Akay, 2009). Like PSO, it remains efficient on large data sets while barely compromising performance (Karaboga and Basturk, 2007; Li et al., 2012). In Zhu et al. (2017), ABC has been applied to SVM and compared with the more classical Least Squares (LS)-SVM combination. Here, it is shown that both optimizations provide nearly the same results. Moreover, in recent years, the standard ABC algorithm, like many other optimization algorithms, has been improved by for instance allowing multiple dimensions being updated in an iteration (Alizadegan et al., 2013). Such improvements are desirable, because the standard ABC algorithm focuses on exploration capacity over exploitation capacity, which may lead to slower convergence speed and sub-optimal solution accuracy (Wang et al., 2022; Karaboga and Akay, 2009).

2.2.4 Model explainability

With increasing complexity of machine learning models, there is also an increasing need to develop methods that explain how these models determine their outputs. This has led to the development of the research field of Explainable Artificial Intelligence (XAI), which focuses on transforming black-box models into so-called white-box models, increasing their transparency and understandability among other things (Turek, 2023). Within this field, many different methods have been developed for a variety of applications, like image detection. However, the methods discussed in this section are described in the context of tabular data classification only, as this is the focus of the current study.

One approach is to calculate the permutation feature importance, which is based on the decrease in a model score when a single feature value is randomly shuffled, thereby breaking the relationship between the feature and the target (Scikit-Learn Developers, 2023a). So, the larger the decrease in the model performance score, the more important the feature. However, this procedure is extremely sensitive to (multi)collinearity, leading to an overestimation of the importance of correlated predictors (Nicodemus et al., 2010; Archer and Kimes, 2008; Strobl et al., 2007). Consequently, several alternatives have been proposed, such as permute-and-relearn and dropped variable importance (Mentch and Hooker, 2016; Lei et al., 2016). However, other studies show that the limitations of the classic permutation importance procedure persist in these alternatives (Hooker et al., 2019; Vorotyntsev, 2020).

A better alternative for extracting feature importance, according to Hooker et al. (2019), may be calculating the Shapley values, given that this procedure is undertaken carefully to avoid the same extrapolation bias present in permutation importance. For instance, Shapley values are unable to handle one-hot encoded data and highly correlated features (Amoukou et al., 2021). The calculation of Shapley values is based on game theory: it quantifies the contribution of each 'player' (i.e. each feature using a single observation) to the 'game' (i.e. a single prediction made by the model) (Lundberg and Lee, 2017). Since each possible combination of features is considered to determine the importance of a particular feature, this set of combinations to be considered can be defined as a power set. That is, a power set of a set C is the set of all subsets of C, including the empty set $\varnothing$ and the original set C itself. More specifically, the SHAP formula, which is Shapley applied in the machine learning context, uses weighted marginal contributions of the features to determine the final feature importance of a particular feature:

$$\text{SHAP}_c(x) = \sum_{C_c \in C} \left(|C| \cdot \binom{N}{|C|} \right)^{-1} \cdot \left(\text{Prediction}_C(x) - \text{Prediction}_{C \setminus c}(x) \right) \tag{2.5}$$

with c the feature, C_c a subset of the original set C, N the total number of features, and Prediction(x) a single prediction (Lundberg and Lee, 2017).

The first part of Equation 2.5 represents the weighting of the marginal contributions, which are represented in the second part of the equation. As indicated here, the marginal distribution is defined by the difference of two predictions where one prediction has one additional feature, ceteris paribus.

The main drawback of this exact procedure is however computationally intractable as, given N features and k samples, the time complexity is $O(k \cdot N \cdot 2^N)$. As such, applications of SHAP

in a machine learning context are usually approximations instead of exact calculations (Lundberg and Lee, 2017; Lundberg, 2018).

While extracting feature importance is one of the most popular approaches to making a machine learning model more explainable, there are other options. One of these options is the *class map*, which is a state-of-the-art visualization tool that provides information on how a classifier has made its decisions. More specifically, the class map gives insights in erroneous predictions in relation to the difficulty of an observation, thus when a model is uncertain (Raymaekers et al., 2020). By using a combination of explainable AI and machine learning tools, individual limitations can be mitigated and artificial models can be made more explainable, interpretable and transparent (Turek, 2023).

Chapter 3

Methods

3.1 The methodological framework

This study uses an adaptation of the SEMMA framework, which is one of the popular methodological frameworks used for data mining. SEMMA stands for the five phases embedded in this methodology: Sample, Explore, Modify, Model, and Assess (Plotnikova et al., 2020). The first stage regards determining the data to be used for analysis. Next, in the *explore* stage, a preliminary analysis on the data is conducted to examine the characteristics, interdependencies, and possible issues of the data. This is also often referred to as *exploratory data analysis*. The results from this stage are then used to clean the data accordingly during the *modification* stage, after which the data are proceeded to the modeling stage. Here, data mining techniques are applied to the data to acquire the desired type of outcome from the data, like binary classification of the observations. Lastly, in the *assess* stage, model evaluation takes place using test data.

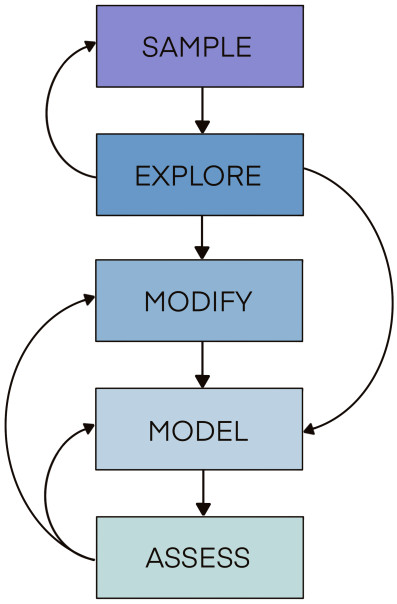

Figure 3.1: The adapted SEMMA framework used in this study

This framework is however not applied one-to-one, because it does not take into account the iterative nature of data-scientific research. As such, this standard methodology is adapted slightly to better fit the characteristics of this study and provide more accurate guidance. More specifically, feedback loops from the exploration stage to the sampling stage and from the assessment stage to the modification and modeling stages are added and a feedforward pathway is added from the exploration to the modeling stage. This is illustrated in Figure 3.1. The first feedback pathway is added to allow changing which data are used for analysis, as exploratory analysis commonly exposes shortcomings in the data used, rendering it unsuitable for certain research goals. The second feedback pathway – from assessment back to modification and modeling – allows for making changes in the data modification and modeling procedure based on the findings from testing the models with the given modified data. For example, these findings might indicate errors in a data scaling procedure in the modification stage or show shortcomings in the used model. The feedforward pathway supports the use of information gathered in the exploratory analysis phase to develop suitable models for the data. For instance, if the data are imbalanced, one may choose to use an ensemble method (e.g. Random Forest) instead of a single sampling method (e.g. Decision Tree), as the former is shown to be

more robust to imbalances in the data.

The results are then primarily used to analyze the importance of the observed features for predicting a given target. Next to this, to provide a more complete and in-depth analysis given that this study mainly functions as a basis for further research, the results are also used to investigate the potential link between feature importance, model performance, and missingness of the observed and target variables.

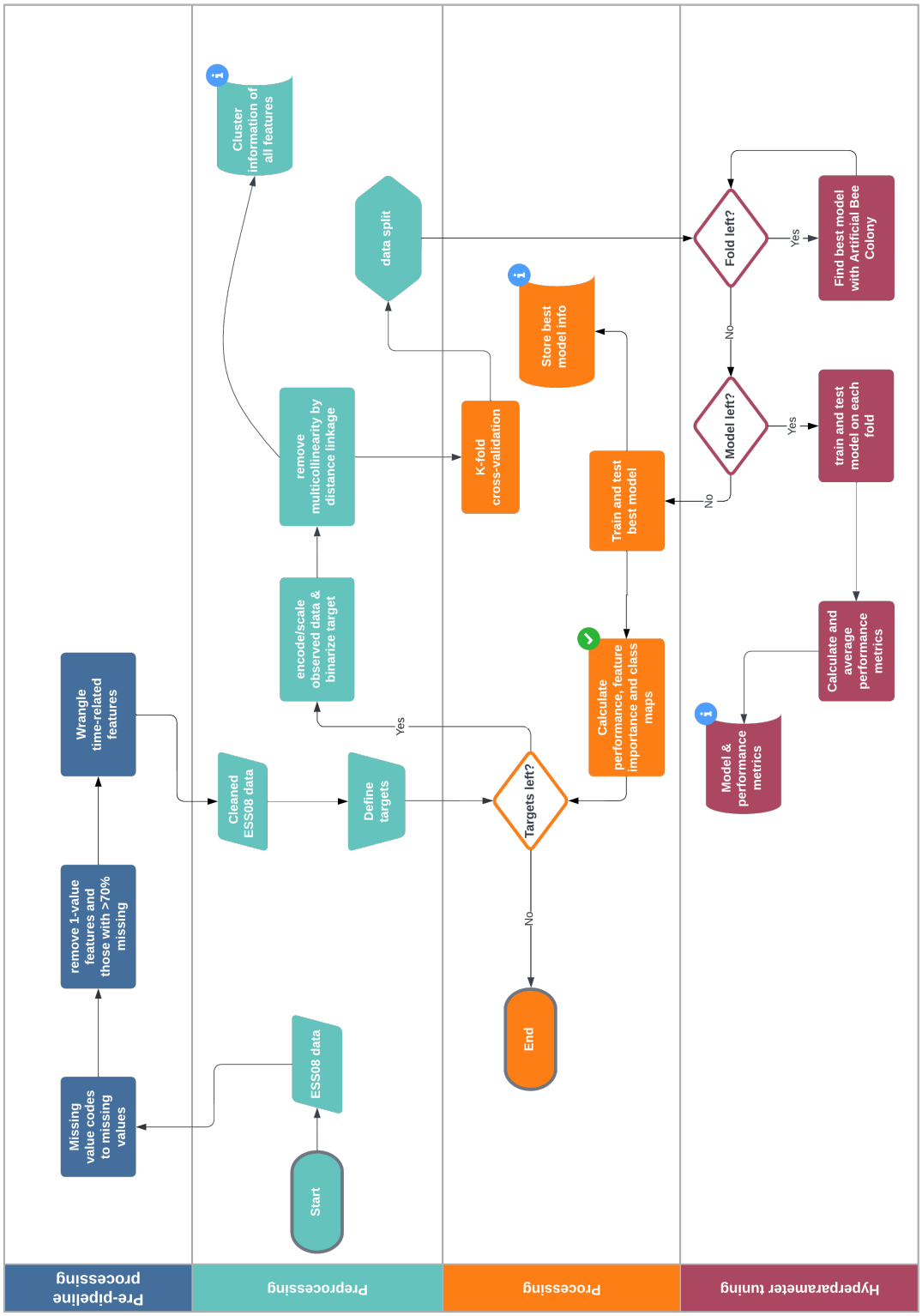

Figure 3.2: Flow chart illustration of the processing pipeline implemented in this study

3.2 Processing pipeline

In this study, the process starts by preprocessing the original ESS08 data set to make it suitable for classification of the missingness. This procedure is split into two parts, one part of which takes place before the actual processing pipeline of the classification models (see Figure 3.2). The other part takes place inside this pipeline due to the multiverse nature of the study. That is, as described in more detail in Section 4.2, first a general cleaning is applied to the original data, where the missing value codes are converted to actual missing values, the time-related features are wrangled, and features with a constant value and those with more than 70% missing values are removed. The 70% threshold is chosen by a combination of a qualitative and quantitative analysis of the features' missingness and a feature's potential importance.

After this *pre-pipeline processing*, indicated by the dark blue boxes in Figure 3.2, the targets are defined. This can be seen as the start of the actual processing pipeline, where each target definition represents a different data set, in accordance with the multiverse approach taken. For each of these data sets, the set is first pushed through a second round of data cleaning, where the observed data are scaled and encoded based on the nature of the variable, and the values of the target feature are binarized into missing (0) and non-missing (1). For the categorical variables, target or mean encoding as described in Section 2.2.3 is applied, while for the few numerical features standard scaling is implemented. The main reasons for using target encoding instead of one-hot encoding or feature collapsing are that one-hot encoding feeds the curse of dimensionality problem in these data sets – due to the large number of features – and that feature collapsing is undesirable in the context of multiverse analysis – having the goal to minimize reasearcher degrees of freedom – and from an ethical point of view, due to the added subjectivity when only using a self-defined share of the most frequent or most important categories while simplifying the remainder to one residual category.

After this, to mitigate the curse of dimensionality problem and more importantly, to remove any multicollinearity that may bias the classification models and feature importance calculations, feature selection is applied by using hierarchical clustering (with a threshold of 1.0) in combination with Spearman's correlation as described in more detail in Section 2.2.3. Then, three-fold cross-validation is performed to balance computational expenses and acquiring more stable estimates of the validation error. Here, the data are first randomly, non-stratified, split into 33% test data and 67% training data, after which stratified K-fold cross-validation is performed on the training data. Here, stratification is used to ensure each fold is representative of all strata in the data, therewith mitigating potential bias in the case of imbalanced data (Refaeilzadeh et al., 2009). In each fold, the best model, measured by the Matthews correlation coefficient, among a range of models is searched for by the Artificial Bee Colony algorithm. After doing this three times, i.e. for three folds, among the three best models, the overall best model is extracted and used for final testing on the left-out test set. Lastly, the feature importances from the overall best model are calculated by applying SHAP (see Section 2.2.4). For the Random Forest classifier, this is done by the tree explainer using the complete data. However, for the Multi-Layer Perceptron and the Support Vector Machine classifiers, a kernel explainer is required, rendering it impossible to use the full data set due to time constraints. Hence, a subset of the data is used. For the Multi-Layer Perceptrons, a random sample from the training set with 500 observations is used to train the

kernel explainer and 300 randomly drawn observations from the test set are used to determine the feature importance. These quantities are based on time-wise feasibility as this procedure has to be run for each of the sets in the collection of data sets. Similarly, for the Support Vector Machines randomly drawn samples of size 75 and 35, respectively, have been used for the kernel explainer. The latter sample sizes likely limit generalizability of the findings, but it is infeasible to increase these sizes due to the long time it takes to run the SHAP kernel explainer on the Support Vector Machine models. The pseudocode of the three-fold cross-validation and the adapted version of the ABC algorithm can be found in Algorithms 1 and 2.

Throughout the processing pipeline, important findings are stored for later use. This regards information on the clusters defined in the feature selection procedure, performance metrics of and information on the best models of each fold and the best model overall, the best models themselves, and the feature importances. More specifically, the performance metrics stored are Matthews correlation coefficient, F1 score, ROC score, balanced accuracy, and accuracy. Moreover, class maps are made to get more insight into the uncertainty and difficulty of observations when it comes to the classification process, as described in more detail in Section 2.2.4. First, the observations are assigned a color based on the class predicted by the model. Next, the probability of an observation belonging to the opposite class (PAC) and the localized farness (LF) are calculated. This is done by extracting the fitted model probabilities for each of the two classes.[1] The local farness is computed by defining the nearest neighbors using the kernel density tree, which is a K-NN algorithm designed for fast conversion regarding N-point problems, with Euclidean distance as metric and using the Epanechnikov kernel weighting function to weigh the local distances and calculate the corresponding class probability ($\mathrm{P}(i \in g_i)$) (Epanechnikov, 1969). Then, the localized farness is calculated as $1.0 - \mathrm{P}(i \in g_i)$. Lastly, for the purpose of visualizing the class maps later on, in the dashboard (see Appendix B), for each model, the class, PAC, LF, color, model, and class performance are stored for each observation. The (Python) implementation for the binary class version of the class maps can be found in Appendix D. Regarding the feature importance, the results are stored in both a raw format, where only the subset of features for each target is present, and a wrangled version, where the cluster information is used for de-clustering to ensure all features are present. Here, features within the same cluster are assigned the same importance as they are similar in pattern and therewith have more or less the same explanation of missingness of the target.

This information is used to compare the results of the different models to provide an initial set of (potential) predictors and to gain insight into how this multiverse analysis can contribute to the identification and archiving of important predictors of non-response.

Table 3.1: Hyperparameter configurations of the SVM models

Hyperparameter	Min value	Max value	Pre-set value
C	0	1	–
γ	0	1	–
Number estimators	–	–	20
Max samples	–	–	0.05

[1] This can be extended to the multiclass case as well.

Table 3.2: Hyperparameter configurations of the RF models

Hyperparameter	Min value	Max value
Number estimators	1	500
Max number features	1	number features - 1
Max depth	2	50
Min samples for split	2	10
Min samples for leaf	2	10

3.2.1 Specifications of the models and the ABC algorithm

The Random Forest classifier has several hyperparameters. In this study, a subset of the most important hyperparameters to tune is defined, namely 1) the number of estimators, 2) the maximum number of features, 3) the maximum depth for each tree, 4) the minimum number of samples needed for a split, and 5) the minimum number of samples needed for a leaf (a single node). The parameter space for each hyperparameter within which the ABC algorithm searches for the best model, for the Random Forests and the other two classes of models, can be found in Tables 3.1, 3.2, and 3.3. The tables provide the ranges for the automatically tuned hyperparameters and the pre-set value for hyperparameters that are neither searched for by the ABC algorithm nor use the default value. These ranges are as wide as possible along with taking into account plausible ranges, therewith balancing the time needed to run the algorithm and acquiring a well-performing model. Moreover, for hyperparameters that are not in the 0-to-1 scale, the values are scaled to be between 0 and 1 when used as input to the ABC algorithm because it leads to improved performance of the searching process. The output values are then unscaled to their original value and plugged into the model for training.

This is also done for the *number of units* (i.e. the sizes of the hidden layers) hyperparameter in the Multi-Layer Perceptrons. Other hyperparameters tuned by the ABC algorithm with respect to the Multi-Layer Perceptrons are the dropout rate, the learning rate, the exponential decay rate for first moment vector estimates, and the exponential decay rate for second moment vector estimates. Moreover, the activation function is set to the sigmoid function and the optimizer is set to Adam, one of the go-to stochastic gradient-based optimizers commonly used for classification, especially for large data sets. The number of hidden layers is set to four, the batch size is set to 2048, and the number of epochs is set to 150. These are determined by manually examining different sets of values which have been chosen based on the characteristics of the data and the objective: there are around 20 000 samples used for training the model, so a relatively large batch size is possibly most suitable and the objective is binary classification, indicating relatively shal-

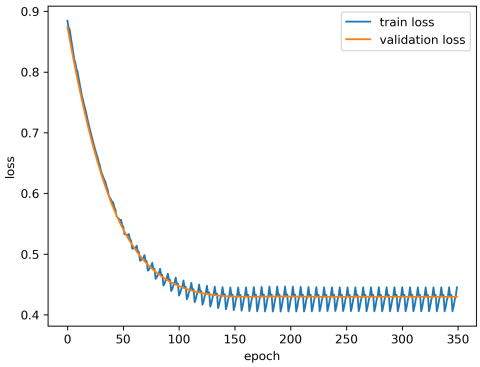

Figure 3.3: Plot of the model loss of a preliminary MLP model (classifying *occf14b*) during the construction process, used to determine a suitable combination of batch size and number of epochs

Table 3.3: Hyperparameter configurations of the MLP models

Hyperparameter	Min value	Max value	Pre-set value
Hidden layer sizes	20, 10, 5, 3	200, 100, 50, 25	–
L2 regularization	0.00001	0.01	–
Initial learning rate	0.0001	0.01	–
β_1	0.6	0.999	–
β_2	0.6	0.999	–
Activation function	–	–	sigmoid
Optimizer	–	–	Adam
Batch size	–	–	2048
Epochs	–	–	150

low networks should already be able to solve it. Moreover, a suitable number of epochs – given a certain batch size and number of hidden layers – has been determined by plotting the number of epochs versus the loss and analyze at what point the performance starts to plateau. In Figure 3.3, it can be seen that – given a batch size of 2048 and four hidden layers – the loss plateaus at approximately 150 epochs. Additionally, early stopping is turned on, which leads to 10% of the data being used for validation within the training process.

Lastly, for the Support Vector Machines, the hyperparameters tuned by the ABC algorithm are the only two influential hyperparameters: the regularization parameter C and the radial basis function kernel coefficient γ. As described in Section 2.2.3, here, bagging in the form of pasting is applied to minimize bias and overfitting in the model. Moreover, the number of base estimators is set to 20 and the maximum number of samples is set to 0.05 (i.e. 5% of the observations), where all features can be used for training a base estimator. These values are based on a manual trial-and-error procedure, again to balance performance and the time it takes to run the models.

For the ABC algorithm, a hive size (i.e. the number of *bees*) and the maximum number of iterations is set to 50. This is also determined using a manual trial-and-error process, where likewise the focus has been on acquiring the best possible performance without getting infeasible run times.

