How to Predict Recidivism in Young Offenders: Comparing Logistic Regression, Random Forest, and Boosted Classification Trees and Examining Possible Risk Factors.

Leonie Waldeck

Master's thesis Forensic Psychology

Final version

Student number: i6190915

Internal supervisor: Dr. Jan Schepers, Faculty of Psychology and Neuroscience Maastricht University

External supervisor: Dr. Maja Meischner-Al-Mousawi, Kriminologischer Dienst des Freistaates Sachsen

Internship location: Kriminologischer Dienst des Freistaates Sachsen

13,719 words

24.07.2020

Abstract

One important goal of a young offender institution in Germany is to encourage the young offenders (YO) to not commit another crime in the future. In order to enhance the effectiveness of interventions that aim at this goal, research is needed concerning the question under what circumstances a YO recidivates. Traditionally, logistic regression is used in order to analyze recidivism data. Here it is argued that the use of tree-based statistical methods might yield new insights for the prediction of recidivism. It is hypothesized that boosted classification trees yield better predictions than random forest, and that random forests yield better predictions than logistic regression. This study involved 643 YO that were released from the German young offender institution Regis-Breitingen. Predictors included in this study were demographic information, the interventions a YO participated in, an assessment by professionals and a self-assessment. The outcome criterion was whether or not the YO received a new court decision (except for acquittal) within two years after release. The results showed that there was no difference in predictive performance between the methods. All methods performed poorly. Static risk factors for recidivism were younger age and a shorter amount of time spent in the young offender institution. Dynamic risk factors for recidivism included the YO having no place in work, vocational training or school after release, having a need to continue structured transition management after release as well as having participated in delict- or problem specific measures. A possible reason for poor predictive performance is heterogeneity of the YO. Implications for further research and policy making are discussed.

Acknowledgements

I am extremely grateful to Dr. Maja Meischner-Al-Mousawi, Sven Hartenstein and Sylvette Hinz: Thank you for allowing me to work on the dataset, for inspiring discussions, clear feedback, and a fantastic internship that encouraged me to pursue a career in the field of forensic psychology.

I also would like to thank Dr. Jan Schepers for agreeing to be my second supervisor, for valuable advice and feedback, and for answering all my questions.

Writing this thesis during the Covid-19 lockdown would have been a very different experience without housemates full of ideas on how to end working days while staying at home. Thank you!

Finally, I wish to give special thanks to Birgit, Thomas, Nelly and Thorge for their optimism and loving support throughout the whole of my studies.

Table of contents

Introduction

Both crime- and recidivism rates are higher among juveniles than among adults (Heinz, 2004). This emphasizes the question of how to deal with young offenders (YO). According to German law, the goal of young offender institutions (YOI) in Saxony is to enable the YO to live a socially responsible life without committing offenses in the future (§ 2 Sächsisches Jugendstrafvollzugsgesetz from 2007). In order to achieve these goals, research concerning the risk factors for recidivism in YO is essential (Walter, 2006). This is due to the fact that knowledge of these risk factors can help professionals make informed decisions about both interventions and estimations about recidivism risk (Jimerson et al., 2004). To gain that knowledge, one question that needs to be answered is what the most appropriate methods are for the prediction of recidivism in YO. Another question is what variables are most relevant for the prediction of recidivism. The aim of this paper is to examine both the predictability of and the risk factors for recidivism in YO that have been incarcerated in the YOI of Regis-Breitingen in the German federal state of Saxony.

Methods for the prediction of recidivism

The first goal of this paper is to find an appropriate method for the prediction of recidivism. Below, the requirements for the methods are described. This is followed by an explanation of why the methods used in this study were chosen.

Interpretability versus flexibility

"Everything should be made as simple as possible, but no simpler" is a quote of Albert Einstein. As one of many scientists in history he endorsed the principle of parsimony, which states that if there are multiple explanations for a phenomenon, the simplest is to be preferred in science (Vandekerckhove et al., 2015). Thus, there are two sides of a coin when choosing between different methods. On the one hand, the method needs to explain the phenomenon. In order to explain the phenomenon, the method needs to fit the data, which means that it needs to be flexible. Flexibility refers to the degree to which a method can handle different types of associations in order to fit the structure of the data at hand (G. James et al., 2013). Generally, flexible methods yield higher prediction performances than inflexible methods. On the other hand, a method should be as interpretable as possible. Interpretability is defined as the ability to understandably explain or present the relationship between a predictor variable and an outcome to a human being (Doshi-Velez & Kim, 2017). When choosing a method, a trade-off (compromise) has to be made between interpretability and flexibility of a method: Methods that are high on interpretability tend to be low on flexibility and vice versa. Thus, the higher the flexibility of a method is, the better are the predictions but the worse is the interpretability and vice versa (Alonso et al., 2015). Methods that are highly flexible are referred to as “black boxes”: They provide an output for the input, but what exactly happens in between is poorly understood (Adadi & Berrada, 2018).