Algorithm 1: K-fold cross-validation with the Artificial Bee Colony algorithm

Input: k (number of folds), X (the observed data), y (the target data), $max_{iterations}$ (maximum number of iterations), *hive* (number of bees), *dim* (number of parameters to search), l (lower bounds parameter space), u (upper bounds parameter space)

Output: $v_{importance}$ (feature importances)

$X_{train} \subset X$, $X_{test} \subset X \wedge X_{test} \cap X_{train} = \emptyset$ s.t. $X_{test} \cup X_{train} = X$
$y_{train} \subset y$ s.t. $X_{X_{train,i}} = y_{y_{train,i}}$, $y_{test} \subset y \wedge y_{test} \cap y_{train} = \emptyset$ s.t. $y_{test} \cup y_{train} = y$
$c = 0$, $m = 0$
for $c = 0$ *to* $c = k - 1$ **do**
    $X_{train_c} \subset_{stratified} X_{train}$, $X_{test_c} \subset_{stratified} X_{test}$
    $y_{train_c} \subset_{stratified} y_{train}$ s.t. $X_{X_{train_c,i}} = y_{y_{train_c,i}}$, $y_{test_c} \subset_{stratified} y_{test}$
    pars_c = ArtificialBeeColony($hive$, $max_{iterations}$, l, u)
for $m = 0$ *to* $m = k - 1$ **do**
    for $c = 0$ *to* $c = k - 1$ **do**
        train and validate model m with pars_m on X_{train_c}, y_{train_c}, X_{test_c} and y_{test_c}
for $m = 0$ *to* $m = k - 1$ **do**
    train and test model m with pars_m on X_{train}, X_{test}, y_{train}, y_{test}
    $v_{importance,m} = |\text{SHAP}(\text{model } m, X_{train}, X_{test}, y_{train}, y_{test})|$
return $v_{importance}$

3.3 Visualization methods

One of the difficulties that comes with multiverse analyses is how to best present the many results and insights that are produced by the analyses. To communicate the collection of results in a clear way while retaining as much detail as possible, the results are presented in an interactive dashboard.[2] This dashboard consists of different graphs and plots suitable for the results at hand, following the principles described by Munzner (2014) and others (see Section B.1 for a summary of the main principles). Moreover, guidance on how to use the dashboard and how to interpret the results is provided in the form of textual descriptions. For completeness, the main results are also provided in a (sortable) tabular format, for easy exploration and browsing. The dashboard is coded using *Plotly* and *Dash* components, combined with HTML markup language.[3] This combination allows the desired level of flexibility in designing a custom dashboard, given the large collection of results. Before coding the dashboard, first a sketch is made to pre-visualize desired visualizations, their possible interactions and encodings. This allows verifying and justifying different designs without spending too much valuable time on the coding part (Munzner, 2014; Ware, 2014). The final sketch can be found in Appendix A.1.

Next to the dashboard, where possible, results are aggregated to summarize the overall findings. Similar to the approach taken with the dashboard, these aggregated results are visualized if this enhances comprehension of the findings and corresponding interpretation. Again, the decisions behind the choices made in terms of design and type of visualization are mainly based on Munzner (2014).

[2]The dashboard can be accessed here: `http://multiverseanalysisvisualizer.pythonanywhere.com/`.
[3]The code for building this dashboard, along with all other code made for this project can be found here: `https://github.com/Lieve2/ADSthesis_multiverse_analysis.git` (Göbbels, 2023).

Algorithm 2: Pseudocode of the adapted Artificial Bee Colony algorithm

Input: $max_{iterations}$ (maximum number of iterations), *hive* (number of bees), *dim* (number of parameters to search), l (lower bounds parameter space), u (upper bounds parameter space)

Output: solution (best set of values for hyperparameters)

$m = 0$, $h = 0$, $dim = 3$, $\text{trial}_{max} = 0.6 \cdot hive \cdot dim$
for $h = 0$ *to* $h = (hive + hive \bmod 2) - 1$ **do**
  $\text{colony}_i = \text{bee}_h$
$\text{solution}_i^{\mathbb{C}} i = 1, ..., hive$ s.t. $l \leq \text{solution}_i \leq u$ `// D-dim initial solution vector`
$\text{fitness}_i = 1 - MCC_i$, $\text{fitness}_{best} = \text{fitness}_i$ `// initialize fitness`
$l = 0$ `// set cycle number to 0`
$\text{bees}_{employed} = \text{colony}_{0,hive/2}$, $\text{bees}_{onlooker} = \text{colony}_{hive/2,hive}$ `// classic 50/50 split`
for $s = 0$ *to* $s = max_{iterations} - 1$ **do**
  for $bee_{employed}$ *in* $bees_{employed}$ **do**
    $d =$ randomly selected dimension `// produce new solution`
    $v_{i,d} =$ mutation on d-th dimension of $\text{bee}_{employed}$ and randomly selected other bee s.t. $l \leq v_{i,d} \leq u$
    if $dim > 2$ **then**
      `// adaptation from original:`
      `// update 3 dimensions each iteration`
      $d_2 =$ randomly selected second dim, $d_3 =$ randomly selected third dim
      $v_{i,d2} =$ mutation on $d2$-th dim of $\text{bee}_{employed}$ and randomly selected other bee s.t. $l \leq v_{i,d2} \leq u$
      $v_{i,d3} =$ mutation on $d3$-th dim of $\text{bee}_{employed}$ and randomly selected other bee s.t. $l \leq v_{i,d3} \leq u$
    $\text{fitness}_{i,bee_{employed}} = 1 - MCC_{bee_{employed_j}}$ `// fit model on current solution`
  $p_i = 0.9 \cdot \text{fitness}_i / \max(\text{fitness}) + 0.1$ `// calculate solution probabilities`
  for $bee_{onlooker}$ *in* $bees_{onlooker}$ **do**
    $\text{solution}_{new} = \text{solution}_i$ where $i = \max(p)$ `// select new solution based on` p_i
    `// produce new solution`
    $v_{i,d} =$ mutation on d-th dim of $\text{bee}_{onlooker}$ and randomly selected other bee s.t. $l \leq v_{i,d} \leq u$
    if $dim > 2$ **then**
      `// adaptation from original`
      $d_2 =$ randomly selected second dim, $d_3 =$ randomly selected third dim
      $v_{i,d2} =$ mutation on $d2$-th dim of $\text{bee}_{onlooker}$ and randomly selected other bee s.t. $l \leq v_{i,d2} \leq u$
      $v_{i,d3} =$ mutation on $d3$-th dim of $\text{bee}_{onlooker}$ and randomly selected other bee s.t. $l \leq v_{i,d3} \leq u$
    if $fitness_{i,bee_{onlooker}} > fitness_{colony_i}$ **then**
      $\text{fitness}_{i,bee_{onlooker}} = 1 - MCC_{bee_{onlooker_j}}$ `// update fitness if best so far`
      $t = 0$ `// reset trials counter`
    else
      $t = t + 1$
  if $t_{bee} > trial_{max}$ **then**
    $v_{bee} =$solution$_i$ where $i = \max(p)$ `// abandon solution exceeding trial limit`
if $fitness_{bee} > fitness_i$ **then**
  $\text{fitness}_{best} = \text{fitness}_{bee}$ `// store current best`
  $\text{solution} = \text{solution}_{bee}$
  $s = s + 1$
return *solution*

Chapter 4

Data

4.1 The European Social Survey 2016

The European Social Survey (ESS) is an academically motivated, international survey across a multitude of European countries, where face-to-face interviews are conducted every two years using newly selected, cross-sectional samples. The goal of this survey is to measure different kinds of behavioral and social patterns across European nations (European Social Survey, 2023a).

The main objectives of ESS are assembling, interpreting and publicizing accurate data on the social conditions in European countries, providing access to these data in a timely and free manner, and continuously improving methods and analysis for quantitative social measurement. All of the data are collected on a national level by Computer-Assisted Personal Interviewing (CAPI) interviews, which are then aggregated. During the data collection, each country is required to provide case-level information on fieldwork progress (i.e. the most up-to-date data set available) on a weekly basis to allow fast response to potential problems that may be occurring. After the data collection period, the final data sets, including meta- and paradata (contact forms), are collected and finalized by the ESS Core Scientific Team (European Social Survey, 2023b).

The ESS 2016 version is the eighth execution of this project and covers 23 different countries (ESS Round 8: European Social Survey Round 8, 2016). The focus of the face-to-face questionnaire is on questions regarding the topics of politics and trust, attitudes toward sexual and ethnic minorities, social behavior, religion, background, energy supplies and climate (change), social benefits (e.g. pension or child care) and employment, attitudes toward the European Union, and education (European Social Survey, 2023c).

4.2 Exploratory analysis and pre-processing

In total there are 535 variables and nearly 45 thousand observations present in the original data set. Most of these variables are categorical, representing classes of attitudes towards social queries or groups to which the respondents belong. For example, one variable contains information on how much control the respondent has on how their daily work is organized, on a scale from zero (no influence) to ten (complete control). Besides categorical variables, there are 13 numerical variables, including time spent on the internet in minutes and the duration of the interview in minutes.

After converting missing value codes to missing values in a python-readable format, following the descriptions in the accompanying codebook, the average amount of missing data is 53.5%. Moreover, of the 535 variables there are in total 50 variables that have no missingness at all. As can be seen in Figure 4.1, variables that contain missing values generally have either a low proportion of

missingness or an extremely high one. Since too high proportions of missingness render the variable uninformative regarding the task of classification, even though the focus of this study is on the missingness within variables, it is decided to remove all variables that have more than 70% missing values. Moreover, there are several variables that have one constant value, which is likewise uninformative. After removing the variables with high missingness and with constant values, 277 variables remain.

Next, additional cleaning is done with respect to the variables related to interview duration. Here, missing durations are imputed by calculating the duration from the start- and end-of-interview time variables where possible. For a few observations this has been impossible, so that for these cases random imputation is used to impute the missing duration. This means that, from the list of possible durations, for each observation a duration value is chosen at random and imputed into the data. After this imputation, only the interview duration variable is kept in the data and the start- and end-of-interview time variables are removed since the information is already present in the duration variable. However, the variables regarding the start and end day, month, and year are kept as these might carry information about the missingness of certain variables. For example, in early May 2016 Italy legalized same-sex civil unions, which may have led to respondents in Italy being more open to answering such questions compared to before early May (Squires, 2016). This pre-pipeline processing leads to a total of 273 features being used in the remainder of the processing pipeline, as illustrated in Figure 3.2.

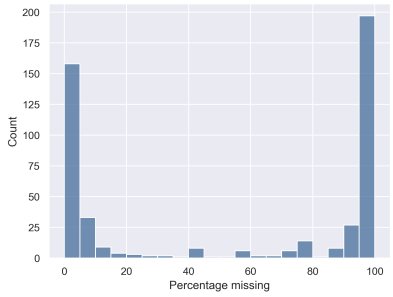

Figure 4.1: Count plot of the missingness distributions in the original data, excluding complete variables

The first step of the in-process preprocessing phase (the lightblue, second lane in Figure 3.2) is to define the targets, which is done by extracting all features that have more than 5% missing, corresponding to 74 features. Each target feature then represents a different data set and will lead to a different model used for the classification. Next, for each data set, the observed variables are encoded, imputed and scaled, and the values of the target are converted into missing (0) and non-missing (1). For the numerical variables, missing values are imputed by random imputation and scaled using z-score normalization. For the categorical variables, missing values are assigned as a new class and encoded using the target encoding approach as explained in Section 2.2.3. Moreover, to minimize multicollinearity in the data and reduce the dimensionality of the data, hierarchical clustering with distance linkage is ap-

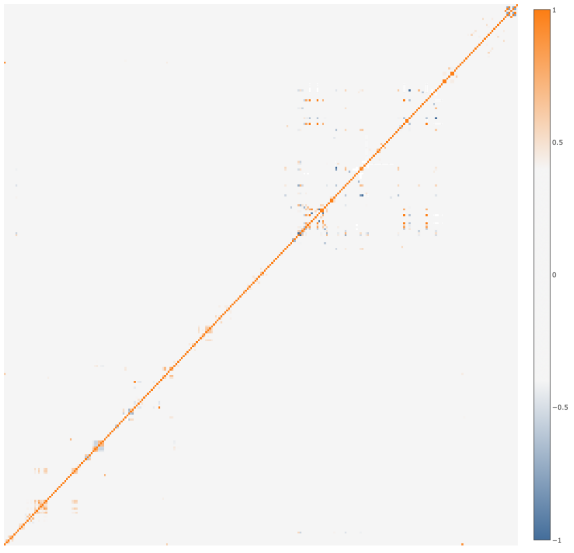

Figure 4.2: Correlation heat map of the original data

plied. For this, Spearman rank order correlation is calculated on the data, after which a distance matrix is constructed and the hierarchical clustering with Ward's method is conducted. Then, of each cluster, only the first feature is kept and used for the classification task. This is necessary, because there indeed exists multicollinearity in the data, as can be seen in Figure 4.2.[1] This figure shows highly correlated features in dark orange (high positive correlation) or dark blue (high negative correlation), indicating (multi)collinearity.

[1]Note that in this heat map, the names of the features are left out due to legibility issues that would occur otherwise. However, a similar heat map is present in the dashboard, where the feature names are visible, to allow for more detailed analysis of the (multi)collinearity.

Chapter 5

Results

An overview of the most important results for each target can be found in Appendix E, where the results are summarized in a tabular format. Moreover, the meanings of the feature abbreviations and the proportion of missingness for each feature can be found in Appendices G and H, respectively. To allow assessing how overlap in missingness between the predictors and the target may be of influence to the results, in Appendix I, a list with the average and maximum overlap regarding each target is provided. The key findings from this collection of results will be described in the following sections.

Table 5.1: Average performance RF, SVM, and MLP models

Metric	**RF**	SVM	MLP
Accuracy	**0.943**	0.921	0.937
Balanced accuracy	**0.823**	0.784	0.799
ROC score	**0.823**	0.784	0.799
F1 score	**0.955**	0.934	0.951
Matthews coefficient	**0.713**	0.624	0.641

5.1 Model performance

As described in Chapter 3, the performance of each model is measured using several metrics: accuracy, balanced accuracy, ROC score, F1 score, and Matthews correlation, with a focus on the Matthews correlation due to its unbiased representation of performance. Assessing and comparing model performance allows for a more complete interpretation of (the validity of) the feature importance later on.

Overall, the Random Forest models perform best, with an average Matthews correlation coefficient of 0.71. This is followed by the Multi-Layer Perceptrons, which have an average Matthews correlation coefficient of 0.64, and the Support Vector Machines, with an average coefficient of 0.62. An overview of the average performance for all of the metrics for each group of models can be found in Table E.1.

Looking at the performance for each of the 74 models in Figures 5.2, 5.3, and 5.4, it can be seen that for the Random Forests, 24 models have a performance that is lower than the defined threshold (0.7) for which a model is considered having decent performance. In other words, 50 out of the 74 models (68%) in the Random Forest class have decent performance. For the Multi-Layer Perceptron class, there are 45 models (61%) with decent performance and for the Support Vector Machine class there are merely 29 models (39%) performing good enough. Moreover, it can be seen that all three classes of models have a similar pattern, indicating that they generally perform poorly on the same targets. However, where the Multi-Layer Perceptrons and the Support Vector

Machines sometimes have a Matthews correlation coefficient of exactly zero, implying a random prediction, Random Forest models never have a Matthews coefficient that is greatly lower than 0.2. For the Support Vector Machine class, this occurs for three targets: *bnlwinc* (Attitude towards "social benefits should only be for people with the lowest incomes"), *lbenent* (Attitude towards "many with very low incomes get less benefit than legally entitled to"), and *occf14b* (Occupation of father of respondent when respondent was 14). Furthermore, for the Multi-Layer Perceptrons, this occurs for 13 different data sets, including *bnlwinc* and *lbenent*, but not *occf14b*. For the targets where the Multi-Layer Perceptrons have a Matthews coefficient of zero, the average percentage missing values is 9.3 and none of the targets have more than 19% missing values, representing a highly imbalanced data set. Similarly, the three targets where the Support Vector Machines have a coefficient of zero, have an average missingness of 10.6%, never surpassing 14.5% missing values.

Looking at the class maps for the models predicting the target *bnlwinc*, it can be seen that when predicting class 0 (missing), all of the observations are relatively difficult, with a localized farness ranging between the 50% quantile and the 95% quantile (see Figure 5.5). This is not the case for the observations belonging to class 1 (non-missing), that all have relatively high proximity to each other. Here, the models are able to relatively accurately predict the true class 1 observations. However, similar to what is indicated by the Matthews correlation coefficient, the class maps show that the Multi-Layer Perceptron and Support Vector Machine assign all observations to be of class 1 and the Multi-Layer Perceptron does so with equal probability for all observations (i.e. they all have a 75% chance of belonging to class 1, indicating random assignment of a class instead of actually predicting the class given the observed information).

The average overlap between the missing values in the target and those in a predictor are between 0.5% and 4.4%, with the highest percentage overlap a predictor has with its target being 56.3%. Moreover, there are merely five occurrences of a predictor having 40% or more overlap with its target, while each of these five targets has a proportion of missingness ranging between 56% and 67%. One of the targets with relatively low overlap (2.9% on average) is *rlgdnm*. This target – where 41% of the values are missing – is accurately classified by the models (see Figure 5.7).

Regarding the models with good performance, there is a set of 29 common targets for which the models all perform well. The missingness proportions of these targets range from nearly 60% to just under 7%, as can be seen in Figure 5.1. One of the targets that is accurately predicted by all three models (0.97-0.99 Matthews correlation coefficient) is what religion or denomination the respondent has ever belonged to, if any (i.e. *rlgblge*). Here, the class maps for class 0 show that all three models have some observations that they are uncertain about (close to the midpoint line), while occasionally observations are wrongly assigned to class 1 with high certainty (probability close to 1), as can be seen in Figure 5.6. For the Random Forest, these are merely the more difficult observations, while the other two models also misclassify simple observations. However, most

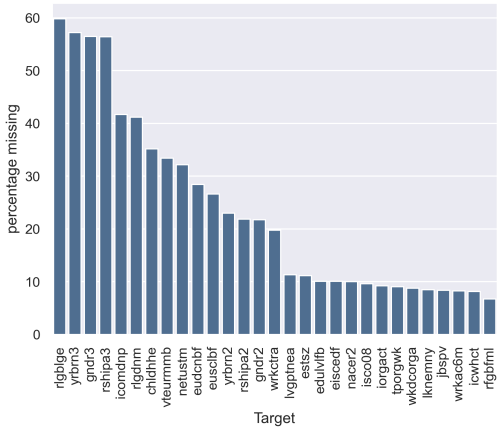

Figure 5.1: Proportions of missing values for the 29 well-classified targets

observations are located in the bottom, indicating a correct prediction with high certainty (probability close to 0), even for some of the more difficult observations (i.e. those that lie more to the right). Moreover, for the Random Forest model, a clear cluster with uncertain observations can be distinguished and, compared to the Support Vector Machine and Multi-Layer Perceptron, it has relatively few incorrect predictions with high certainty. Looking at class 1, both the Multi-Layer Perceptron and the Random Forest have hardly any incorrect predictions with high certainty, while the Support Vector Machine does. Additionally, the Multi-Layer Perceptron correctly predicts most, both easy and difficult, predictions with extremely high certainty. Both the Random Forest and the Multi-Layer Perceptron only misclassify relatively difficult observations, but this is not the case for the Support Vector Machine. Similar to the class-0 maps, here too, most observations are located at the bottom, indicating correct predictions with high certainty.

Table 5.2: Table of the five most informative features on average for each model type when all 74 targets are considered, ranked in descending order

	RF	**SVM**	**MLP**
Rank	Feature (SHAP value, % missing)	Feature (SHAP value, % missing)	Feature (SHAP value, % missing)
1	region (0.038, 0%)	dvrcdeva (0.034, 0.6%)	eiscedf (0.041, 10.1%)
2	regunit (0.038, 0%)	yrbrn2 (0.034, 23.0%)	dvrcdeva (0.040, 0.6%)
3	psppsgva (0.037, 2.2%)	rshpsts (0.033, 41.7%)	yrbrn2 (0.039, 23.0%)
4	rfgbfml (0.037, 6.7%)	eisced (0.029, 0.3%)	wkhtotp (0.038, 67.3%)
5	pdwrk (0.035, 0%)	domicil (0.029, 0.1%)	uemplap (0.038, 0%)

5.2 Feature importance

Similar to creating an overview of the performance of each model by averaging the results across all models, this is also done for the feature importance. When taking into account all targets and the de-clustered features (see Chapter 3), the three overall most important features for the Random Forests are *region* (region of the respondent), *regunit* (regional unit of the respondent), and *psppsgva* (attitude towards "the political system allows people to have a say in what the government does"). For the Support Vector Machines, these are *dvrcdeva* (whether ever divorced), *yrbrn2* (birth year of second person in household), and *rshpsts* (relationship status). Similarly, for the Multi-Layer Perceptrons, the *dvrcdeva* and *yrbrn2* features are in second and third place, respectively, but the on average most importance feature is *eiscedf* (highest level of education of father in ES-ISCED scale). All of these features have SHAP values around 0.05, indicating that they – on average – change the prediction by approximately 0.05 compared to using a baseline value of the respective feature. An overview of the five most important features for each model can be found in Table 5.2.

Figure 5.2: Performance in Matthews correlation coefficient and F1 score for the 74 Random Forest models

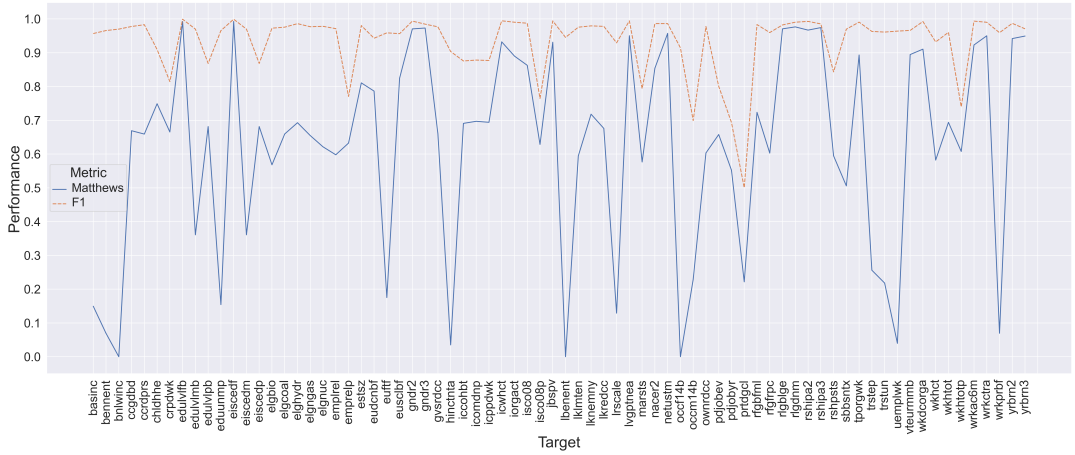

Figure 5.3: Performance in Matthews correlation coefficient and F1 score for the 74 Support Vector Machine models

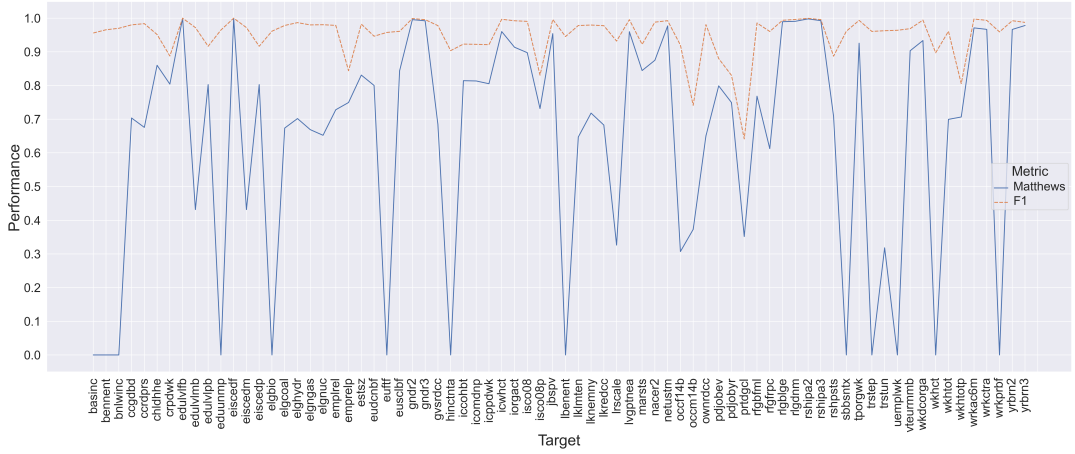

Figure 5.4: Performance in Matthews correlation coefficient and F1 score for the 74 Multi-Layer Perceptron models

Figure 5.5: Class maps of the *bnlwinc* target

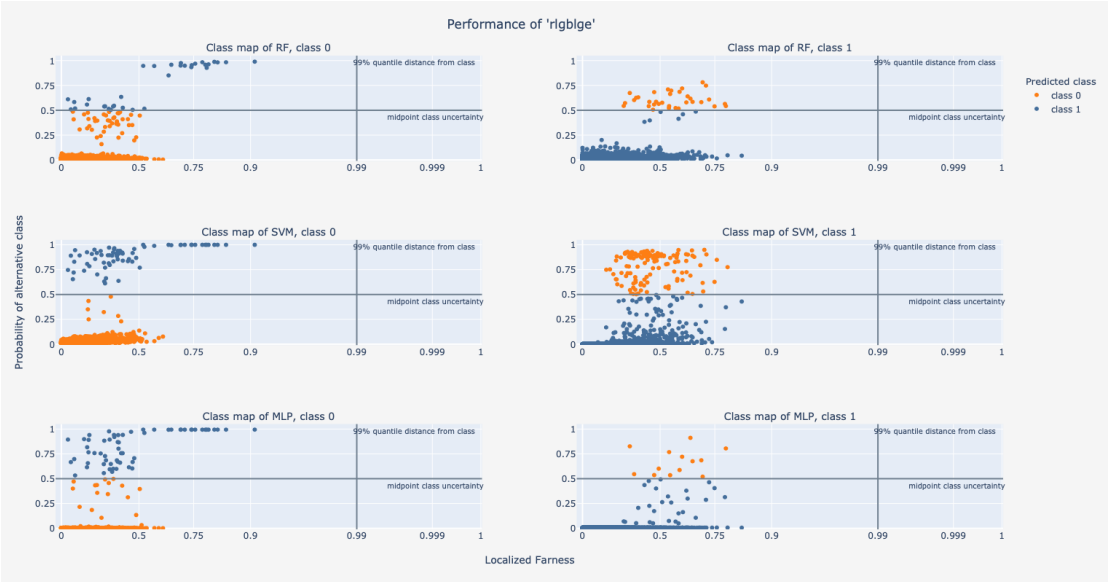

Figure 5.6: Class maps of the *rlgblge* target

For the class-0 maps (left), blue is wrongly assigned to class 1 with a certainty defined by the value on the y-axis. The closer to the 0.5 line, the higher the uncertainty. The difficulty of an observation is indicated by the location on the x-axis, where further to the right indicates higher difficulty. Similarly, for the class-1 maps, orange-colored plots are wrongly assigned to class 0, a y-value of 0.5 indicates high uncertainty and a large value on the x-axis indicates high difficulty.

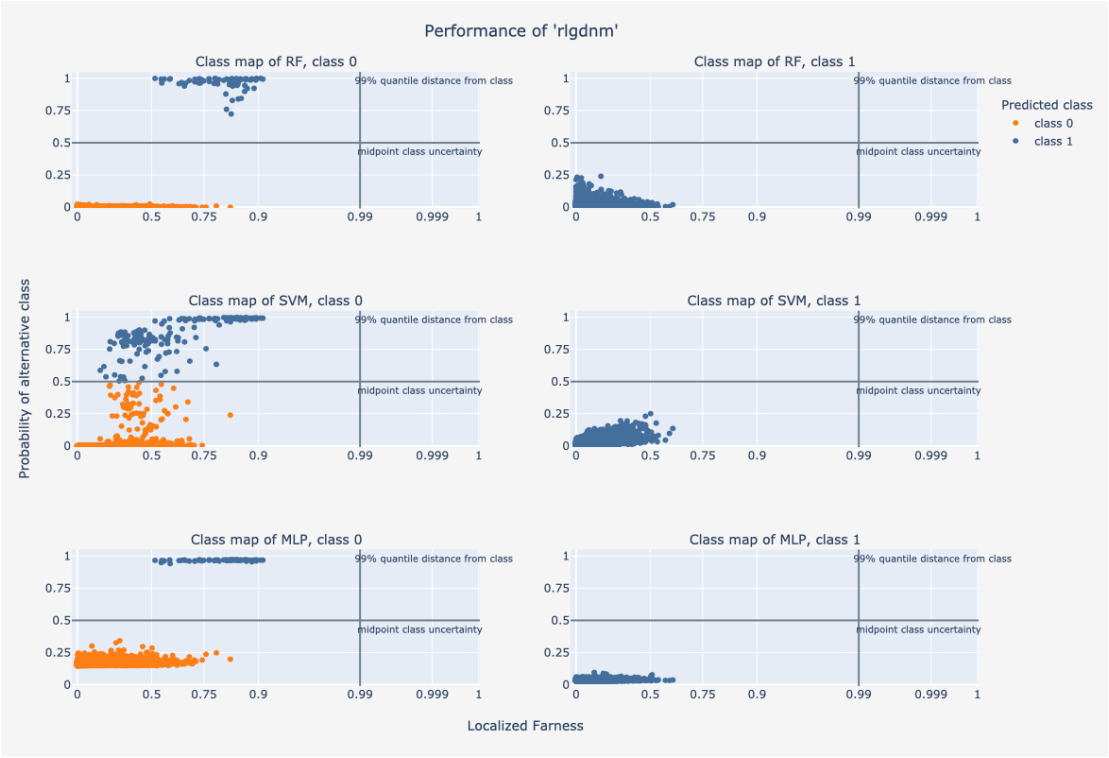

Figure 5.7: Class maps of the *rlgdnm* target

As indicated in Section 5.1, however, not all models perform well, and this may corrupt SHAP calculations. So, the average has also been calculated on the subset of targets with good performance for all three models. Here, the five most important features are identical, albeit in a different order, for all three models: *yrbrn2*, *pdwrk* (whether respondent has done paid work in 7 days before the interview), *dvcdeva*, *eduyrs* (years of fulltime education completed), and *icpart2* (whether the respondent lives with husband, wife or partner, in interviewer code). The average SHAP values for the features of this subset of classifications are slightly higher, ranging from 0.043 to 0.065. An overview of the top five important features and the corresponding SHAP values can be found in Table 5.3.

Using again the *bnlwinc* target as example for assessing the feature importance of poorly performing models, the results show that neither the Support Vector Machine nor the Multi-Layer Perceptron have any features that are important to their classification decisions. More specifically, the feature with the highest SHAP value for the Support Vector Machine is *netustm* (time spent using the internet on a typical day) with an importance of 0.003 and for the Multi-Layer Perceptron, all features have an importance of exactly zero, which coincides with the previously described corresponding class maps. That is, since the Multi-Layer Perceptron classifies all observations to class 1 with equal probability, there is no distinction between classifications of the individual observations. For the Support Vector Machine, however, there is still some variation in the probabilities of belonging to a certain class, indicating that some information from the observed variables is used to distinguish between individual observations. Conversely, for the Random Forest, there are two visibly important features: *psppsgva* and *gndr* (gender of the respondent). These features

Table 5.3: Table of the five most informative features on average for each model type when only good (Matthews coefficient ≥ 0.7) models are considered, ranked in descending order

	RF	**SVM**	**MLP**
Rank	Feature (SHAP value, % missing)	Feature (SHAP value, % missing)	Feature (SHAP value, % missing)
1	pdwrk (0.049, 0%)	pdwrk (0.059, 0%)	eiscedf (0.068, 10.1%)
2	eiscedf (0.048, 10.1%)	yrbrn2 (0.058, 23.0%)	dvrcdeva (0.065, 0.6%)
3	eisced (0.047, 0.3%)	dvrcdeva (0.053, 0.6%)	yrbrn2 (0.065, 23.0%)
4	domicil (0.047, 0.1%)	eduyrs (0.047, 1.0%)	wkhtotp (0.063, 67.3%)
5	icpart2 (0.045, 0.4%)	icpart2 (0.043, 0.4%)	uemplap (0.062, 0%)

have a SHAP value of 0.036 and 0.023, respectively.

Looking at the feature importance of a well-classified target, *rlgblge*, all three models assign high importance to the observed feature *dscrrce* (whether or not respondent's group is discriminated in terms of color or race), which has a SHAP value ranging between 0.442 and 0.475, indicating that this feature changes the prediction by nearly 0.5 compared to using the baseline value. Moreover, this observed feature has no missing values, meaning that, for certain, the true values are important to the classifications of the models. Similarly, the feature *psppsgva* (SHAP value of 0.247), which is the second-most important feature for the Random Forest model, has a low missingness rate of 2.15%. The other models do not have any other visibly important features for classifying the missingness of the target regarding what religion or denomination the respondent has ever belonged to.

Chapter 6

Discussion

6.1 Discussion of the results

The results regarding the model performances show that the Random Forest class of models performs best overall, with a Matthews coefficient of 0.71. This is the only class of models that, on average, meets the 0.7 threshold for which a model is considered decent enough, as the Support Vector Machine and the Multi-Layer Perceptron have a Matthews coefficient of 0.62 and 0.64, respectively. Moreover, assessing whether or not the 'decent performance' threshold is met for each specific target, it turns out that 50 out of the 74 models meet this threshold for the Random Forests, while the Multi-Layer Perceptrons and Support Vector Machines are only decent in 45 and 29 cases, respectively. While this confirms the theory stating that Random Forests are especially suitable for binary classification, these results also indicate the apparent difficulty of predicting item non-response.

To further investigate this apparent difficulty regarding predicting the missingness of a share of the set of investigated targets, class maps have been constructed, and results to analyze the potential link between model performance, proportion and overlap of missingness of the targets and the observed variables have been gathered and described in Chapter 5. For the models that have a Matthews correlation coefficient that is equal to exactly zero, the observations have only been assigned to one class, which indicates random behavior rather than evident use of the information from the observations. This may be due to essential information regarding the missingness of the target being unobserved, rendering the MAR assumption – which is by definition assumed during the classification process – implausible. This relates to a second potential cause indicated by the results: fading of the associations due to the categorical imputation. That is, by introducing unobserved values in the form of an additional class, for features with a large proportion of missingness, this additional class and thus the missingness, may diminish the overall association between the missingness of the target and the respective feature, leading to lower predictive ability of the observed values in this feature.

For most of the models, the class maps for class 0 show that the observations are relatively hard and the models are uncertain about most of the observations, which at least partly explains the low Matthews coefficients for misclassified targets. However, comparing this to the class 1 maps, the observations are generally considered easy, and the models are uncertain about a much smaller proportion of the observations. This is especially the case when the data are imbalanced towards class 1 (i.e. when the proportion of class 1 is much higher than class 0), such as for the target *rlgdnm*, where 41% of the observations is class 0 and 59% class 1, which class maps show no erroneous and generally confident classifications for class 1, but some misclassifications regarding class 0. However, the inverse does not seem to hold entirely. When the proportion of class 0 is

higher than class 1, some uncertainty for a proportion of the observations remains and the class-0 observations are still seen as difficult. This is for example the case for the target *rlgblge*, which has 60% class-0 observations and 40% class-1. Here, the proportion of uncertain classifications is larger for class 0 than for class 1 and the class-0 observations seem to have an overall higher level of difficulty. On average, the predictors of *rlgdnm* have 1.9% overlap in missing values with the target, with the maximum overlap in missingness between a predictor and the target being 24.3%. On the other hand, the average overlap of missing values with *rlgblge* is 2.9%, with a maximum overlap of 33.0%. So, the remaining uncertainty may be due to the predictors having more missing data for observations when the outcome is also missing compared to observed cases. However, since there are merely four targets that have more class-0 than class-1 outcomes and because the maximum percentage missing of the targets is 59%, this cannot be stated with certainty and should be investigated further.

Interestingly, when looking at the table in Appendix E, many of the targets with bad classification regard attitude-related questions, especially regarding social benefits and income, education, and the European Union, rather than fact-related questions – these are often classified well. This may be because it can be more difficult to collect sufficient information about emotions, state of mind, and other subjective and personal aspects of the respondent.

For most of the models with bad performance, there are no features that have SHAP values above 0.05, indicating that the observed information is not profoundly influential to the classifications made. When considering only the well-performing models, the most informative features regard employment, education level, domicile, and household and partner information.[1]

So, these results are promising in the sense that they provide an initial set of possibly informative predictors concerning targets related to the topics of politics and trust, attitudes toward sexual and ethnic minorities, social behavior, religion, background, energy supplies and climate (change), social benefits (e.g. pension or child care) and employment, attitudes toward the European Union, and education.

6.2 Remaining limitations

While the aforementioned results are promising, some limitations of the study remain. One limitation is that using SHAP for determining the feature importances, while a popular approach, is not perfect. That is, it is highly sensitive with respect to (multi)collinearity and, for the non-tree-based models, uses approximations on a subset of data rather than calculations on the entire provided data set. While the former is taken care of by applying a clustering approach as feature selection, the latter is still likely to influence the results. Interestingly, however, the feature importances of the different model classes still often coincide in terms of ranking. So, this could be considered a minor limitation. Nonetheless, for future reseach it can be useful to implement multiple ways to extract feature importance and compare these results. This will further increase generalizability of the findings, which will benefit the gathering of important predictors across different contexts

[1] Within the thesis project group, similar studies have been conducted concerning different data sets. Interesting to note is that the study by Albert Banke (2023) also indicates that features carrying spatial or location information are important in predicting missingness for certain social science-related targets. This not only demonstrates the potential importance of including such features, but at a more general level, also illustrates the long-term purpose of these studies: to combine findings of different studies for determining the importance of certain features.

and domains.

Another limitation remains in the type of imputation used to complete the predictor variables before the modeling takes place. In this study, simple but limited imputation has been applied by adding an additional class to account for missing values in the categorical predictors and random imputation has been applied to the numeric predictors. However, this likely biases the estimated classifiers and could therefore fade the true associations between predictor and target. As such, for future research it should be taken into account that the choice of imputation method influences the acquired results. Thus, it may be desirable to apply a variety of imputation methods and compare the results to achieve a more robust collection of (potential) predictors of missingness.

A third limitation is that, even though a multiverse approach is taken, some researcher degrees of freedom are not (sufficiently) accounted for, which limits generalizability of the findings. That is, besides only using one data set, each model is only run on a single seed, so the results may differ when using different seeds. Moreover, while as many of the important hyperparameters are tuned automatically by the Artificial Bee Colony algorithm, this is not done for all hyperparameters and not on the entirety of the possible parameter space. Additionally, the hyperparameters of the ABC algorithm are set manually, which may lead to finding better models when being initialized differently, even though each decision is made taking into account existing theory and aims to balance the performance-time trade-off, as described in Chapters 3 and 7. Similarly, the choice of different models could also lead to different results. While the latter described researcher degrees of freedom are harder to account for, further research could for instance focus on applying similar multiverse procedures as the one proposed in this study to different but related data sets (to verify the feature importances) or use different seeds (to verify model performance and feature importances). It is especially vital that the feature(s) (types) suggested to be important in the ESS08 data set are verified with other, related data sets before an archive can be constructed, because the importance lies not in the applicability on a single data set, but on general (multi)domain applicability, so that new, non-existing data sets can be constructed where the MAR mechanism can be safely assumed. This also mitigates the current limitation, that MAR can not always be safely assumed in the ESS08 data, by including different information and possibly combining the (important) features of the data used in this study with related data to obtain a more exhaustive set of features.

A more general limitation of the multiverse approach is that it limits going into detail with respect to the analysis of the results due to the large collection of results. So, results are mostly described in a more aggregated or summarized format, which may lead to interesting target-specific patterns being missed, even though the results of the 74 data sets have been studied in as much detail as possible given the available time. Nevertheless, while this can be considered a limitation, at the same time it meets the purposes of this study, which is to provide an initial set of (potentially) essential predictors for a range of target variables and to show how multiverse analysis can contribute to the creation of an archive of essential predictors. By enabling constructing a collection of models in a flexible and streamlined way due to the possibility of using different types of models on a sequence of data sets, the multiverse approach turns out to be an overall suitable method for guiding the process of identifying predictors of non-response.

6.3 Moving forward

Since the results indicate possible bias due to the type of imputation methods used, as a first subsequent analysis, the random forest pipeline is adjusted to use multiple imputation and is run for a subset of 64 targets from the total of 74 targets. This provides a first idea of how the choice of imputation method can influence the process of identifying predictors of non-response in a multiverse analysis context. More specifically, instead of adding an additional class and performing random imputation, MICE with Bayesian ridge regression (the default in the used package), three imputation rounds and *most frequent* as initial imputation strategy, is implemented. The results, which can be found in Appendix F, show an average Matthews correlation coefficient of 0.97. Moreover, all models have good performance as the lowest Matthews coefficient is 0.90, which is for the model with the target *wkhtot*. Here, the most important features over the 64 classification models regard again employment (*jbspv*, *estsz*), education (*eisced*), household and partner information (*gndr2*, *agea*, *gndr*, *maritalb*). However, the prime important feature is stated to be *stfedu*, which carries information on how satisfied the respondent is with the educational system in their country. So, where the models with single imputation seem to be unable to extract informative patterns from the features carrying subjective information, such as attitudes or beliefs, the models using multiple imputation, on the contrary, appear to be able to do so. Similar to the models using the data imputed with the simple imputation methods, the (ten) most important features of Random Forest models using multiple imputed data have low missingness: 4.7%, on average. Thus, the idea of imputations fading the actual associations between potential predictor and target still remains a possibility and should be investigated further.

Accordingly, while the type of features considered informative for predicting missingness does not seem to differ much between the (Random Forest) classifications done with the single and the multiple imputation approach – even though the specific variables often differ – this basic postliminary analysis demonstrates the relevance of conducting multiple, comparable studies to acquire the level of robustness in the results that is needed for the successful construction of an archive of essential predictors of non-response.

Chapter 7

Ethical considerations and implications

Ethics in data science projects, like this study, regards among other things a correct use of data and models, and transparency regarding used methods and acquired results. With more advanced machine learning models being deployed every day, the importance of taking into account ethics during data science-related projects becomes increasingly clear (Mittelstadt et al., 2016; Mantelero, 2022). As such, during each stage of this study, from gathering the data to assessing the different models, the ethical aspects have been taken into consideration and appropriate action has been taken where applicable. In the following, the different ethical aspects considered in this study are described, following roughly the structure of the Fundamental Rights and Algorithms Impact Assessment (FRAIA) (Gerards et al., 2022).