This raises the question of what properties are appropriate when analyzing recidivism data. Regarding the flexibility of the method, self-evidently it is desirable to obtain predictions that are as accurate as possible. Regarding the interpretability of the method, using a method that lacks a thorough understanding of how the predictions are made is not suitable for the field of risk prediction for two reasons. First, it would be unethical to make decisions that have an impact on peoples’ lives based on an algorithm that is poorly understood (Rudin et al., 2018; Tonry, 2013). Second, an understanding of risk factors is essential in order to develop and evaluate interventions. Therefore, a “black box” method is not suitable for risk prediction. In summary, an effective method for predicting recidivism risk is as accurate as possible, while interpretability needs to remain present.

Logistic Regression

Logistic Regression (LR; see explanation below) is a traditional statistical method that is often used in scientific studies addressing risk factors (McReynolds et al., 2010; Mulder et al., 2011). One advantage of LR is its high interpretability: Explaining the relationship between a predictor and an outcome is quite simple (G. James et al., 2013). Furthermore, LR directly calculates the probability that a respondent belongs to a particular class, which is useful in many contexts. However, a disadvantage of LR is its low flexibility, which leads to poorer predictions especially when a complex data structure is involved. Additionally, LR relies on a number of assumptions that need to be met in order to allow for the inference of valid conclusions. The first assumption is that the continuous variables need to be linearly related to the log odds of the outcome variable (Stoltzfus, 2011). The second assumption is the absence of multicollinearity, which means that the independent variables must not be strongly associated with each other. Third, outliers can strongly influence the results of LR, therefore they need to be eliminated. These assumptions are often not met in real-world data, which leads to incorrect inferences (Hosmer et al., 1991; Ottenbacher et al., 2004; Westreich et al., 2011). Other disadvantages of LR are that dummy variables need to be created and that interaction effects are often neglected. This raises the question of what method might be better suited for the prediction of recidivism.

Tree models

Compared to LR, tree models (see method section for an explanation) are more flexible (G. James et al., 2013). Thus, tree tend to have a low bias, which is the error that is introduced by approximating a complex problem by a simple model (G. James et al., 2013). Another advantage is that they relax the assumptions about the data: They can handle highly nonlinear data structures as well as outliers and missing values, and the predictor variables can be correlated with each other (Buskirk & Kolenikov, 2015; Mendez et al., 2008). Additionally, the prediction of tree models does not depend on the measurement scales or the distributions of the predictors (Elith et al., 2008). Therefore, tree models are straightforward to implement. As explained before, the increased flexibility of tree models compared to LR comes with a decreased interpretability. Nonetheless the interpretability of tree models is considered good because tree models mirror human decision-making and they can be well displayed graphically (G. James et al., 2013). Single classification trees are not a powerful tool for prediction because they often overfit the training dataset: They are instable in the sense that small deviations in the data can lead to significant changes in the tree and thus in its interpretation (Austin et al., 2013; G. James et al., 2013). Therefore, they tend to have a high variance, which is the amount to which the model would change if a different training dataset had been used. The predictive performance of classification trees can largely increase when multiple trees are built and combined. Hereinafter, two of these approaches are discussed.

Random Forest

A Random Forest (RF; see method section for an explanation) is a modified tree model. By combining multiple trees, flexibility of the model is increased. All advantages concerning data structure and assumptions of tree models explained above apply to RF (G. James et al., 2013). Additionally, RF is well suited for calculations with a large number of variables compared to the number of respondents, and the predictions of RF are more stable than those of a single tree. Another advantage of RF is that the default settings for the hyperparameters[1] often work well, which makes it quite easy to tune the model (Sun et al., 2016; Svetnik et al., 2000). As mentioned before, a disadvantage of RF is that the increased flexibility compared to tree models comes with a decreased interpretability. Furthermore, since interaction effects are included automatically in RF, the interpretation of the relationship between a variable and an outcome is less straightforward than in LR. In addition, due to the fact that a large number of trees are built for a RF, it is no longer possible to display the model in one simple graph. However, a summary of the importance of each predictor can be obtained, which can then be displayed graphically. Generally, the interpretability of RF is considered acceptable (Ali et al., 2012; Robinson et al., 2017). Its combination of interpretability and flexibility has been empathized in the literature (Gao et al., 2017; Qi, 2012).