7.1 Stage 1: Why?

This first stage regards validating the use of an algorithm by defining the objectives of the study and why an algorithm is necessary. For this study, the aim is to work towards the creation of an archive of essential predictors of missingness in a wide range of data sets to guide researchers in the design of their studies. An algorithm is desirable, if not required, to perform the embedded task of classification in determining which predictors are essential in a given data set. Moreover, the predetermined approach – the multiverse analysis – requires analysis of a large body of data, rendering it a complex and multi-step process, for which the use of algorithms is the more desirable and efficient choice.

7.2 Stage 2: What?

The second stage regards assessing the algorithm and its input, that is, the data. While the overarching class of algorithms has already been defined by the task at hand, there remains a wide range of algorithms or models still available. More specifically, the implemented prediction models are all binary classification models. From this class of models, the Support Vector Machine, Random Forest, and Multi-layer Perceptron have been chosen, because they have different deficiencies and levels of complexity, allowing for an interesting comparison as one aspect of the multiverse approach. However, a major drawback is that all three methods are considered black-box models, meaning that their inherent calculations and decision-making is incomprehensible for humans and thus ethical aspects like transparency and explainability are corrupted. To counteract this, be-

sides calculating feature importances which is required given the objective of the study, class maps have been added to further enhance explainability and transparency of these black-box models. As explained in more detail in Section 2.2.4, this provides information on both the uncertainty of the model for each separate observation in the data and relates this to the difficulty of that observation. This allows gaining insight into with what certainty misclassifications are made and on what type(s) of observations. Moreover, to minimize biased results due to class imbalances, the Matthews correlation coefficient has been used as main performance metric, but at the same time, multiple other, more common metrics have been provided. This is to ensure the results are comprehensive for researchers and students from different fields that are used to different metrics and to provide an exhaustive picture of the model performances in a more general sense. As explained in Sections 2.2.3 and 3.2, the Matthews correlation coefficient is one of few metrics that provides an unbiased indication of model performance, taking into account performance on all four classes of the confusion matrix and thus also unbalanced classification targets.

Regarding the ethical aspects surrounding the data, it has been decided to use open-source, anonymous data to avoid any potential issues regarding privacy and security. Moreover, the methodology used by ESS explicitly aims to minimize non-response bias by enhancing balanced response rates across groups and interviewer effects by researching how interviewer behavior can affect measurement and representation of the target population and continuously improve the surveying upon these results (European Social Survey (ESS), 2023a). Nonetheless, since the data collected only spans the European population, it is likely there is some extent of (geographical) bias present in the data, which should be taken into account when interpreting the results provided by this study. That is, for instance, features that are considered essential for predicting a certain (group of) target(s) may not at all be essential for continents other than Europe, and vice versa, due to differences in cultures, ways of living, and social rules and norms. Moreover, bias may also be introduced in the translation process when combining the different national data sets into one European data set. That is, in each participating country the questions and responses are (mostly) asked in the national language and translated to English, which may introduce subjectivity in the meaning of a question or response by the translator. However, it is unlikely this translation introduces large levels of bias, because most questions require an answer given based on a certain scale and the few open questions all regard quantitative answers, like the time spent on the internet in minutes. Moreover, ESS is aware of this potential bias and therefore conducts ongoing research on the best ways to provide translations with minimal bias (European Social Survey (ESS), 2023b).

7.3 Stage 3: How?

This stage regards what is done with the results, including what decisions are based on them. As indicated before, the results are meant to give an indication of what features can be important in predicting missingness patterns in a range of data sets with the aim to ensure researchers design their studies in a correct way so that applying state-of-the-art imputation techniques for missing data handling provides unbiased imputations. The main potential effect of the algorithm is thus higher data quality in social science-related studies (and possibly beyond). Moreover, the long-term aim is to validate these findings with other, similar studies conducted on different data sets to provide more robust results regarding feature importance.

Additionally, to adhere to open-source principles and open communication about the functioning of the algorithm, the entire project code is made available. This allows researchers, students, and other interested individuals to assess the code and decide how to use the findings of this study for themselves. With similar reasoning, the dashboard has been constructed with a focus on explainability, interpretability, and transparency of the results and minimizing difficulties in legibility for individuals suffering from color blindness.

7.4 Stage 4: Fundamental rights

The fourth and last stage of the FRAIA road map regards the assessment of safeguarding fundamental rights. Since 1) there are no personal data used, 2) the data are collected in an ethical way, according to the European Social Survey, 3) lack of model explainability and transparency is restored as much as possible by including additional methods that provide information on the reasoning of the classification models, 4) a dashboard is constructed to provide explainable, interpretable, and transparent results for a wide range of potential users, and 5) all code is freely accessible, the algorithm, and this study in a more general sense, are unlikely to have a direct and significant influence on fundamental rights.

Chapter 8

Conclusion

This study has investigated which variables are predictive of missingness in certain target variables in the context of European behavioral and social pattern-data and how this multiverse approach can contribute to the long-term goal of building an archive of universally informative predictors of missingness to ensure researchers from the social sciences and beyond can design their studies in such a way that MAR can be assumed safely and advanced imputation techniques can be used validly. To do so, a multiverse analysis approach is taken to predict the missingness in a multitude of targets from the ESS08 data set and extract the most informative features with respect to this missingness. More specifically, for each target the best model for each type of model – Random Forest, Support Vector Machine, and Multi-Layer Perceptron – is sought by the automatic Artificial Bee Colony tuning algorithm and cross-validation, and the feature importance is calculated using SHAP. With an average Matthews correlation coefficient of 0.71, the Random Forest models perform best, but all three classes provide approximately similar results with respect to the feature importance. Even though these classification models do not always perform well, a consensus on the types of variables that are informative for a set of targets – which regard the topics of politics and trust, attitudes toward sexual and ethnic minorities, social behavior, religion, background, energy supplies and climate (change), social benefits (e.g. pension or child care) and employment, attitudes toward the European Union, and education – can thus be distinguished from the results. More specifically, the results suggest the importance of including variables concerning employment, education level, domicile, and household and partner information, especially when dealing with more fact-related variables.

Though limitations remain in accounting for the researcher degrees of freedom and the missing data in the observed variables, this study provides an initial set of important predictors in the context of social science-related data and shows that multiverse analysis can adequately guide the process of identifying predictors of non-response by enabling constructing a collection of models in a flexible but streamlined way, due to the possibility of using different types of models on a sequence of data sets. This can benefit the successful construction of an archive of informative predictors, due to multiverse pipelines, like the one proposed in this study, being easily adaptable to different contexts and purposes, allowing researchers from different fields to contribute to the construction of this archive.

Bibliography

B. Akkaya. Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases. *y-BIS 2019 Conference: Recent Advances in Data Science and Business Analytics*, 1 2019. URL `https://www.academia.edu/41940316/Comparison_of_Multi_class_Classification_Algorithms_on_Early_Diagnosis_of_Heart_Diseases`. 8

E. C. Alexander. Don't Know or Won't Say? Exploring How Colorblind Norms Shape Item Nonresponse in Social Surveys. *Sociology of Race and Ethnicity*, 4(3), 2017. doi: 10.1177/2332649217705145. 3

A. Alizadegan, B. Asady, and M. Ahmadpour. Two modified versions of artificial bee colony algorithm. *Applied Mathematics and Computation*, 225:601–609, 12 2013. ISSN 0096-3003. doi: 10.1016/J.AMC.2013.09.012. 14

S. I. Amoukou, N. J.-B. Brunel, and T. Salaün. The Shapley Value of coalition of variables provides better explanations. 3 2021. URL `https://arxiv.org/abs/2103.13342v3`. 15

K. J. Archer and R. V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 1 2008. ISSN 0167-9473. doi: 10.1016/J.CSDA.2007.08.015. 15

C. Blumenberg, D. Zugna, M. Popovic, C. Pizzi, A. J. D. Barros, and L. Richiardi. Questionnaire Breakoff and Item Nonresponse in Web-Based Questionnaires: Multilevel Analysis of Person-Level and Item Design Factors in a Birth Cohort. *Journal of Medical Internet Research*, 20(12), 2018. doi: 10.2196/11046. 3

M. D. Bormida. *The Big Data World: Benefits, Threats and Ethical Challenges*. Emerald Publishing Limited, 12 2021. ISBN 978-1-80262-414-4. doi: 10.1108/S2398-601820210000008007. 1

L. Breiman. Bagging Predictors. 24:123–140, 1996. 8

L. Breiman. Pasting small votes for classification in large databases and on-line. *Machine Learning*, 36(1):85–103, 1999. ISSN 08856125. doi: 10.1023/A:1007563306331/METRICS. URL `https://link.springer.com/article/10.1023/A:1007563306331`. 8

J. Brownlee. How to use Data Scaling Improve Deep Learning Model Stability and Performance, 2020. 11

O. Bulut, J. Xiao, M. C. Rodriguez, and G. Gorgun. An Empirical Investigation of Factors Contributing to Item Nonresponse in Self-Reported Bullying Instruments. *Journal of School Violence*, 19(4):539–552, 2020. doi: 10.1080/15388220.2020.1770603. 3

D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):1–13, 1 2020.

ISSN 14712164. doi: 10.1186/S12864-019-6413-7/TABLES/5. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7. 9

D. Chicco and G. Jurman. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(1), 12 2023. ISSN 17560381. doi: 10.1186/s13040-023-00322-4. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9938573#id-name=PMChttp://doi.org/10.1186/s13040-023-00322-4. 9

L. M. Collins, J. L. Schafer, and C. M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(3):330–351, 2001. ISSN 1082989X. doi: 10.1037/1082-989X.6.4.330. 4, 6

Y. Dong and C. Y. J. Peng. Principled missing data methods for researchers. *SpringerPlus*, 2(1): 1–17, 2013. ISSN 21931801. doi: 10.1186/2193-1801-2-222. URL /pmc/articles/PMC3701793//pmc/articles/PMC3701793/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/. 5

M. N. Elliott, C. Edwards, J. Angeles, K. Hambarsoomians, and R. D. Hays. Patterns of Unit and Item Nonresponse in the CAHPS® Hospital Survey. *Health Services Research*, 40(6 Pt 2): 2096, 12 2005. ISSN 00179124. doi: 10.1111/J.1475-6773.2005.00476.X. URL /pmc/articles/PMC1361246//pmc/articles/PMC1361246/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1361246/. 3

C. K. Enders. *Applied missing data analysis*. Guilford Press, 2 edition, 10 2022. ISBN 9781462549863. URL https://www.routledge.com/Applied-Missing-Data-Analysis/Enders/p/book/9781462549863. 4

V. A. Epanechnikov. Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability and Its Applications*, 14(1):153–158, 7 1969. ISSN 0040-585X. doi: 10.1137/1114019. URL https://epubs.siam.org/doi/10.1137/1114019. 21

ESS Round 8: European Social Survey Round 8. Data file edition 2.2, 2016. 26

European Social Survey. About ESS, 2023a. URL https://www.europeansocialsurvey.org/about/. 26

European Social Survey. Data Collection, 2023b. URL https://www.europeansocialsurvey.org/methodology/ess_methodology/data_collection.html. 26

European Social Survey. European Social Survey Round 8, 2023c. URL https://ess-search.nsd.no/en/study/f8e11f55-0c14-4ab3-abde-96d3f14d3c76. 26

European Social Survey (ESS). Interviewer Behaviour and Effects, 2023a. URL https://www.europeansocialsurvey.org/methodology/methodological_research/interviewer_behaviour_effects.html. 41

European Social Survey (ESS). Translation Procedures, 2023b. URL https://www.europeansocialsurvey.org/methodology/methodological_research/translation_procedures.html. 41

J. J. Faraway. *Practical Regression and Anova using R*. Lawrence Erlbaum Associates, Mahwah, N.J., 2 edition, 2002. ISBN 978-0-8058-4037-7. URL https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf. 13

E. Fitkov-Norris, S. Vahid, and C. Hand. Evaluating the Impact of Categorical Data Encoding and Scaling on Neural Network Classification Performance: The Case of Repeat Consumption of Identical Cultural Goods. *Communications in Computer and Information Science*, 311:343–0352, 2012. ISSN 18650929. doi: 10.1007/978-3-642-32909-8{_}35. URL https://www.researchgate.net/publication/262173733_Evaluating_the_Impact_of_Categorical_Data_Encoding_and_Scaling_on_Neural_Network_Classification_Performance_The_Case_of_Repeat_Consumption_of_Identical_Cultural_Goods. 10, 11

J. Gerards, M. Schaefer, A. Vankan, and I. Muis. Fundamental Rights and Algorithms Impact Assessment. Technical Report March, Data School,Ministry of Interior and Kingdom Relations, Utrecht, 2022. URL https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms. 40

L. Göbbels. GitHub Repository Applied Data Science Thesis "Using Multiverse Analysis for Estimating Response Models", 2023. URL https://github.com/Lieve2/ADSthesis_multiverse_analysis.git. 24

S. B. Goldberg, D. M. Bolt, and R. J. Davidson. Data Missing Not at Random in Mobile Health Research: Assessment of the Problem and a Case for Sensitivity Analyses. *Journal of Medical Internet Research*, 23(6), 6 2021. ISSN 14388871. doi: 10.2196/26749. URL /pmc/articles/PMC8277392//pmc/articles/PMC8277392/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8277392/. 4

J. W. Graham. *Missing data: Analysis and design*. Springer New York, 1 2012. ISBN 9781461440185. doi: 10.1007/978-1-4614-4018-5/COVER. 1, 3, 4

T. Gupta. Continuous Data and Zero-Frequency Problem in Naive Bayes Classifier, 2020. URL https://towardsdatascience.com/continuous-data-and-zero-frequency-problem-in-naive-bayes-classifier-7784f4066b51. 6

G. Hooker, L. Mentch, and S. Zhou. Unrestricted Permutation forces Extrapolation: Variable Importance Requires at least One More Model, or There Is No Free Variable Importance. *Statistics and Computing*, 31(6), 5 2019. ISSN 15731375. doi: 10.1007/s11222-021-10057-z. URL https://arxiv.org/abs/1905.03151v2. 15

Inside BigData. Data Science is Changing the World for the Better: Here's How , 4 2020. URL https://insidebigdata.com/2020/04/14/data-science-is-changing-the-world-for-the-better-heres-how/. 1

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer New York, NY, 1 edition, 2013. ISBN 978-1-4614-7138-7. doi: 10.1007/978-1-4614-7138-7{_}1. URL https://link.springer.com/book/10.1007/978-1-4614-7138-7#bibliographic-information. 3, 4, 13

D. Karaboga and B. Akay. A comparative study of Artificial Bee Colony algorithm. *Applied Mathematics and Computation*, 214(1):108–132, 2009. doi: 10.1016/j.amc.2009.03.090. 14

D. Karaboga and B. Basturk. Artificial Bee Colony (ABC) optimization algorithm for solving constrained optimization problems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4529 LNAI:789–798, 2007. ISSN 16113349. doi: 10.1007/978-3-540-72950-1{_}77/COVER. URL `https://link.springer.com/chapter/10.1007/978-3-540-72950-1_77`. 14

C. Kern, B. Weiss, and J.-P. Kolb. Predicting Nonresponse in Future Waves of A Probability-Based Mixed-Mode Panel With Machine Learning. *Journal of Survey Statistics and Methodology*, 11 (1):100–123, 2023. doi: 10.1093/JSSAM/SMAB009. 3

P. Kutschar and M. Weichbold. Interviewing elderly in nursing homes: Respondent and survey characteristics as predictors of item nonresponse. *Survey Methods: Insights from the Field*, 2019. doi: 10.13094/SMIF-2019-00015. 3

T. Lanigan, S. Raschka, and A. Amor. Importance of Feature Scaling, 2023. 10

S. Lee, M. Liu, and M. Hu. Relationship Between Future Time Orientation and Item Nonresponse on Subjective Probability Questions: A Cross-Cultural Analysis. *Journal of Cross-Cultural Psychology*, 48(5), 2017. doi: 10.1177/0022022117698572. 3

J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-Free Predictive Inference For Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 4 2016. ISSN 1537274X. doi: 10.1080/01621459.2017.1307116. URL `https://arxiv.org/abs/1604.04173v2`. 15

G. Li, P. Niu, and X. Xiao. Development and investigation of efficient artificial bee colony algorithm for numerical function optimization. *Applied Soft Computing*, 12(1):320–332, 1 2012. ISSN 1568-4946. doi: 10.1016/J.ASOC.2011.08.040. 14

O. Lipps and G.-A. Monsch. Effects of Question Characteristics on Item Nonresponse in Telephone and Web Survey Modes. *Field Methods*, 34(4), 2022. doi: 10.1177/1525822X221115838. 3

Z. C. Lipton. The Mythos of Model Interpretability. *Communications of the ACM*, 61(10):35–43, 6 2016. ISSN 15577317. doi: 10.1145/3233231. URL `https://arxiv.org/abs/1606.03490v3`. 14

R. J. A. Little. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, 83(404):1198, 12 1988. ISSN 01621459. doi: 10.2307/2290157. 4

S. Lundberg. Simple Kernel SHAP, 2018. URL `https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/model_agnostic/SimpleKernelSHAP.html?highlight=approximation`. 16

S. M. Lundberg and S. I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017-December:4766–4775, 5 2017. ISSN 10495258. URL `https://arxiv.org/abs/1705.07874v2`. 12, 15, 16

A. Mantelero. *Beyond data: Human Rights, Ethical and Social Impact Assessment in AI*. Springer, 2022. ISBN 9789462655300. 40

S. Masís. Interpretable Machine Learning with Python. *Packt Publishing*, page 737, 2021. URL https://www.oreilly.com/library/view/interpretable-machine-learning/9781800203907/. 12

L. Mentch and G. Hooker. Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research*, 17:1–41, 2016. 15

D. Micci-Barreca. A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems. *ACM Special Interest Group on Knowledge Discovery in Data Explorations Newletter*, 3(1):27–32, 2001. doi: https://doi.org/10.1145/507533.507538. URL https://doi-org.proxy.library.uu.nl/10.1145/507533.507538. 11, 12

G. Mignogna, C. E. Carey, R. Wedow, N. Baya, M. Cordioli, N. Pirastu, R. Bellocco, M. G. Nivard, B. M. Neale, R. K. Walters, and E. Ingelsson. Patterns of item nonresponse behavior to survey questionnaires are systematic and have a genetic basis, 2022. 3

B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms: Mapping the debate. *Big Data and Society*, 3(2):1–21, 2016. ISSN 20539517. doi: 10.1177/2053951716679679. 40

J. Mockus. *Bayesian Approach to Global Optimization*. Springer Netherlands, Dordrecht, 1 edition, 1989. ISBN 978-94-009-0909-0. URL https://link.springer.com/book/10.1007/978-94-009-0909-0. 14

T. Montvilas. How To Approach Data Governance To Avoid Poor Data Quality, 2022. URL https://www.forbes.com/sites/forbesbusinessdevelopmentcouncil/2022/01/07/how-to-approach-data-governance-to-avoid-poor-data-quality/?sh=5113712411a2. 1

T. Munzner. *Visualization Analysis and Design*. Taylor & Francis Group, 2014. ISBN 978-1-4665-0891-0. 24, 52, 54

J. L. Myers and A. D. Well. *Research Design and Statistical Analysis*. 2003. 13

K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(1):1–13, 2 2010. ISSN 14712105. doi: 10.1186/1471-2105-11-110/FIGURES/6. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-110. 15

J. Peng, J. Hahn, and K.-W. Huang. Handling Missing Values in Information Systems Research: A Review of Methods and Assumptions. *Information Systems Research*, 34(1), 3 2022. ISSN 1047-7047. doi: 10.1287/ISRE.2022.1104. URL http://pubsonline.informs.orghttp://www.informs.org:Theonlineappendicesareavailableathttps://doi.org/10.1287/isre.2022.1104. 1, 3

V. Plotnikova, M. Dumas, and F. Milani. Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science*, 6:1–43, 2020. ISSN 23765992. doi: 10.7717/PEER J-CS.267/SUPP-2. URL /pmc/articles/PMC7924527//pmc/articles/PMC7924527/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7924527/. 17

J. Raymaekers, P. J. Rousseeuw, and M. Hubert. Class maps for visualizing classification results. *Technometrics*, 64(2):151–165, 7 2020. doi: 10.1080/00401706.2021.1927849. URL http://arxiv.org/abs/2007.14495http://dx.doi.org/10.1080/00401706.2021.1927849. 16

P. Refaeilzadeh, L. Tang, and H. Liu. Cross-Validation. *Encyclopedia of Database Systems*, pages 532–538, 2009. doi: 10.1007/978-0-387-39940-9{_}565. URL https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_565. 20

D. B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/BIOMET/63.3.581. URL https://dash.harvard.edu/handle/1/3408223. 3

Scikit-Learn Developers. Permutation Importance with Multicollinear or Correlated Features, 2023a. URL https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html. 13, 15

Scikit-Learn Developers. sklearn.ensemble.BaggingClassifier, 2023b. URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html. 8

T. Shin. The 10 Best Data Visualizations of 2021, 10 2021. URL https://towardsdatascience.com/the-10-best-data-visualizations-of-2021-fec4c5cf6cdb. 53

D. Singh and B. Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, 12 2020. ISSN 1568-4946. doi: 10.1016/J.ASOC.2019.105524. 10, 11

N. Squires. Italian parliament gives gay unions the green light, 5 2016. URL https://www.telegraph.co.uk/news/2016/05/11/italian-parliament-gives-gay-marriages-the-green-light/. 27

C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):1–21, 1 2007. ISSN 14712105. doi: 10.1186/1471-2105-8-25/FIGURES/11. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-25. 15

L. Tani and C. Veelken. Comparison of Bayesian and particle swarm algorithms for hyperparameter optimisation in machine learning applications in high energy physics, 1 2022. URL https://arxiv.org/abs/2201.06809v1. 14

A. Tharwat. Classification Assessment Methods. *Applied Computing and Informatics*, 17(1): 168–192, 2021. doi: 10.1016/jaci2018.08.003. 8

K. M. Ting. *Encyclopedia of machine learning*. Springer, 2011. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8. 8

M. Turek. Explainable Artificial Intelligence (XAI), 2023. URL `https://www.darpa.mil/program/explainable-artificial-intelligence`. 15, 16

S. Van Buuren. *Flexible imputation of missing data*. Routledge, 2 edition, 2018. ISBN 9780429492259. URL `https://stefvanbuuren.name/fimd/`. 1, 3, 4, 5

D. Vorotyntsev. Stop Permuting Features, 7 2020. URL `https://towardsdatascience.com/stop-permuting-features-c1412e31b63f`. 15

C. Wang, P. Shang, and P. Shen. An improved artificial bee colony algorithm based on Bayesian estimation. *Complex and Intelligent Systems*, 8(6):4971–4991, 12 2022. ISSN 21986053. doi: 10.1007/S40747-022-00746-1/TABLES/16. URL `https://link.springer.com/article/10.1007/s40747-022-00746-1`. 14

C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, San Francisco, 2 edition, 2014. ISBN 1558608192. URL `https://www.researchgate.net/publication/224285723_Information_Visualization_Perception_for_Design_Second_Edition`. 24

M. Weinhardt. Ethical Issues in the Use of Big Data for Social Research. *Historical Social Research*, 45(3):342–368, 2020. doi: 10.12759/hsr.45.2020.3.342-368. URL `https://www.researchgate.net/publication/342851817_Ethical_Issues_in_the_Use_of_Big_Data_for_Social_Research`. 1

Wikipedia. Evaluation of Binary Classifiers, 2023. URL `https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers`. 9

I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Inc., 11 2016. ISBN 9780128042915. doi: 10.1016/c2009-0-19715-5. 6, 7, 8, 9, 10, 12, 13, 14

C. Zhang and Y. Ma. *Ensemble Machine Learning: Methods and Applications*. Springer New York, NY, 1 edition, 2012. ISBN 978-1-4419-9326-7. doi: https://doi.org/10.1007/978-1-4419-9326-7. URL `https://link.springer.com/book/10.1007/978-1-4419-9326-7#bibliographic-information`. 8

Z. H. Zhou. *Ensemble methods: Foundations and algorithms*. CRC Press, 1 edition, 6 2012. ISBN 9781439830055. doi: 10.1201/B12207/ENSEMBLE-METHODS-ZHI-HUA-ZHOU. URL `https://www.taylorfrancis.com/books/mono/10.1201/b12207/ensemble-methods-zhi-hua-zhou`. 8

M. Zhu, A. Hahn, Y. Q. Wen, and A. Bolles. Comparison and optimization of the parameter identification technique for estimating ship response models. In *2017 3rd IEEE International Conference on Control Science and Systems Engineering, ICCSSE 2017*, pages 743–750. Institute of Electrical and Electronics Engineers Inc., 10 2017. ISBN 9781538604847. doi: 10.1109/CCSSE.2017.8088033. 14

Appendix A

Dashboard sketch

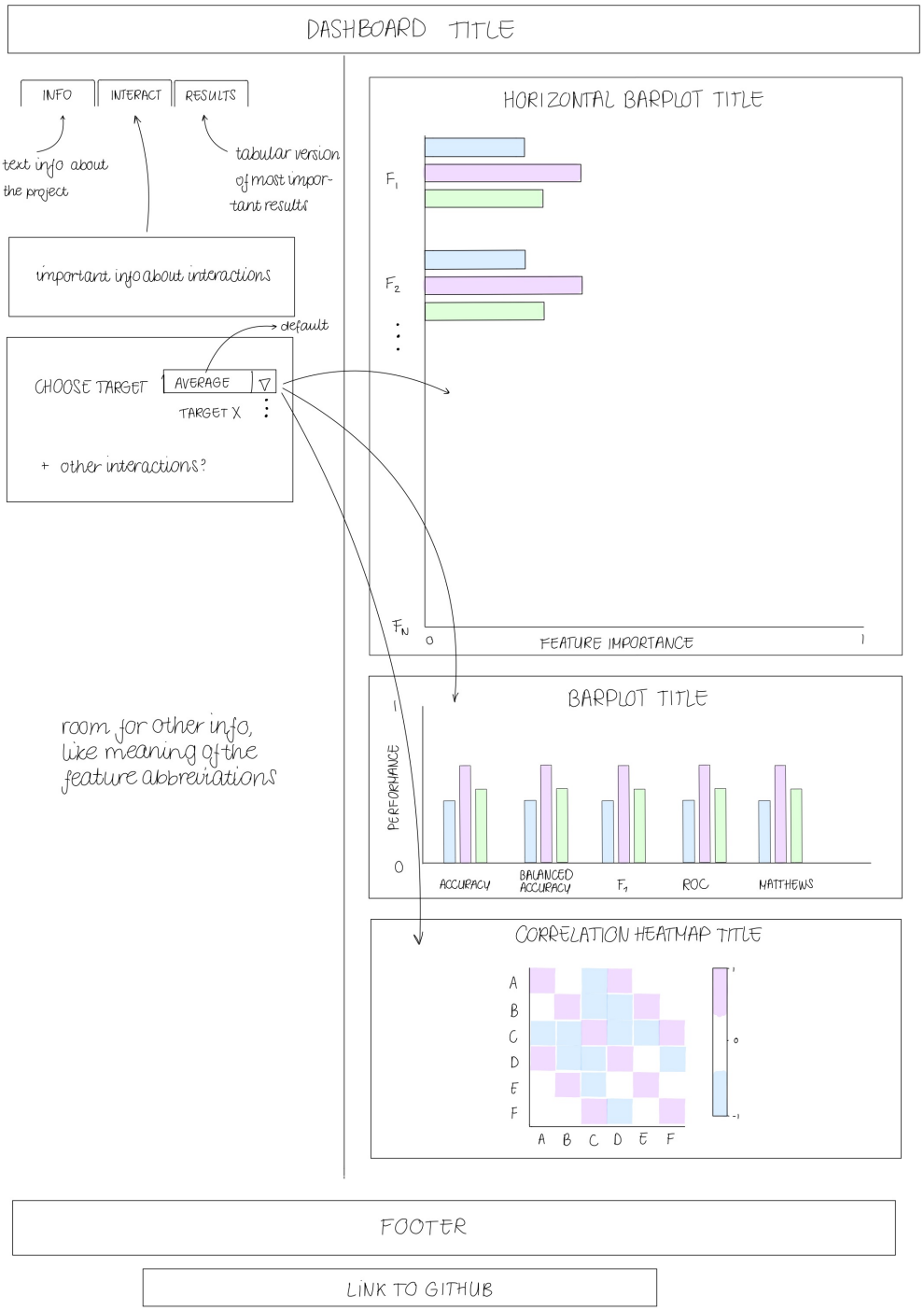

Figure A.1: Final sketch of the dashboard, made in the pre-visualization phase

Appendix B

Supplementary information regarding the dashboard

B.1 Best practices of visualization

Visualization is the art of making the unseen visible, as numbers in itself do not usually tell the whole story and require interpretation. As such, Munzner describes the concept of visualization as follows: "[c]omputer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively" (Munzner, 2014). Then, data visualization is not just making 'pretty pictures', it is about making insightful pictures. Consequently, there are different considerations that need to be made regarding trade-offs and applicability of tools in visualization design and validation of designs, and the limitations of computers, humans and displays (Munzner, 2014). Munzner distinguishes three main goals of visualization: exploration (everything is unknown), verification (hypotheses are present), and communication (everything is known). These goals influence the suitability of particular visualization practices and should therefore be clear when entering the process of visualization design (Munzner, 2014).

To provide some guidance on designing valid visual encodings, Munzner (2014) describes eight rules of thumb. The first rule is to use three-dimensional visualizations as sparingly as possible due to humans perceiving the world more closely to two-dimensional instead of three-dimensional, increasing the risk of severe misperception caused by line-of-sight ambiguity and perspective distortion. Similarly, the second rule is to only use two-dimensional visualizations when justified, so when understanding the topological aspect is essential. The third rule is to minimize cognitive load as human perception easily suffers from change blindness – the inability to notice (drastic) changes if one's attention is directed elsewhere. The fourth rule regards the trade-off between resolution, defined as the number of available pixels divided by the display area, and immersion, which refers to the feeling of presence. That is, immersion comes at the cost of resolution. The next rule regards yet another trade-off: showing an overview versus providing details. Here, it is argued that one should first provide an overview and add zooming, filtering and details on demand. Related to this is the rule stating that successful interaction design requires a good match between latencies (response time) of the low-interaction mechanism, the visual feedback mechanism, the system update time and the operational load, which will lead to the possibility of exploring a larger information space compared to a static image. In other words, here one should think about desirable maximum durations an interactive process can take so that human attention is maintained and the visualization continues to be insightful. The last two rules regard the more formal aspects: color and layout. According to Munzner, even in black-and-white the

most crucial aspects must be legible. Moreover, it is argued that one should focus on function first before beautifying the visualization, as effectiveness is more crucial than attractiveness. Here, a ranking of different channels depending on the type of attribute being visualized is provided. For ordered attributes, position on a common scale, followed by one one an unaligned scale and length respectively, are most effective. For categorical attributes, these are spatial region, followed by color hue and motion.

Lastly, related to Munzner's 'black-and-white' rule and the advice to favor effectiveness over attractiveness, it is vital to determine a suitable color palette by taking into account color theory and limitations in vision in the case of colorblindness. Even though red-green colorblindness is most common, there are several other types of color blindness that cause different colors to be rendered indistinguishable. To guide choosing the right color palette, simulation software has been developed, to visualize how the chosen colors are perceived by the respective types of colorblindness. One such simulator is the *jsColorblindSimulator*, which allows simulating protanopia (red blindness), protanomaly (red weakness), deuteranopia (green blindness), deuteranomaly (green weakness), tritanopia (blue blindness), tritanomaly (blue weakness), achromatopsia (complete color blindness), and achromatomaly (complete color weakness) . As a general rule, it is advised to not combine contrasting and closely related colors, like red and green, or grey and blue, but finding suitable color combinations usually remains an act of trial and error.

Keeping in mind these rules of thumb will lead to clean and clear visualizations. For example, Shin (2021) describes ten excellent visualizations made in 2021, in which Munzner's rules of thumb are embedded even though these are not explicitly mentioned (Shin, 2021) . These visualizations, among other things, do not use color unnecesarily and minimize the data-to-ink ratio, keep the information density in the visualizations low, and only use the three-dimensional space when justified, for example for visualizing the earth's submarine cable network.

B.2 From sketch to fully-functioning dashboard

After the pre-visualization process and after gathering all of the desired results, as explained in more detail in Chapters 3 and 5, the envisioned dashboard has been coded to a deployable, interactive visualization tool.[1] In the final dashboard, effort is put into providing essential information in an exhaustive and transparent way, without compromising clarity, for a wide audience, ranging from beginning data science students to established researchers. In the following sections, the different aspects of the dashboard will be explained in more detail to give insight in the different decisions made and the reasoning behind these decisions.

B.2.1 Arranging tabular data

In this section, the different plots and graphs used in the dashboard and the reasoning behind them will be explained. These choices all relate to the different possibilities of arranging tabular data, which is the format in which all of the results are stored. The first two plots shown are variations of each other: a horizontal bar plot and the standard (vertical) bar plot. This type of plots is ideal for the tasks of looking up and comparing values of one quantitative value attribute,

[1]The dashboard can be accessed here: `http://multiverseanalysisvisualizer.pythonanywhere.com/`.

like feature importance and model performance – given that the performance metrics operate on the same scale (Munzner, 2014). For visualizing the feature importance, it is decided to use a horizontal bar plot instead of the standard version used for the model performance, because of the large number of features (keys) present. By transposing the bar plot and adding an option to vertically scroll through the barplot, clarity and legibility are maximized and comparison between the feature importances of the different models and between different features is possible.

Immediately below the bar plots visualizing feature importance and model performance, a visualization of the six different class maps is presented. These are scatter plots that show the uncertainty of the model for each observation in relation to the difficulty of the observation. Scatter plots are most suitable here as the goal is to find some form of trends – that is, for instance, if the model is uncertain about its predictions of difficult observations or if it is wrong, but certain, about simple observations – regarding two quantitative variables, namely the model's probability of an observation belonging to an alternative class (model uncertainty) and localized farness (difficulty of the observation) (Munzner, 2014). There is one scatter plot for each prediction class, so two class maps per model as the task is binary classification. To allow for easier comprehension, in each plot two helper lines are added to indicate the 99% quantile distance from the class (i.e. extremely difficult observations) and the midpoint of class uncertainty (i.e. the area where the model is most uncertain).

The last graph present in the dashboard is a correlation heat map, to allow assessing the existence of multicollinearity for the different data sets used during the processing and of the original (cleaned) data. This is beneficial for visualizing the large quantity of data in a compact space, providing a clear overview despite the relatively high information density due to the large number of features (Munzner, 2014). This type of graph uses two categorical key attributes (the features) and one quantitative value (the correlation) as input. The features are presented on the two axes while the correlation is visualized by color encoding the value with a continuous color scale (Munzner, 2014).

Besides graphs, several tables are added in the *Results* tab, that can be found in the top left of the dashboard. Adding these tables ad a higher degree of exhaustiveness and possibilities for analysis in the dashboard. Moreover, one may be looking for particular values, which can sometimes be easier to find in (sortable) tables, like the ones included. More specifically, there are tables with the feature importance and model performance results for each model, providing a total of six tables.

B.2.2 Interacting with the dashboard

The dashboard does not merely provide static graphs, but allows modifying the views to particular needs, further enhancing a high level of both clarity and detail and mitigating the fallacy illustrated by Anscombe's Quartet: a single summary is often an oversimplification that does not show the true structure of the data (Munzner, 2014). The first option for interaction is a dropdown menu, including search bar, to define a particular target (i.e. model and data set) to the views or choose to view the average over all models. The latter option is also the default setting. This dropdown menu is linked to all four plots to ensure cohesion of the different results. The dropdown format only allows choosing one particular key at a time, which is in this case desirable as otherwise the

legibility and clarity of the bar plots would be compromised, or at least significantly more complex, and visualizations of the class maps and correlation heat map would be rendered impossible.

The horizontal bar plot allows for two more customizations: choosing which model is used to order the bars (from high to low) – this is the random forest by default – and choosing whether to show the subset of the features used for classification for the chosen target or whether to show all of the features, where features belonging to the same cluster are assigned the same importance as explained in Chapter 3. In this case, the so-called *radio-items* interaction is preferred over the dropdown, because even though there is still only one option at a time possible, there are now only three (choosing model) and two (choosing subset versus all) options to choose from, instead of the 75 possibilities when choosing a target. Where a dropdown collapses all options into a single row if not interacting with it, a radio-items button keeps all options visible, which provides more clarity on the different options to choose from.

Next, for the standard bar plot, a multi-select option is added to show or hide one or more performance metrics. By default, all five calculated metrics are shown, but if one for example wants to compare accuracy and balanced accuracy of the models with each other, one can unselect the *ROC*, *F1*, and *Matthews* options.

Besides the graph-specific interactions,there is also a possibility to hover over the plots and view results for that particular point. For the bar plots this means that the exact value for a certain bar is shown. Similarly, for the class maps this mean the exact values uncertainty and farness (difficulty) for a particular observation are shown and for the heat map the feature pair and corresponding correlation value are presented.

B.2.3 Other features in the dashboard

Besides graphs and tables, there is also space dedicated to essential descriptions in a textual format, with the aim to help users to comprehend the visualizations and interact with the dashboard to acquire the views they desire. Moreover, since the features are presented in an abbreviated format, their descriptions are too long to add to the visualizations, and because going through the codebook each time is tedious and may even be impossible if the codebook is unavailable, a description of the meaning of all of the features present in the data used for processing is added. Moreover, this description is also linked to a radio-items interaction. That is, if a target is selected from the dropdown menu, one can choose to either show the feature information or show the cluster information of the chosen target. When the cluster information option is selected, the list of selected targets is shown, together with non-selected features for each of the respective clusters.

Moreover, to organize the descriptions and avoid making the visualization page getting too dense, three different tabs are included: *Info*, *Interact*, and *Results*. The first tab includes general information on the dashboard and the study. The second tab contains the interaction options, including descriptions to guide the user in how to use them, and the feature and cluster information. Lastly, the third tab contains the six tables containing the same results shown in the plots, but in a tabular format.

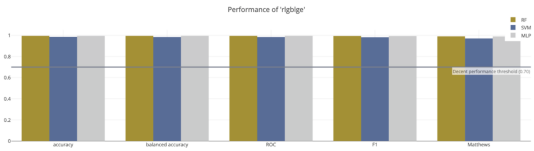

Figure B.1: Simulation of viewing the standard bar plot with protanopia

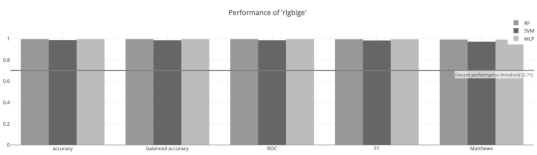

Figure B.2: Simulation of viewing the standard bar plot with achromatopsia

A little bit of color theory

As described in Section B.1, hue (or color) is an important factor in the success or failure of a visualization (dashboard). As such, in this section, the decisions made regarding hues, in combination with other encodings, are described. The background boxes of the visualizations are colored alternately between white and a light-grey to clearly distinguish the different spaces dedicated to each of the graphs and avoid accidental merging of the results when interpreting the graphs. The color theme present in the graphs themselves is based on a trial-and-error process of choosing colors that are distinguishable for all of the different types of color blindness while remaining as aesthetically pleasing as possible. As indicated in Section B.1, the colorblindness assessment has been done using the jsColorblindSimulator. Figures B.1 and B.2 show what the chosen color theme looks like for the standard bar plot for protanopia – the most common form of color blindness – and achromatopsia.