RF has been found to deliver superior prediction results compared to LR in multiple studies, especially studies that involve real-world data (Couronné et al., 2018; Feng & Wang, 2017; Guo et al., 2004; Hsieh et al., 2011; Maroco et al., 2011; Muchlinski et al., 2016), even though contradicting results have also been found (Ruiz & Villa, 2008; Weng et al., 2015). Ultimately, the performance of a method depends on the structure of the dataset at hand. However, it is not yet known which method to choose when analyzing a certain data structure (Kirasich et al., 2018). Since the majority of previous research articles favors RF, it is expected that RF yields a higher prediction performance than LR on recidivism data. Thus, even though its interpretability is lower than that of LR, RF might be better suited for analyzing recidivism data due to higher flexibility and good-enough interpretability.

Boosted Classification Trees

Just as RF, Boosted Classification Trees (BCT; see method section for an explanation) are modified tree models. However, the way in which the trees are combined makes BCT more flexible than RF. The advantages of tree models with regard to data structure and assumptions explained above apply to BCT as well (G. James et al., 2013). A disadvantage of BCT is that the model performance strongly depends on the tuning of the model, which increases the complexity of using BCT. Furthermore, the increased flexibility of BCT again comes with a decreased interpretability. There is no operationalization of interpretability and

---

[1] Hyperparameters are parameters that are not learned by the model, but that need to be determined before the learning process takes place. They have an influence on the way a method makes its predictions (see appendix C).

therefore no defined measure that could tell whether interpretability of BCT is sufficient for contexts in which interpretability is regarded important. BCT are used in situations in which flexibility is considered more important than interpretability (Lee et al., 2019) and in situations in which flexibility and interpretability are considered approximately equally important (Yang et al., 2016). Furthermore, research currently aims at finding techniques that help interpret the results yielded by BCT, as for instance by designing clear graphs (Gill et al., 2020). Thereby the interpretability of BCT might be increased in the future. Taking these findings together, it might be argued that the interpretability of BCT is just good-enough for the prediction of recidivism which would mean that the flexibility that BCT offers is the highest flexibility that can be used responsibly on a recidivism dataset.

Research has shown that BCT are powerful predictors (Hutchinson et al., 2011), especially when the structure of the data is complex and nonlinear (Pittman & Brown, 2011). Many machine learning[2] competitions have been won using BCT (Liu et al., 2017). Studies comparing RF and BCT have reached different conclusions: Often, BCT yielded a higher prediction performance than RF (G. James et al., 2013; Naghibi et al., 2016; Ogutu et al., 2011). However, some authors suggest that their results might be quite similar (Brillante et al., 2015; Wainer, 2016) and some favor RF (Andersson, 2017; Khalilia et al., 2011). Based on earlier research, it is expected that BCT will lead to better predictions than RF.

Hypotheses

In summary, the combination of interpretability and flexibility in RF is remarkable and therefore using RF on recidivism data could potentially yield new insights. The **first hypothesis in this paper states that RF yields a higher prediction performance than LR on recidivism data**[3]. Next, the interpretability of BCT might be considered just good enough for research on recidivism. While keeping that limitation in mind, it is interesting to examine how one of the most flexible methods that can be used on recidivism data without the sacrifice of too much interpretability performs. The **second hypothesis states that BCT yields a higher prediction performance than RF on recidivism data**. Thus, in terms of prediction it is expected that BCT yields the best results, followed by RF and LR respectively.

[2] Machine learning takes place when an algorithm learns the associations between predictors and an outcome variable based on an observed dataset. Subsequently, this algorithm can be used in order to make predictions for datasets in which the predictors are observed and the outcome is unknown.

[3] The prediction performance of a method depends on the amount of YO that are correctly classified as YO who recidivates and YO who does not recidivate. For a detailed explanation, see the method section.

Variables relevant for the prediction of recidivism

The second goal of this paper is to evaluate which predictor variables are chosen by the methods. An elaborate literature review on the risk factors for recidivism in youth delinquents is beyond the scope of this thesis, since the predictors were chosen from variables that were already available. Nonetheless, a brief literature review is given in the following, in order to better understand the predictors that the different methods choose.

Table 1 illustrates the variables that were associated with recidivism in a meta-analysis involving 23 studies and 15,265 YO (Cottle et al., 2001). Almost all these variables are static, which means that they are unchangeable by intervention. Static risk factors of recidivism in YO have been frequently addressed in the literature, and there is a widespread agreement in the scientific community that the variables displayed in table 1 are risk factors for recidivism in YO (Hong et al., 2013; Mulder et al., 2011).