Appendix C

Dashboard interaction view

Figure C.1: Image of the dashboard with the interaction tab open

Appendix D

Code for the class maps

Listing D.1: Calculating uncertainty and localized farness for class maps

```
def compPAC(model, X, y):
    """
    :param model: sklearn model fitted to training data.
                  model must have "probability=True" when initialized.
    :param X:     dataset for prediction, usually a held-out test set
    :param y:     labels corresponding to X; must be numpy array
    :return:      Probability of Alternative Classification (PAC) from the trained
        classifier
    """

    # parameters
    n = X.shape[0] # number of data points in X
    PAC = np.array([0.0]*n) # initialize PAC array
    nlab = len(np.unique(y)) # number of classes

    # get fitted model probabilities
    model_probs = model.predict_proba(X)

    # case: two classes
    if nlab == 2:
        altint = 1 - y # y will take values 0 or 1
        for i in range(n):
            PAC[i] = model_probs[i, altint[i]]
        return PAC

def compLocalFarness(X, y, k, metric='euclidean'):
    """
    :param X:       dataset for prediction, should be the same as what was used for
        PAC
    :param y:       corresponding labels of X
    :param k:       number of nearest neighbors to consider for localized farness
        computation
    :param metric:  distance metric for nearest neighbor search.
    :return:        localized farness computed from the data, independent of
        classifier
    """

    # find nearest neighbors with KD Tree
    kdt = KDTree(X, metric=metric)
    dist, ind = kdt.query(X, k=k) # get the nearest neighbor distances and indices
```

```
    # array of epsilon_i (widths of epanechnikov kernels)
    epislon_arr = [dist[i][(k-1)] for i in range(len(dist))]

    # epanechnikov kernel weighting function
    ep_kernel = lambda x: (3/4)*(1 - (x*x))*(int(abs(x) <= 1))
    kernel_wt = lambda ep, d: (1/ep) * ep_kernel(x=(d/ep))

    # compute localized farness
    n = X.shape[0] # number of rows in the data
    local_farness = np.array([0.0]*n) # initialize local farness

    for i in range(n):
        local_dists = dist[i] # distances from point i to its neighbors
        wts = [kernel_wt(ep=epislon_arr[i], d=local_dists[ii]) for ii in range(len(
            local_dists))]
        wts = wts / sum(wts) # weight the local distances. wts should sum to 1.
        class_prob = sum(wts[y[ind[i]] == y[i]]) # Pr(i \in g_i)
        local_farness[i] = 1.0 - class_prob # LF(i) = 1 - Pr(i \in g_i)

    # round to 4 decimal places, for simplicity
    local_farness = np.abs(np.round(local_farness, 4))
    return local_farness
```

Appendix E

Overview results and qualitative assessment

Table E.1: Overview results and qualitative assessment

Target	Performance (Matthews)	Important features
basinc	RF: 0.430[1] SVM: 0.149 MLP: 0.0	RF: gndr , *gndr2*[2], *agea* (0.080), pspwght, *ipcrtiv*, *pweight* (0.055) SVM: – MLP: –
bennent	RF: 0.221 SVM: 0.069 MLP: 0.0	RF: – SVM: – MLP: –
bnlwinc	RF: 0.170 SVM: 0.0 MLP: 0.0	RF: – SVM: – MLP: –
ccgdbd	RF: 0.721 SVM: 0.669 MLP: 0.704	RF: – SVM: – MLP: banhhap, *dfincac*, *uemplwk*, *smdfslv* (0.053)
ccrdprs	RF: 0.718 SVM: 0.659 MLP: 0.676	RF: – SVM: – MLP: –
chldhhe	RF: 0.907 SVM: 0.749 MLP: 0.860	RF: yrbrn2, *icpart2*, *rshipa2*, *pdwrk*, *lvgptnea*, *dvrcdeva*, *rshpsts*, *rshipa3* (0.397), iccohbt, *domicil*, *edulvlpb*, *uemplap*, *edctnp*, *eisced*, *isco08p*, *edulvlb*, *eiscedf*, *dngdkp*, *wkhtotp*, *eduyrs*, *iccohbt*, *emprelp*, *pdwrkp*, *marsts* (0.105), netusoft, *anctry1*, *hswrkp*, *hincfel*, *hinctnta*, *actrolga*, *dngref* (0.056) SVM: yrbrn2 (0.335), netusoft (0.051) MLP: yrbrn2 (0.353), iccohbt (0.066), netusoft (0.064)

[1]Features are considered important when SHAP$\geq$ 0.05

[2]Italicized features refer to features not directly used in the the analysis but present in the same cluster as the important feature during feature selection, hence having the same importance. For each target, these 'clustered' features are only indicated once per target, but the same clusters hold for each model as the subset selection is done before the modeling phase is started

Target	Performance (Matthews)	Important features
crpdwrk	RF: 0.834 SVM: 0.665 MLP: 0.804	RF: gndr3, *wrkctra*, *estsz*, *jbspv*, *dsbld*, *emplrel* (0.142), netusoft, *icpart1*, *hinctnta*, *hincfel*, *hswrkp*, *anctry1*, *dngref*, *actrolga*, *rshpsts* (0.095), dngna, *pdjobyr*, *yrbrn3* (0.092), yrbrn3, *rshipa2*, *dngnapp* (0.069) SVM: netusoft (0.211), netustm, *uempli*, *uempla* (0.052), gndr3 (0.081) MLP: netusoft (0.170), gndr3 (0.125), wkdcorga, *mbtru*, *wkhtot*, *tporgwk*, *iorgact*, *wkhct*, *icwhct*, *nacer2*, *uemp3m*, *isco08*, *wrkac6m* (0.066)
edulvlfb	RF: 1.0 SVM: 0.993 MLP: 1.0	RF: emprf14, *eiscedm*, *occf14b* (0.090) SVM: emprf14 (0.202) MLP: emprf14 (0.126)
edulvlmb	RF: 0.506 SVM: 0.361 MLP: 0.432	RF: – SVM: – MLP: edulvlfb, *emprm14*, *emprf14* (0.058)
edulvlpb	RF: 0.827 SVM: 0.681 MLP: 0.803	RF: uemplap, *eiscedf*, *wkhtotp* (0.197), maritalb, *chldhm*, *uemplip* (0.107), yrbrn2, *pdwrkp*, *rshpsts*, *dngdkp*, *isco08p*, *emrelp*, *edulvlb*, *chldhhe*, *marsts*, *domicil*, *edctnp*, *iccohbt*, *eisced*, *dvrcdeva*, *rshipa2* (0.068) SVM: yrbrn2 (0.130), uemplap (0.112), cntry, *dweight* (0.076), nwspol, *edctn*, *occm14b*, *atncrse* (0.058) MLP: uemplap (0.220), yrbrn2 (0.148)
eduunmp	RF: 0.286 SVM: 0.154 MLP: 0.0	RF: – SVM: – MLP: –
eiscedf	RF: 1.0 SVM: 0.993 MLP: 1.0	RF: emprf14, *eiscedm*, *occf14b* (0.089) SVM: emprf14 (0.202) MLP: emprf14 (0.126)
eiscedm	RF: 0.505 SVM: 0.361 MLP: 0.432	RF: – SVM: – MLP: edulvlfb, *emprm14*, *emprf14* (0.058)
eiscedp	RF: 0.826 SVM: 0.681 MLP: 0.803	RF: uemplap, *wkhtotp*, *eiscedf* (0.204), maritalb, *chldhm*, *uemplip* (0.099), yrbrn2, *marsts*, *eisced*, *emprelp*, *pdwrkp*, *rshpsts*, *isco08p*, *iccohbt*, *edulvlb*, *dngdkp*, *dvrcdeva*, *rshipa2*, *edctnp*, *domicil*, *chldhhe* (0.067) SVM: yrbrn2 (0.130), uemplap (0.112), cntry, *dweight* (0.076), nwspol, *edctn*, *occm14b*, *atncrse* (0.058) MLP: uemplap (0.220), yrbrn2 (0.148)

Target	Performance (Matthews)	Important features
elgbio	RF: 0.583 SVM: 0.568 MLP: 0.0	RF: – SVM: wrdpimp (0.055) MLP: –
elgcoal	RF: 0.709 SVM: 0.660 MLP: 0.674	RF: – SVM: – MLP: wrpwrct, *wrdpimp*, *wrinspw*, *wrntdis* (0.055)
elghydr	RF: 0.713 SVM: 0.693 MLP: 0.702	RF: – SVM: – MLP: –
elgngas	RF: 0.698 SVM: 0.655 MLP: 0.669	RF: – SVM: wrpwrct, *wrdpimp*, *wrinspw* (0.076) MLP: –
elgnuc	RF: 0.651 SVM: 0.621 MLP: 0.652	RF: – SVM: wrdpimp, *wrntdis*, *wrdpfos* (0.057) MLP: –
emplrel	RF: 0.799 SVM: 0.598 MLP: 0.728	RF: – SVM: iorgact, *wrkac6m*, *mbtru*, *nacer2*, *uemp3m*, *icwhct*, *wkhtot*, *isco08*, *wkhct*, *tporgwk* (0.059) MLP: iorgact (0.130)
emprelp	RF: 0.757 SVM: 0.633 MLP: 0.750	RF: dvrcdeva, *edctnp*, *uemplap*, *iccohbt*, *dngdkp*, *pdwrkp*, *isco08p*, *wkhtotp*, *eisced*, *marsts*, *eiscedf*, *edulvlb*, *dvrcdeva*, *domicil*, *chldhhe*, *edulvlpb* (0.183), maritalb, *chldhm* (0.083), netusoft, *hswrkp*, *hincfel*, *dngref* (0.076), lvgptnea, *eduyrs*, *pdwrk*, *icpart2* (0.067) SVM: dvrcdeva (0.142), netusoft (0.115), netustm, *uempli*, *uempla* (0.078), lvgptnea, *icpart2*, *eduyrs*, *pdwrk* (0.062) MLP: dvrcdeva (0.256), netusoft (0.128)
estsz	RF: 0.832 SVM: 0.811 MLP: 0.831	RF: – SVM: iorgact, *wrkac6m*, *wkhtot*, *nacer2*, *icwhct* (0.082) MLP: iorgact (0.092)
eudcnbf	RF: 0.807 SVM: 0.786 MLP: 0.800	RF: psppsgva, *regunit*, *rfgbfml*, *region*, *ipcrtiv* (0.658), wrkprty, *pray*, *imwbcnt*, *wrkorg* (0.070) SVM: psppsgva (0.184) MLP: psppsgva (0.280)
euftf	RF: 0.394 SVM: 0.175 MLP: 0.0	RF: – SVM: – MLP: –
eusclbf	RF: 0.847 SVM: 0.825 MLP: 0.844	RF: psppsgva, *ipcrtiv*, *rfgbfml*, *region*, *regunit* (0.710) SVM: psppsgva (0.189) MLP: psppsgva (0.305)

Target	**Performance (Matthews)**	**Important features**
gndr2	RF: 0.996 SVM: 0.971 MLP: 0.996	RF: yrbrn3, *marsts*, *edulvlb*, *wkhtotp*, *emprelp*, *isco08p*, *dngdkp*, *edctnp*, *edulvlpb*, *domicil*, *chldhhe*, *eiscedf*, *dvrcdeva*, *eisced*, *iccohbt*, *uemplap*, *rshpsts*, *pdwrkp* (0.131), rshipa3, *lvgptnea*, *eduyrs*, *pdwrk*, *icpart2* (0.064) SVM: yrbrn3 (0.284) MLP: yrbrn3 (0.284)
gndr3	RF: 0.993 SVM: 0.973 MLP: 0.992	RF: yrbrn3, *pdwrk*, *eduyrs*, *lvgptnea*, *icpart2* (0.502) SVM: yrbrn3 (0.446) MLP: yrbrn3 (0.491)
gvsrdcc	RF: 0.696 SVM: 0.660 MLP: 0.685	RF: – SVM: – MLP: inctxff, *dfincac*, *uemplwk*, *banhhap*, *smdfslv* (0.053)
hinctnta	RF: 0.272 SVM: 0.035 MLP: 0.0	RF: psppsgva, *rfgbfml*, *psppsgva*, *regunit*, *region* (0.097) SVM: – MLP: –
iccohbt	RF: 0.838 SVM: 0.691 MLP: 0.814	RF: uemplap, *wkhtotp*, *eiscedf* (0.209), chldhm, *uemplip*, *chldhhe* (0.113), yrbrn2, *marsts*, *emprelp*, *edulvlpb*, *rshipa2*, *domicil*, *eisced*, *dngdkp*, *pdwrkp*, *edctnp*, *rshpsts*, *isco08p*, *dvrcdeva*, *edulvlb*, *maritalb* (0.068) SVM: yrbrn2 (0.128), uemplap (0.106), cntry, *dweight* (0.076), nwspol, *edctn* (0.055) MLP: uemplap (0.227), yrbrn2 (0.148)
icomdnp	RF: 0.841 SVM: 0.697 MLP: 0.814	RF: uemplap, *wkhtotp*, *eiscedf* (0.220), maritalb, *chldhm*, *uemplip* (0.112), yrbrn2, *rshpsts*, *domicil*, *edulvlpb*, *edctnp*, *dngdkp*, *emprelp*, *edulvlb*, *marsts*, *pdwrkp*, *chldhhe*, *iccohbt*, *rshipa2*, *eisced*, *dvrcdeva* (0.053) SVM: yrbrn2, (0.138), uemplap (0.108), cntry, *dweight* (0.076), nwspol, *edctn* (0.056) MLP: uemplap (0.221), yrbrn2 (0.155)
icppdwk	RF: 0.834 SVM: 0.694 MLP: 0.801	RF: uemplap, *wkhtotp*, *eiscedf* (0.223), maritalb, *chldhm*, *uemplip* (0.123) SVM: yrbrn2, *rshpsts*, *domicil*, *edulvlpb*, *edctnp*, *dngdkp*, *emprelp*, *edulvlb*, *marsts*, *pdwrkp*, *chldhhe*, *iccohbt*, *rshipa2*, *eisced*, *dvrcdeva* (0.131), uemplap (0.107), cntry, *dweight* (0.077), nwspol, *edctn*(0.056) MLP: uemplap (0.201), yrbrn2 (0.165)

Target	Performance (Matthews)	Important features
icwhct	RF: 0.960 SVM: 0.932 MLP: 0.960	RF: wkdcorga, *mbtru*, *tporgwk*, *isco08*, *nacer2*, *uemp3m*, *wkhct*, *wkhtot*, *iorgact*, *wrkac6m* (0.063) SVM: wkdcorga (0.084) MLP: wkdcorga (0.155)
iorgact	RF: 0.914 SVM: 0.890 MLP: 0.914	RF: wkdcorga, *mbtru*, *icwhct*, *wkhct*, *isco08*, *wkhtot*, *tporgwk*, *wrkac6m*, *uemp3m*, *nacer2* (0.051) SVM: wkdcorga (0.083) MLP: wkdcorga (0.143)
isco08	RF: 0.898 SVM: 0.863 MLP: 898	RF: – SVM: wkdcorga, *iorgact*, *icwhct*, *tporgwk*, *wrkac6m*, *mbtru*, *uemp3m* (0.069) MLP: wkdcorga (0.074), wkhct, *wkhtot*, *nacer2* (0.057)
isco08p	RF: 0.746 SVM: 0.629 MLP: 0.732	RF: dvrcdeva, *uemplap*, *dngdkp*, *emprelp*, *wkhtotp*, *eiscedf*, *edulvlb*, *domicil*, *chldhhe*, *edulvlpb*, *edctnp*, *pdwrkp*, *iccohbt*, *marsts*, *eisced* (0.195), netusoft, *hincfel*, *dngref*, *hswrkp* (0.084), lvgptnea, *pdwrk*, *icpart2*, *eduyrs* (0.074), maritalb, *chldhm* (0.073) SVM: dvrcdeva (0.143), netusoft (0.116), netustm, *uempla*, *uempli* (0.072), lvgptnea (0.063) MLP: dvrcdeva (0.252), netusoft (0.123)
jbspv	RF: 0.955 SVM: 0.931 MLP: 0.954	RF: iorgact, *nacer2*, *uemp3m*, *icwhct*, *wkhct*, *tporgwk*, *wrkac6m*, *mbtru*, *wkhtot*, *isco08* (0.062) SVM: iorgact (0.085) MLP: iorgact (0.110)
lbenent	RF: 0.205 SVM: 0.0 MLP: 0.0	RF: wrkprty, *imwbcnt*, *wrkorg* (0.144), psppsgva, *regunit*, *region*, *rfgbfml* (0.051) SVM: – MLP: –
lklmten	RF: 0.654 SVM: 0.595 MLP: 0.647	RF: – SVM: – MLP: lkredcc, *gvsrdcc*, *inctxff*, *ownrdcc*, *sbsrnen* (0.068)
lknemny	RF: 0.721 SVM: 0.718 MLP: 0.718	RF: psppsgva, *regunit*, *region*, *rfgbfml* (0.058) SVM: psppsgva (0.098) MLP: psppsgva (0.072)
lkredcc	RF: 0.712 SVM: 0.676 MLP: 0.683	RF: – SVM: – MLP: inctxff, *smdfslv*, *uemplwk*, *dfincac*, *banhhap* (0.059)
lrscale	RF: 0.345 SVM: 0.129 MLP: 0.326	RF: psppsgva, *rfgbfml*, *regunit*, *region*, *dscrrce* (0.072), wrkprty, *pray*, *imwbcnt*, *wrkorg* (0.069) SVM: – MLP: psppsgva (0.058)

Target	**Performance (Matthews)**	**Important features**
lvgptnea	RF: 0.961 SVM: 0.950 MLP: 0.960	RF: maritalb, *edulvlb*, *eisced*, *domicil* (0.853) SVM: maritalb (0.276) MLP: maritalb (0.162)
marsts	RF: 0.878 SVM: 0.577 MLP: 0.845	RF: chldhm, *chldhhe* (0.582), uemplap, *wkhtotp*, *eiscedf* (0.087) SVM: netusoft, *anctry1*, *actrolga*, *hincfel*, *dngref*, *pdjobyr*, *jbspv*, *hswrkp* (0.104), cntry, *dweight* (0.062), yrbrn2, *iccohbt*, *eisced*, *rshpsts*, *maritalb*, *domicil*, *edulvlpb*, *edctnp*, *dngdkp*, *isco08p*, *emprelp*, *edulvlb*, *pdwrkp*, *dvrcdeva*, *rshipa2* (0.060) MLP: uemplap (0.191), chldhm (0.144), yrbrn2 (0.134), netusoft (0.111)
nacer2	RF: 0.887 SVM: 0.854 MLP: 0.875	RF: – SVM: wkdcorga, *tporgwk*, *icwhct*, *wkhct*, *wkhtot* (0.082) MLP: wkdcorga (0.138)
netustm	RF: 0.978 SVM: 0.957 MLP: 0.977	RF: actrolga, *anctry1*, *hinctnta* (0.261) SVM: actrolga (0.417) MLP: actrolga (0.436)
occf14b	RF: 0.359 SVM: 0.0 MLP: 0.307	RF: – SVM: – MLP: edulvlfb, *emprm14*, *emprf14*, *edulvlmb* (0.069)
occm14b	RF: 0.394 SVM: 0.229 MLP: 0.373	RF: psppsgva, *region*, *atncrse*, *regunit*, *rfgbfml* (0.122), netusoft, *pdwrk*, *rshpsts*, *actrolga*, *icpart1*, *hswrkp* (0.069) SVM: netusoft (0.087) MLP: psppsgva (0.115), netusoft (0.079)
ownrdcc	RF: 0.645 SVM: 0.603 MLP: 0.651	RF: – SVM: – MLP: lkredcc, *gvsrdcc*, *sbsrnen*, *lklmten* (0.062)
pdjobev	RF: 0.817 SVM: 0.658 MLP: 0.799	RF: gndr3, *emplrel*, *dsbld*, *wrkctra*, *estsz*, *jbspv* (0.152), netusoft, *icpart1*, *hinctnta*, *actrolga*, *hincfel*, *hswrkp*, *anctry1*, *rshpsts*, *dngref* (0.105), dngna, *pdjobyr* (0.081), yrbrn3, *rshipa2*, *dngnapp* (0.054) SVM: netusoft (0.215), gndr3 (0.079), netustm, *uempli*, *uempla* (0.052) MLP: netusoft (0.174), gndr3 (0.122), wkdcorga, *uemp3m*, *icwhct*, *wkhtot*, *wrkac6m*, *mbtru*, *nacer2*, *tporgwk*, *wkhct*, *isco08* (0.072)

Target	Performance (Matthews)	Important features
pdjobyr	RF: 0.779 SVM: 0.552 MLP: 0.750	RF: gndr3, *emplrel*, *dsbld*, *wrkctra*, *estsz*, *jbspv* (0.134), dngna, *pdjobyr*, *dngnapp* (0.114), netusoft, *icpart1*, *hinctnta*, *actrolga*, *hincfel*, *hswrkp*, *anctry1*, *rshpsts*, *dngref* (0.077) SVM: netusoft (0.201), netustm, *uempli*, *uempla* (0.052) MLP: netusoft (0.160), gndr3 (0.118), wkdcorga, *wkhtot*, *icwhct*, *wkhct*, *tporgwk*, *uemp3m*, *mbtru*, *wrkac6m*, *nacer2*, *isco08* (0.087)
prtdgcl	RF: 0.363 SVM: 0.222 MLP: 0.352	RF: trstplc, *trstun*, *trstep*, *trstplt*, *trstprt*, *trstplc* (0.072), pbldmn, *stfedu*, *stfhlth* (0.067) SVM: netusoft, *pdjobyr*, *hswrkp*, *hincfel*, *dngref*, *rshpsts*, *pdwrk*, *icpart1* (0.077) MLP: trstplc (0.084), pbldmn (0.072)
rfgbfml	RF: 0.770 SVM: 0.24 MLP: 0.768	RF: – SVM: – MLP: elghydr, *elghydr*, *basinc*, *eusclbf* (0.056)
rfgfrpc	RF: 0.623 SVM: 0.603 MLP: 0.613	RF: – SVM: elghydr, *slvuemp* (0.130) MLP: elghydr (0.054)
rlgblge	RF: 0.990 SVM: 0.971 MLP: 0.990	RF: dscrrce, *dscretn*, *dsrlng*, *dscrntn*, *dscrrlg* (0.442) SVM: dscrrce (0.457) MLP: dscrrce (0.475)
rlgdnm	RF: 0.991 SVM: 0.977 MLP: 0.991	RF: dscrrce, *dscretn*, *dsrlng*, *dscrntn*, *dscrrlg* (0.433) SVM: dscrrce (0.460) MLP: dscrrce (0.381)
rshipa2	RF: 0.999 SVM: 0.967 MLP: 0.998	RF: yrbrn2, *lvgptnea*, *rshipa3* (0.112), icpart1, *eduyrs*, *pdwrk*, *dvrcdeva*, *icpart1* (0.057) SVM: yrbrn2 (0.269) MLP: yrbrn2 (0.341)
rshipa3	RF: 0.995 SVM: 0.974 MLP: 0.993	RF: yrbrn2, *eduyrs*, *icpart1*, *dvrcdeva*, *pdwrk* (0.504) SVM: yrbrn2 (0.451) MLP: yrbrn2 (0.492)
rshpsts	RF: 0.764 SVM: 0.595 MLP: 0.708	RF: maritalb, *uemplip*, *chldhm* (0.188), yrbrn2, *icpart2*, *dngdkp*, *edulvlb*, *eiscedf*, *uemplap*, *wkhtotp*, *emprelp*, *isco08p*, *edctnp*, *marsts*, *eisced*, *edulvlpb*, *domicil*, *lvgptnea*, *yrbrn2*, *chldhhe*, *pdwrkp* (0.119) SVM: yrbrn2 (0.148), cntry, *dweight* (0.081) MLP: yrbrn2 (0.249)
sbbsntx	RF: 0.527 SVM: 0.506 MLP: 0.0	RF: – SVM: uentrjb, *admub*, *bnlwinc* (0.055) MLP: –

Target	Performance (Matthews)	Important features
tporgwk	RF: 0.927 SVM: 0.893 MLP: 0.926	RF: – SVM: wkdcorga, *wkhtot*, *icwhct*, *wkhct*, *nacer2* (0.084) MLP: wkdcorga (0.101)
trstep	RF: 0.350 SVM: 0.256 MLP: 0.0	RF: psppsgva, *regunit*, *gndr*, *region*, *rfgbfml*, *agea*, *gndr2* (0.065) SVM: – MLP: –
trstun	RF: 0.337 SVM: 0.218 MLP: 0.318	RF: – SVM: – MLP: –
uemplwk	RF: 0.204 SVM: 0.040 MLP: 0.0	RF: – SVM: – MLP: –
vteurmmb	RF: 0.905 SVM: 0.895 MLP: 0.903	RF: psppsgva, *regunit*, *region*, *rfgbfml*, *gndr3*, *gndr2* (0.317) SVM: psppsgva (0.333) MLP: psppsgva (0.378)
wkdcorga	RF: 0.933 SVM: 0.911 MLP: 0.934	RF: iorgact, *uemp3m*, *wkhct*, *icwhct*, *isco08*, *tporgwk*, *wrkac6m*, *mbtru* (0.069) SVM: iorgact (0.085) MLP: iorgact (0.155)
wkhct	RF: 0.652 SVM: 0.582 MLP: 0.0	RF: jbspv, *wkdcorga*, *icwhct*, *tporgwk*, *wrkac6m*, *mbtru*, *isco08*, *nacer2*, *iorgact*, *uemp3m*, *wkhtot* (0.090) SVM: jbspv (0.161) MLP: –
wkhtot	RF: 0.710 SVM: 0.694 MLP: 0.700	RF: jbspv, *wkdcorga*, *icwhct*, *tporgwk*, *wrkac6m*, *mbtru*, *isco08*, *nacer2*, *iorgact*, *uemp3m*, *wkhtot* (0.083) SVM: jbspv (0.076) MLP: jbspv (0.140)
wkhtotp	RF: 0.725 SVM: 0.610 MLP: 0.707	RF: lvgptnea, *icpart2*, *eisced*, *chldhhe*, *domicil*, *edulvlpb*, *edctnp*, *isco08p*, *emprelp*, *eiscedf*, *eiscedp*, *pdwrkp*, *iccohbt*, *chldhm*, *dngrefp*, *icppdwk* (0.190), netusoft, *hinctnta*, *cmsrvp*, *dngoth* (0.082), marsts, *maritalb* (0.078), polint, *rfgfrpc*, *regunit*, *region* (0.054) SVM: lvgptnea (0.135), netusoft (0.111), netustm, *uempla*, *wrkac6m*, *tporgwk*, *edctn* (0.071), rshpsts, *edulvlb*, *dvrcdeva*, *rshpsts*, *eduyrs* (0.066) MLP: lvgptnea (0.217), netusoft (0.106)

Target	**Performance (Matthews)**	**Important features**
wrkac6m	RF: 0.971 SVM: 0.923 MLP: 971	RF: – SVM: – MLP: icwhct, *tporgwk*, *wkhtot*, *nacer2*, *uemp3m*, *isco08* (0.066), wkdcorga, *mbtru*, *iorgact*, *wkhct*, *isco08* (0.065)
wrkctra	RF: 0.968 SVM: 0.950 MLP: 0.967	RF: iorgact, *isco08*, *tporgwk*, *wrkac6m*, *mbtru*, *nacer2*, *uemp3m*, *wkhct*, *icwhct*, *wkhtot* (0.164) SVM: iorgact (0.322) MLP: iorgact (0.193)
wrkprbf	RF: 0.202 SVM: 0.069 MLP: 0.0	RF: – SVM: – MLP: –
yrbrn2	RF: 0.967 SVM: 0.942 MLP: 0.967	RF: yrbrn3, *wkhtotp*, *domicil*, *eiscedf*, *eisced*, *emprelp*, *isco08p*, *rshipa3*, *dngdkp*, *edctnp*, *dvrcdeva*, *edulvlb*, *edulvlpb*, *chldhhe*, *uemplap*, *pdwrkp*, *marsts*, *iccohbt* (0.119), icpart1, *icpart2*, *eduyrs*, *pdwrk*, *lvgptnea* (0.072) SVM: yrbrn3 (0.272) MLP: yrbrn3 (0.317)
yrbrn3	RF: 0.980 SVM: 0.949 MLP: 0.978	RF: yrbrn2, *pdwrk*, *icpart2*, *yrbrn2*, *icpart1*, *eduyrs* (0.546) SVM: yrbrn2 (0.427) MLP: yrbrn2 (0.475)

Appendix F

Results of the postliminary analysis

Figure F.1: Performance in Matthews correlation coefficient and F1 score for the 64 Random Forest models with multiple imputation

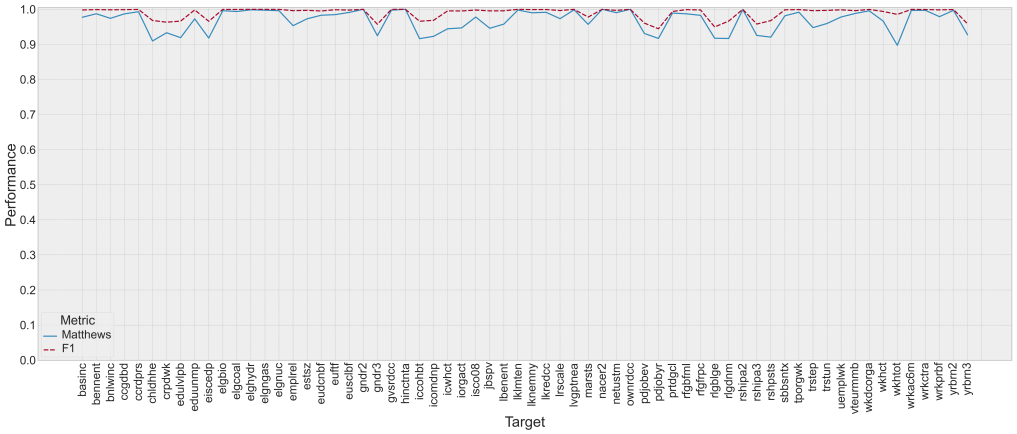

Table F.1: Average performance of the RF models with multiple imputation

Metric	Average
Accuracy	0.989
Balanced accuracy	0.979
ROC score	0.979
F1 score	0.990
Matthews coefficient	0.978

Table F.2: Table of the five most informative features on average for the Random Forests with multiple imputed data for the subset of 64 targets, ranked in descending order

Rank	Feature	SHAP value	% missing
1	stfedu	0.077	3.5%
2	eisced	0.067	0.3%
3	gndr2	0.065	21.8%
4	jbspv	0.063	8.4%
5	agea	0.061	0.3%
6	dscrntn	0.060	0.0%
7	estsz	0.054	11.1%
8	dscrrce	0.052	0.0%
9	gndr	0.051	0.0%
10	maritalb	0.050	2.0%

Appendix G

Descriptions of the features

Table G.1: Feature abbreviations and their descriptions

Feature	Description
cntry	Country
pspwght	Post-stratification weight including design weight
anweight	Analysis weight
nwspol	Time spent consuming news about politics and current affairs
netusoft	Frequency of internet use
netustm	Time spent using the internet on a typical day
pplfair	Attitude towards "most people try to take advantage of you vs. try to be fair"
psppsgva	Attitude towards "the political system allows people to have a say in what the government does"
trstplc	Trust in the police
contplt	Whether respondent has contacted a politician or government official in the last 12 months
wrkprty	Whether respondent has worked in a political party or action group in the last 12 months
stfeco	How satisfied with the present state of the economy in the country
stfgov	How satisfied with the national government
mnrgtjb	Attitude towards "men should have more right to a job than women when jobs are scarce"
freehms	Attitude towards "gays and lesbians are free to live life as they wish"
imsmetn	Attitude towards allowing many immigrants of the same race or ethnic group as majority
imdfetn	Attitude towards allowing many immigrants of different races or ethnic groups from the majority
inprdsc	With how many people can intimate and personal matters be discussed with
aesfdrk	Feeling of safety of walking alone in the local area after dark
health	Subjective general health
hlthhmp	How often hampered in daily activities by illness, disability, infirmity, or mental problem
rlgdnm	Current religion or denomination
dscrrce	Whether or not respondent's group is discriminated in terms of color or race
dscrgnd	Whether or not respondent's group is discriminated in terms of gender
dscrref	Whether or not respondent refuses to answer the questions about discrimination
dscrna	Whether or not respondent has not given an answer to the questions about discrimination
ctzcntr	Country of which respondent is a citizen
brncntr	Country of birth
rfgfrpc	Attitude towards "most refugee applicants are not in real fear of persecution in their own countries"

Feature	Description
elgcoal	Attitude towards how much electricity should be generated from coal in a particular country
elgngas	Attitude towards how much electricity should be generated from natural gas in a particular country
elgnuc	Attitude towards how much electricity should be generated from nuclear power in a particular country
wrenexp	How worried about energy being too expensive for many people
wrdpfos	How worried about the country being too dependent on fossil fuels
clmthgt2	How much respondent has thought about climate change before the day of the interview
sbsrnen	Whether or not respondent favors subsidizing renewable energy to reduce climate change
gvslvue	Attitude towards "the standard of living for the unemployed is the responsibility of the government"
sbprvpv	Attitude towards "social benefits and services prevent widespread poverty"
sbeqsoc	Attitude towards "social benefits and services lead to a more equal society"
sblwcoa	Attitude towards "social benefits and services make people less willing to care for one another"
eduunmp	Whether or not more money should be spent on education for the unemployed at the cost of unemployment benefit
eudcnbf	Attitude towards "more decisions made by the EU will lead to a higher level of benefits of a particular country"
lknemny	Likelihood of not having enough money for the household necessities in the next 12 months
gndr	Gender of the respondent
gndr3	Gender of third person in the household
agea	Calculated age of the respondent
dvrcdeva	Whether or not the respondent has ever been divorced or a civil union dissolved
icpart2	Whether or not the respondent lives with husband, wife or partner (interviewer code)
edulvlb	Highest level of education
uempli	Whether or not one is unemployed and not actively looking for a job in the 7 days before the interview
dsbld	Whether or not one is permanently sick or disabled in the 7 days before the interview
hswrk	Whether or not is doing housework, looking after children or other in the 7 days before the interview
dngoth	Whether or not is doing something other in the 7 days before the interview
dngref	Whether or not respondent refuses to answer what done in the 7 days before the interview
dngdk	Whether or not respondent does not know what done in the 7 days before the interview
dngna	Whether or not no answer given for question what done in the 7 days before the interview
wrkctra	Type of contract (duration)
wkdcorga	How much influence respondent has on organizing their own daily work
edctnp	Whether or not partner of respondent has followed an education in the 7 days before the interview
edulvlfb	Highest level of education of father of respondent
occf14b	Occupation of father of respondent when respondent was 14
occm14b	Occupation of mother of respondent when respondent was 14
imprich	Attitude towards "it is important to be rich, and to have money and expensive things"
ipeqopt	Attitude towards "it is important that people are treated equally and have equal opportunities"

Feature	Description
impsafe	Attitude towards “it is important to live in secure and safe surroundings”
ipudrst	Attitude towards “it is important to understand different people”
inwmms	Month of the start of the interview
inwyys	Year of the start of the interview
trstun	Trust in the United Nations
prtdgcl	How close is respondent to a political party
stfdem	How satisfied with the way democracy works in the country
imueclt	Attitude towards “the cultural life of the country is enriched by immigrants”
rlgatnd	Frequency of attending religious services apart from special occasions
rdcenr	How often doing something to reduce energy use
wrinspw	How worried that energy supply is interrupted by insufficient power generated
uemplwk	How many are unemployed and looking for work, of every 100 working age
gndr2	Gender of second person in the household
rshipa2	Relationship to respondent of second person in household
rshipa3	Relationship to respondent of third person in household
eisced	Highest level of education in ES-ISCED scale
ipgdtim	Attitude towards “it is important to have a good time”
stflife	How satisfied with life as a whole
elgwind	How much electricity in a particular country should be generated from wind power
smdfslv	Attitude towards “for a fair society, differences in standard of living should be small”
bennent	Attitude towards “many manage to obtain benefits or services they are not entitled to”
yrbrn	Birth year of respondent
edctn	Whether or not is following education in the 7 days before the interview
uemplap	Whether or not partner of respondent is unemployed and not actively looking for a job in the 7 days before the interview
dsbldp	Whether or not partner of respondent is permanently sick or disabled in the 7 days before the interview
ipshabt	Attitude towards “it is important to show abilities and be admired”
lklmten	Attitude towards “it is likely that large numbers of people limit their energy use”
slvpens	Attitude towards “the standard of living of pensioners is good”
lbenent	Attitude towards “many with very low incomes get less benefit than legally entitled to”
maritalb	Legal marital status, post coded
estsz	Establishment size
impdiff	Attitude towards “it is important to try new and different things in life”
inwdde	Day of the month of the end of the interview
slvuemp	Attitude towards “the standard of living of unemployed is good”
polintr	How interested in politics
actrolga	Whether or not respondent is able to take an active role in a political group
psppipla	Attitude towards “the political system allows people to have an influence on politics”
cptppola	Whether or not respondent is confident in their own ability to participate in politics
trstplt	Trust in politicians
sclact	Relative frequency of taking part in social activities compared to others of the same age
elghydr	How much electricity in a particular country should be generated from hydroelectric power
wrntdis	How worried about energy supply being interrupted by natural disasters or extreme weather

Feature	Description
ccnthum	Attitude towards "climate change is caused by natural processes, human activity or both"
lkredcc	Attitude towards "when large numbers of people limit their energy use, it is likely that climate change will reduce"
gvcldcr	Attitude towards "child care services for working parents are the responsibility of the government"
wrkprbf	Attitude towards "there should be benefits for parents to combine work and family, even if this means higher taxes"
rshpsts	Relationship with husband, wife or partner currently living with
ipadvnt	Attitude towards "it is important to seek adventures and have an exciting life"
trstep	Trust in the European Parliament
stfhlth	Attitude towards "the current state of the health services in the country is good"
gvsrdcc	Attitude towards "it is likely that governments in enough countries take action to reduce climate change"
gvslvol	Attitude towards "the standard of living for the old is the responsibility of the government"
imsclbn	When should immigrants obtain rights to social benefits or services
ipfrule	Attitude towards "it is important to do what is told and follow the rules"
ppltrst	Attitude towards "most people can be trusted"
euftf	Attitude towards "European unification can go further"
impcntr	Attitude towards allowing immigrants from poorer countries outside Europe
atcherp	How emotionally attached to Europe
ipmodst	Attitude towards "it is important to be humble and modest and to not draw attention"
wrclmch	How worried about climate change
uempla	Whether or not respondent is unemployed and actively looking for a job in the 7 days before the interview
emprf14	Employment status of father of respondent when respondent was 14
edulvlmb	Highest level of education of mother of respondent
rlgblg	Whether respondent belongs to a particular religion or denomination
wrdpimp	How worried about the country being too dependent on energy imports
ipsuces	Attitude towards "it is important to be successful and that people recognize achievements"
rfgbfml	Attitude towards "granted refugees should be entitled to bring close family members"
ccrdprs	To what extent does respondent feel personal responsibility to reduce climate change
elgsun	How much electricity in a particular country should be generated from solar power
stfedu	Attitude towards "the current state of education in the country is good"
crmvct	Whether the respondent or a household member was a victim of burglary or assault in the last 5 years
dscroth	Whether or not respondent's group is discriminated in terms of other grounds
gincdif	Attitude towards "the government should reduce differences in income levels"
eusclbf	Against or in favor of an EU-wide social benefit scheme
lkuemp	How likely is it for the respondent to be unemployed and looking for work in the next 12 months
trstprt	Trust in political parties
atchctr	How emotionally attached to the country
jbspv	Whether respondent is responsible for supervising other employees
hmsfmlsh	Attitude towards "ashamed if close family member is gay or lesbian"

Feature	Description
rlgblge	Ever belonging to a particular religion or denomination
yrbrn3	Birth year of third person in household
cmsrv	Whether or not partner of respondent is doing community or military service in 7 days before interview
tporgwk	Type of organization working or worked for
emprelp	Employment relation of partner
icwhct	Whether respondent has a set basic or contracted number of hours
cflsenr	How confident respondent can use less energy than currently
sblazy	Attitude towards "social benefits or services make people lazy"
iorgact	How much influence respondent has on policy decisions about activities of an organization
banhhap	Attitude towards "sale of the least energy efficient household appliances should be banned to reduce climate change
uentrjb	Attitude towards "most unemployed people do not really try to find a job"
trstlgl	Trust in the legal system
vteurmmb	Whether respondent would vote for the country to remain a member of the EU or leave
impfree	Attitude towards "it is important to make your own decisions and be free"
ccgdbd	Attitude towards "climate change has a good impact across the world"
vote	Whether or not respondent has voted in the last election
sgnptit	Whether or not respondent has signed a petition in the last 12 months
lvgptnea	Whether respondent has ever lived with a partner, without being married
edulvlpb	Highest level of education partner
dngdkp	Whether or not respondent does not know what partner has done in the 7 days before the interview
pplhlp	Attitude towards "most of the time people are helpful"
pray	How often praying apart from at religious services
eneffap	How likely is it for the respondent to buy the most energy efficient home appliance
iphlppl	Attitude towards "it is important to help people and care for others well-being"
bnlwinc	Attitude towards "social benefits should only be for people with the lowest incomes"
domicil	Domicile of respondent
wrpwrct	How worried about power cuts
hincfel	Feeling about income of household nowadays
elgbio	How much electricity in a particular country should be generated from biomass energy
icpart1	Whether or not the respondent lives with husband, wife or partner (interviewer code)
dweight	Design weight
pweight	Population size weight
ipcrtiv	Attitude towards "it is important to think of new ideas and be creative"
rtrdp	Whether or not partner of respondent is retired in the 7 days before the interview
blgetmg	Whether or not respondent belongs to a minority ethnic group in the country
region	Region of the respondent
regunit	Regional unit of the respondent
hincsrca	Main source of household income
anctry1	First ancestry in European Standard Classification of Cultural and Ethnic Groups
trstprl	Trust in the parliament of the country
lrscale	Placement on left and right scale (political orientation)

Feature	Description
wrkorg	Whether or not respondent has worked in another (not a political party or action group) organization or association in the last 12 months
badge	Whether respondent has worn or displayed a campaign badge or sticker in the last 12 months
pbldmn	Whether or not respondent has taken part in a lawful public demonstration in the last 12 months
bctprd	Whether or not respondent has boycotted certain products in the last 12 months
pstplonl	Whether or not respondent has posted or shared anything about politics online in the last 12 months
clsprty	Whether or not respondent feels closer to a particular party than all other parties (politics)
sclmeet	How often socially meeting with friends, relatives or colleagues
icpart3	Whether or not the respondent lives with husband, wife or partner (interviewer code)
hmsacld	Attitude towards "gay and lesbian couples have a right to adopt children"
imbgeco	Attitude towards "immigration is good for the economy of the country"
imwbcnt	Attitude towards "immigrants make the country a better place to live"
happy	How happy is respondent
dscrdk	Whether or not respondent does not know if group of respondent is discriminated
rtrd	Whether or not respondent is retired in the 7 days before the interview
rlgdgr	How religious is respondent
dscrgrp	Whether or not respondent is a member of a group discriminated against in the country
dscrntn	Whether or not respondent is discriminated in terms of nationality
dscrrlg	Whether or not respondent is discriminated in terms of religion
dscrlng	Whether or not respondent is discriminated in terms of language
dscretn	Whether or not respondent is discriminated in terms of ethnic group
dscrage	Whether or not respondent is discriminated in terms of age
facntr	Whether or not the father of the respondent is born in the country
dscrsex	Whether or not respondent is discriminated in terms of sexuality
dscrdsb	Whether or not respondent is discriminated in terms of disability
icomdng	How many activities doing by partner coded (interview code)
lnghom1	First most spoken language at home
mocntr	Whether mother is born in the country
gvrfgap	Attitude towards "the government should be generous in judging applications for refugee status"
ownrdcc	How likely is it for the respondent to limit own energy use to reduce climate change
inctxff	Attitude towards "taxes on fossil fuels should be increased to reduce climate change"
wrtcfl	How worried about energy supply being interrupted by technical failures
wrtratc	How worried about energy supply being interrupted by terrorist attacks
clmchng	Attitude towards "the climate of the world is changing"
dfincac	Attitude towards "large differences in income are acceptable to reward talents and efforts"
sbstrec	Attitude towards "social benefits and services place a too great strain on the economy"
admub	Administration of unemployment benefits questions
sbbsntx	Attitude towards "social benefits and services cost businesses too much in taxes and charges"
hhmmb	Number of people living regularly as a member of the household
crpdwk	Whether or not respondent had control paid work in the 7 days before the interview
pdjobev	Whether or not respondent ever had a paid job

Feature	Description
pdjobyr	The last year of the paid job
chldhhe	Whether or not respondent ever had children living in the household
iccohbt	Whether or not respondent is cohabiting (interviewer code)
marsts	Legal marital status
eiscedp	highest level of education of partner in ES-ISCED scale
pdwrkp	Whether or not partner of respondent did paid work in the 7 days before the interview
dngrefp	Whether or not respondent refuses to answer what partner has done in the 7 days before the interview
icppdwk	Whether or not partner of respondent is in paid work (interviewer code)
isco08p	Occupation of partner of respondent in ISCO08 code
yrbrn2	Year of birth of second person in household
icpdwrk	Whether or not respondent is in paid work (interviewer code)
emplrel	Employment relation
hinctnta	Total net income of all sources of household
cmsrvp	Whether or not respondent is in community or military service in 7 days before the interview
uemplip	Whether or not partner is unemployed and not actively looking for a job in the 7 days before interview
chldhm	Whether or not children are living at home or not
eduyrs	Years of fulltime education completed
pdwrk	Whether or not respondent is doing paid work in 7 days before the interview
mbtru	Whether or not respondent is or was a member of a trade union or similar organization
hswrkp	Whether or not partner of respondent is doing housework, looking after children or other in the 7 days before the interview
mnactic	Main activity in the last 7 days of all respondents, post coded
dngothp	Whether or not partner of respondent is doing something other in the 7 days before the interview
dngnapp	Whether or not question about what partner of respondent is doing in the 7 days before the interview is not applicable
icomdnp	Whether or not partner of respondent is doing more than one activity in 7 days before interview (interviewer code)
dngnap	Whether or not question about what respondent is doing in the 7 days before the interview is not applicable
wkhtot	Total hours normally worked per week in main job with overtime
nacer2	Industry type in NACE rev.2
isco08	Occupation of respondent in ISCO08 code
wkhct	Total contracted hours per week in main job excluding overtime
wrkac6m	Whether or not respondent has paid work in another country of at least 6 months in the past 10 years
uemp3m	Whether or not respondent was ever unemployed and seeking work for a period of more than 3 months
wkhtotp	Hours normally worked in a week in main job with overtime, of partner
eiscedf	Highest level of education of father in ES-ISCED
emprm14	Employment status of mother of respondent when respondent was 14
eiscedm	Highest level of education of mother in ES-ISCED

Feature	Description
atncrse	Whether or not respondent has improved knowledge or skills by a course, lecture or conference in the last 12 months
impenv	Attitude towards "it is important to care for nature and the environment"
imptrad	Attitude towards "it is important to follow traditions and customs"
iplylfr	Attitude towards "it is important to be loyal to friends and devote to people that are close"
impfun	Attitude towards "it is important to seek fun and things that give pleasure"
ipstrgv	Attitude towards "it is important that the government is strong and ensures safety"
ipbhprp	Attitude towards "it is important to behave properly"
inwdds	Day of the month of the start of the interview
iprspot	Attitude towards "it is important to get respect from others"
inwmme	Month of the end of the interview
inwyye	Year of the end of the interview
inwtm	Interview length in minutes of main questionnaire
basinc	Attitude towards "there should be a basic income scheme"
dscrnap	Answered 'not applicable' to question discrimination to group of respondent

Appendix H

Proportion of missingness for each feature

column name,	% missing	column name,	% missing	column name,	% missing
gndr12	99.988735%	prtclbis	99.004213%	rlgdngb	97.938586%
rshipa12	99.988735%	yrbrn7	98.997454%	rlgdnach	97.904792%
yrbrn12	99.988735%	gndr7	98.950143%	edupagb2	97.902539%
yrbrn11	99.979724%	rshipa7	98.934373%	edagepgb	97.900286%
rshipa11	99.977471%	prtclgpl	98.909591%	edupbgb1	97.893527%
gndr11	99.977471%	prtclblt	98.875797%	prtclbse	97.893527%
rlgdelt	99.941424%	rlgdnnl	98.754140%	prtvtbit	97.870998%
yrbrn10	99.936918%	rlgdnase	98.736116%	edlvpdch	97.839457%
gndr10	99.925654%	edlvpdis	98.691058%	clmthgt1	97.823687%
rshipa10	99.923401%	prtclfhu	98.641494%	prtclefr	97.807917%
rlgdehu	99.882849%	prtvtesi	98.535607%	edlvpenl	97.794399%
yrbrn9	99.867078%	prtvtbis	98.407191%	rshpsgb	97.780882%
gndr9	99.846802%	prtclept	98.337351%	edlvpdse	97.744835%
rshipa9	99.844549%	edlvpdpt	98.332845%	edlvpdno	97.742582%
edupail2	99.806250%	prtclcit	98.328339%	edlvpepl	97.740329%
rlgdeapl	99.788226%	prtvtfch	98.285534%	prtcldil	97.717800%
yrbrn8	99.653052%	prtvtcpt	98.281028%	eduppl2	97.715547%
rlgdease	99.637281%	prtcldcz	98.231464%	prtvblt2	97.702030%
gndr8	99.628270%	prtclfee	98.208935%	prtclbgb	97.697524%
rshipa8	99.619258%	rlgdnbe	98.195418%	marstfi	97.693018%
rlgdeis	99.567441%	rlgdnhu	98.186406%	prtcldfi	97.693018%
edufail2	99.515624%	prtclfch	98.170636%	prtvblt1	97.686259%
rlgdeafi	99.504359%	rlgdnno	98.145854%	prtclbno	97.650213%
rlgdebat	99.495348%	edlvpesi	98.112060%	prtvtdpl	97.638948%
edumail2	99.484083%	edlvfdis	98.098542%	prtvblt3	97.636695%
rlgdeach	99.407484%	prtclcbe	98.082772%	prtvtcfr	97.614166%
rlgdeno	99.382702%	edlvmdis	98.069255%	rlgdnafi	97.602902%
rlgdegb	99.337644%	prtcldie	98.062496%	prtclees	97.598396%
rlgdeie	99.258792%	prtvtehu	98.044472%	edlvpdfr	97.589384%
rlgdebe	99.215987%	prtclcat	98.042220%	edlvpebe	97.548832%
rlgdnis	99.069547%	edlvdis	98.028702%	edlvpdru	97.539820%
rlgdeade	99.058283%	prtcldru	98.008426%	edlvpeat	97.494762%
prtclesi	99.035754%	prtclenl	98.003920%	prtvtdcz	97.442945%
rlgdenl	99.006466%	edlvpdhu	97.992656%	edlvfdpt	97.415910%

column name,	% missing	column name,	% missing	column name,	% missing
prtvtbno	97.402393%	edlvmepl	96.395341%	edlvfdcz	95.165251%
edlvpdee	97.397887%	edlvdhu	96.375065%	edlvmdru	95.149481%
prtvtfee	97.370852%	edlvfebe	96.372812%	edufbil1	94.991777%
prtvtdru	97.357334%	edupbil1	96.368306%	edlvmdcz	94.987271%
marstgb	97.352829%	edufagb2	96.366053%	edlvdcz	94.888143%
prtvtfnl	97.341564%	edufbgb1	96.348030%	edumbil1	94.692140%
edlvpdfi	97.330299%	edlvmebe	96.320995%	edlvdru	94.525424%
rshpsfi	97.328047%	edlvenl	96.217361%	edlvfdit	94.484872%
edlvpdlt	97.298759%	edumbgb1	96.194832%	edlvmdit	94.421790%
edlvpfes	97.285241%	edumagb2	96.194832%	edlvfdie	94.322662%
edlvmdpt	97.278482%	edlvepl	96.194832%	edlvmdie	94.291121%
edlvfesi	97.197378%	edupl2	96.183567%	edufade3	94.255075%
edlvpdcz	97.188366%	edlvfdfr	96.129497%	edubil1	94.250569%
prtvtbse	97.156825%	edlvebe	96.032622%	edufbde1	94.246063%
edlvdpt	97.145561%	prtvtbat	96.019105%	eduail2	94.241557%
prtvtcbe	97.143308%	edupade3	95.992070%	edlvdit	94.131165%
edlvmesi	97.136549%	edupade2	95.987564%	edumade3	94.090612%
prtvtdes	97.134296%	edlvfdlt	95.983058%	edufade2	94.070336%
edlvesi	97.062203%	edupbde1	95.978552%	edumbde1	93.975714%
yrbrn6	97.021650%	prtvtcil	95.971794%	edumade2	93.865321%
prtvtbgb	96.996868%	edlvfdee	95.962782%	mnactp	93.833780%
prtvtdfi	96.956316%	prtvtbie	95.958276%	edlvdie	93.813504%
gndr6	96.940546%	edlvmdfr	95.917724%	eduade3	93.606236%
rshipa6	96.904499%	edlvfdfi	95.868160%	eduade2	93.601730%
edlvpdit	96.877464%	prtvede1	95.865907%	edubde1	93.594972%
edlvfdse	96.868452%	edlvffes	95.856895%	intewde	93.574695%
edlvfdch	96.789601%	rlgdnlt	95.838872%	yrbrn5	91.087481%
rlgdnbat	96.726519%	prtvede2	95.814090%	gndr5	90.925271%
edlvmdse	96.719760%	vteumbgb	95.802825%	rshipa5	90.873454%
edlvmdch	96.717507%	edlvmdfi	95.771284%	cntbrthc	90.413860%
edlvfdhu	96.694978%	edlvmfes	95.744249%	livecnta	89.965530%
edlvpdie	96.694978%	edlvfeat	95.672156%	rlgdnme	89.875414%
edlvfenl	96.663437%	edlvdfi	95.667650%	emplno	89.481154%
edagefgb	96.649920%	edlvmdlt	95.649627%	ctzshipc	89.071124%
edlvfdno	96.627391%	edlvmdee	95.633857%	mainact	86.469011%
prtclede	96.625138%	eduagb2	95.629351%	mbrncntb	85.937324%
edlvmdno	96.584586%	edubgb1	95.629351%	lnghom2	85.831437%
edlvmdhu	96.577827%	edagegb	95.624845%	fbrncntb	85.358326%
rlgdnapl	96.575574%	edlvges	95.595557%	vteubcmb	82.456575%
edlvdch	96.575574%	edlvfdru	95.573028%	crpdwkp	77.542974%
edlvmenl	96.535022%	edlvmeat	95.570775%	ub50unp	77.466375%
edlvdno	96.521504%	edlveat	95.478406%	ub50edu	77.024805%
edlvdse	96.514745%	edlvdee	95.453624%	ubspunp	76.903147%
edlvfepl	96.512492%	rlgdnie	95.419830%	ub50pay	76.894136%
rlgdnade	96.510239%	edlvdfr	95.340978%	ub20unp	76.785996%
edagemgb	96.503481%	edlvdlt	95.228333%	ubspedu	76.562958%

column name,	% missing	column name,	% missing	column name,	% missing
ubunp	76.508888%	lrscale	13.075901%	wrdpfos	4.095794%
ubsppay	76.389483%	rfgfrpc	11.706130%	wrclmch	3.904296%
ub20edu	76.324149%	lvgptnea	11.320882%	banhhap	3.802915%
ub20pay	76.182216%	estsz	11.138396%	emprf14	3.793904%
ubedu	75.965936%	lbenent	10.469282%	elgsun	3.746592%
ubpay	75.866808%	eiscedf	10.066010%	hmsfmlsh	3.649717%
yrbrn4	75.103071%	edulvlfb	10.066010%	stfedu	3.534819%
gndr4	74.729087%	nacer2	10.007435%	imwbcnt	3.519048%
anctry2	74.720076%	isco08	9.597405%	imbgeco	3.519048%
rshipa4	74.638971%	iorgact	9.214410%	sblwcoa	3.483002%
njbspv	74.258229%	tporgwk	9.063465%	sbeqsoc	3.363597%
uemp5yr	72.050375%	wkdcorga	8.752563%	stfdem	3.359092%
uemp12m	72.045869%	basinc	8.549801%	wrdpimp	3.277987%
wkhtotp	67.253926%	lknemny	8.504742%	hmsacld	3.266722%
pdjobyr	66.156758%	emplrel	8.398856%	impcntr	3.266722%
isco08p	64.489603%	jbspv	8.385338%	sbsrnen	3.163088%
emprelp	63.719107%	wrkac6m	8.252416%	imueclt	3.160835%
rlgblge	59.817064%	euftf	8.135265%	sbprvpv	3.018902%
yrbrn3	57.208192%	icwhct	8.121747%	lkuemp	2.944556%
prtdgcl	57.043729%	wrkprbf	7.927997%	gvrfgap	2.870210%
pdjobev	56.744092%	trstep	7.801834%	slvuemp	2.863451%
gndr3	56.446707%	trstun	7.792822%	imdfetn	2.818393%
rshipa3	56.406155%	sbbsntx	7.488679%	clmthgt2	2.757564%
crpdwk	54.684930%	gvsrdcc	7.454885%	freehms	2.710253%
marsts	49.739789%	elgcoal	7.423345%	clsprty	2.653930%
occm14b	42.758015%	lkredcc	7.333228%	stfgov	2.615631%
edulvlpb	42.523712%	eduunmp	7.249870%	imsmetn	2.593102%
eiscedp	42.523712%	elgbio	7.128213%	inwtm	2.561561%
iccohbt	41.703652%	ccgdbd	7.107937%	eneffap	2.502985%
rshpsts	41.681123%	uemplwk	6.889405%	clmchng	2.473697%
icomdnp	41.649582%	elgngas	6.824070%	uentrjb	2.419627%
icppdwk	41.647329%	rfgbfml	6.727195%	sblazy	2.322752%
rlgdnm	41.147183%	lklmten	6.706919%	ipfrule	2.300223%
chldhhe	35.147678%	edulvlmb	6.641584%	ipstrgv	2.257418%
vteurmmb	33.385901%	eiscedm	6.641584%	iprspot	2.257418%
netustm	32.158064%	elgnuc	6.585261%	cptppola	2.248406%
eudcnbf	28.438507%	bennent	6.537950%	pray	2.198842%
eusclbf	26.584360%	ownrdcc	6.157208%	emprm14	2.189830%
yrbrn2	22.961678%	bnlwinc	5.771960%	psppsgva	2.158290%
rshipa2	21.850992%	ccrdprs	5.542163%	cflsenr	2.144772%
gndr2	21.758623%	elghydr	5.046523%	actrolga	2.106473%
wrkctra	19.746773%	sbstrec	4.873048%	ipsuces	2.090702%
wkhct	18.769009%	ccnthum	4.850519%	ipshabt	2.018609%
hinctnta	17.892626%	inctxff	4.474283%	stfeco	1.996080%
occf14b	15.457228%	imsclbn	4.469777%	ipudrst	1.987068%
wkhtot	14.407372%	elgwind	4.332350%	maritalb	1.978057%

column name,	% missing	column name,	% missing	column name,	% missing
ipbhprp	1.973551%	rlgatnd	0.786266%	inwsmm	0.076599%
trstprl	1.966792%	dscrgrp	0.770496%	inwdds	0.051817%
dfincac	1.953275%	trstplc	0.720932%	inwyys	0.049564%
ipmodst	1.932998%	anctry1	0.718679%	inwmms	0.049564%
trstprt	1.926240%	rdcenr	0.716426%	ctzcntr	0.047311%
sclact	1.926240%	stfhlth	0.714173%	admub	0.042805%
trstlgl	1.912722%	pplfair	0.709667%	brncntr	0.038300%
ipcrtiv	1.908216%	rlgblg	0.678126%	icomdng	0.027035%
psppipla	1.896952%	lnghom1	0.626310%	icpdwrk	0.024782%
ipadvnt	1.890193%	gvslvol	0.590263%	chldhm	0.024782%
impdiff	1.858652%	dvrcdeva	0.585757%	inwdde	0.022529%
impfun	1.856399%	facntr	0.533940%	inwmme	0.020276%
ipeqopt	1.847388%	atchctr	0.529434%	inwyye	0.020276%
ipgdtim	1.838376%	atncrse	0.515917%	gndr	0.020276%
iphlppl	1.804582%	mbtru	0.500146%	region	0.002253%
smdfslv	1.793318%	uemp3m	0.491135%	name	None
impfree	1.793318%	happy	0.484376%	essround	None
imprich	1.786559%	bctprd	0.482123%	edition	None
iplylfr	1.750513%	sclmeet	0.470859%	proddate	None
impenv	1.709960%	sgnptit	0.423547%	idno	None
imptrad	1.698696%	stflife	0.421295%	cntry	None
impsafe	1.680672%	icpart3	0.403271%	dweight	None
wrinspw	1.604073%	icpart2	0.398765%	pspwght	None
wrtratc	1.579291%	icpart1	0.398765%	pweight	None
hincsrca	1.561268%	pplhlp	0.396512%	anweight	None
atcherp	1.550003%	pstplonl	0.396512%	dscrrce	None
gincdif	1.513957%	hlthhmp	0.392007%	dscrntn	None
slvpens	1.484669%	crmvct	0.364972%	dscrrlg	None
trstplt	1.455381%	agea	0.349201%	dscrlng	None
gvcldcr	1.448622%	yrbrn	0.349201%	dscretn	None
wrtcfl	1.257125%	pbldmn	0.304143%	dscrage	None
gvslvue	1.236849%	mnactic	0.299637%	dscrgnd	None
hincfel	1.180526%	wrkorg	0.290626%	dscrsex	None
nwspol	1.180526%	edulvlb	0.290626%	dscrdsb	None
inprdsc	1.173767%	eisced	0.290626%	dscroth	None
mnrgtjb	1.124203%	badge	0.283867%	dscrdk	None
vote	1.054363%	hhmmb	0.279361%	dscrref	None
wrntdis	1.040845%	contplt	0.274855%	dscrnap	None
wrpwrct	0.995787%	ppltrst	0.259085%	dscrna	None
blgetmg	0.993534%	wrkprty	0.252326%	pdwrk	None
wrenexp	0.973258%	mocntr	0.236556%	edctn	None
eduyrs	0.955235%	polintr	0.218532%	uempla	None
inwehh	0.925947%	health	0.132922%	uempli	None
inwemm	0.923694%	domicil	0.112646%	dsbld	None
rlgdgr	0.907923%	netusoft	0.110393%	rtrd	None
aesfdrk	0.889900%	inwshh	0.078852%	cmsrv	None

column name,	% missing	column name,	% missing	column name,	% missing
hswrk	None	uemplap	None	dngnapp	None
dngoth	None	uemplip	None	dngrefp	None
dngref	None	dsbldp	None	dngdkp	None
dngdk	None	rtrdp	None	dngnap	None
dngna	None	cmsrvp	None	regunit	None
pdwrkp	None	hswrkp	None		
edctnp	None	dngothp	None		

Appendix I

Proportion of overlap between the target and predictors

column name,	avg overlap,	max overlap	column name,	avg overlap,	max overlap
wkhtotp	4.36%	41.66%	estsz	0.98%	9.27%
isco08	3.88%	43.14%	sbbsntx	0.92%	5.23%
emprelp	3.73%	42.84%	euftf	0.91%	5.88%
rshipa3	3.59%	56.27%	gvsrdcc	0.91%	5.48%
gndr3	3.29%	56.33%	wrkprbf	0.89%	5.48%
pdjobyr	3.05%	38.00%	elgcoal	0.87%	5.49%
prtdgcl	3.02%	33.79%	basinc	0.87%	6.00%
icomdnp	2.93%	27.94%	lkredcc	0.86%	5.61%
icppdwk	2.93%	27.94%	elgngas	0.86%	5.07%
rshpsts	2.90%	27.94%	eduunmp	0.83%	4.99%
rlgblge	2.87%	32.96%	elgbio	0.82%	5.25%
iccohbt	2.82%	27.95%	trstun	0.82%	5.40%
yrbrn3	2.81%	33.47%	lklmten	0.81%	5.61%
occm14b	2.75%	28.53%	ccgdbd	0.79%	4.98%
pdjobev	2.73%	31.78%	nacer2	0.78%	8.16%
eiscedp	2.70%	28.38%	isco08	0.77%	8.43%
edulvlpb	2.70%	28.38%	occf14b	0.77%	9.32%
crpdwk	2.28%	30.66%	trstep	0.76%	5.50%
netustm	2.19%	23.51%	tporgwk	0.76%	8.15%
eudcnbf	2.03%	15.31%	uemplwk	0.74%	4.96%
vteurmmb	2.03%	20.15%	elgnuc	0.74%	5.15%
marsts	2.03%	26.74%	lbenent	0.73%	6.41%
rlgdnm	1.95%	24.28%	iorgact	0.71%	8.67%
rshipa2	1.91%	21.85%	ownrdcc	0.71%	4.98%
yrbrn2	1.72%	22.81%	bennent	0.70%	4.29%
eusclbf	1.70%	14.32%	wrkac6m	0.68%	8.02%
gndr2	1.63%	21.72%	wkdcorga	0.66%	8.67%
chldhhe	1.52%	22.34%	eiscedm	0.65%	4.60%
hinctnta	1.41%	11.51%	edulvlmb	0.65%	4.60%
wkhct	1.35%	9.60%	bnlwinc	0.65%	3.97%
lrscale	1.12%	8.55%	rfgbfml	0.65%	4.19%
wkhtot	1.09%	8.23%	elghydr	0.64%	3.88%
rfgfrpc	1.02%	7.38%	jbspv	0.62%	8.16%
wrkctra	0.99%	10.12%	eiscedf	0.62%	5.66%

column name,	**avg overlap,**	**max overlap**	**column name,**	**avg overlap,**	**max overlap**
edulvlfb	0.62%	5.66%	ccrdprs	0.58%	4.25%
emplrel	0.60%	8.09%	lvgptnea	0.49%	6.02%
lknemny	0.58%	6.01%			
icwhct	0.58%	7.98%			