For clinicians working in prison however, dynamic factors are of more interest because they are changeable (Hanson & Morton-Bourgon, 2009). Research on dynamic risk factors is scarcer than research on static risk factors, since dynamic factors are more difficult to operationalize. Furthermore, from an age of around 14 on, stable risk factors generally have a higher influence on recidivism than dynamic risk factors (van der Put et al., 2012). Dynamic risk factors proposed in the literature include a lack of treatment adherence, having parents with poor parenting skills and criminal behavior in the family (Mulder et al., 2011). Furthermore, anger or irritability has been found to be predictive of recidivism in YO (Hong et al., 2013). Other dynamic factors proposed in the literature include gang membership, denial, substance abuse and self-depreciation (Benda et al., 2001). Protective dynamic factors include strong social support and strong attachment to prosocial adults (Lodewijks et al., 2010). Even if there is no agreement concerning the precise dynamic risk factors for recidivism in YO, these variables paint a picture of disadvantaged youth in complex situations, who are in need of stability and support (MacRae et al., 2011).

Table 1

Variables significantly (p < 0.01) associated with higher recidivism in YO in a meta-analysis by Cottle et al. (2001)

Category	Significantly related variables
Demographic information	being male
	low socioeconomic background
Offense history	earlier age at first contact with the law
	earlier age at first crime commitment
	more prior arrests
	more previous crime commitments
	longer first incarcerations
	more previous crimes committed
Family and social factors	history of having been physically or sexually abused
	raised in a single-parent home
	greater number of out-of-home placements
	significant family problems
	inefficient use of leisure time
	delinquent peers
Educational factors	history of special education
Standardized test score	lower standardized achievement score
	verbal IQ score
	performance IQ score
	full scale IQ score
Substance use history	substance abuse
Clinical factors	conduct problems
	non-severe pathology
	higher scores on risk assessment instruments

Methods

Explanation of methods used

Logistic regression

The logistic model has been introduced by Verhulst (1838). The logistic function is:

$$p(Y = 1 \mid X) = \frac{e^{\beta 0+\beta 1X1+\cdots+\beta pXp}}{1+e^{\beta 0+\beta 1X1+\cdots+\beta pXp}}.$$

This function can be used as a classifier[4] by estimating its regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ on a training dataset[5] (G. James et al., 2013). Hereby the maximum likelihood approach is used, such that the coefficients correspond as closely as possible to the training data observed for every individual. Next, the estimated parameters are used to predict the probability that an individual from the testing dataset belongs to a class (P(Y=1)). These probabilities can then be transformed into class predictions, for instance with the rule that all cases with probabilities larger than 0.5 belong to class 1, and all cases with probabilities smaller than 0.5 belong to class 0. The predictions are then compared to the observed outcome values in the testing dataset[6].

One coefficient of LR can only compare two categories of a variable even if that variable has more than two categories. Therefore, LR estimates multiple coefficients when dealing with a variable that has multiple categories: One of the categories of that variable is determined to be the reference category and for all other categories, a coefficient is estimated that compares this category against the reference category. For that reason, with LR the number of coefficients that needs to be estimated can largely exceed the number of variables. In order to reduce the number of coefficients that need to be estimated, parameter shrinkage can be used, which means that the coefficients of the less contributive variables are shrunken towards zero. The more parameters are shrunken, the more the variance is reduced (which comes with an increase in bias). Three shrinkage methods are available (G. James et al., 2013). The first option is ridge regression, in which variables that have little contribution are shrunken close to zero. When using ridge regression, variance is reduced. However, the number of parameters is not reduced, since even for small nonzero numbers, parameters need to be estimated. The second option is lasso regression. In lasso regression, variables that have

---

[4] A classifier is an algorithm that can assign a categorical class label to an observation.

[5] The training dataset is one part of the original dataset. It is used to obtain parameter estimates.

[6] The testing dataset is the other part of the original dataset. It provides an unbiased evaluation of the model predictions.

little contribution are shrunken to be exactly zero. As a consequence, variance is reduced and less parameters need to be estimated. The third option is elastic net regression, which is a combination of the first two options: Coefficients with the smallest contribution are shrunken to be exactly zero, while coefficients with a slightly larger contribution are shrunken to be close to zero.

Classification tree models

Classification tree models are hierarchical decision trees which divide the feature space[7] according to splitting rules (G. James et al., 2013). An example of a splitting rule is: "If a value on variable X is smaller than the threshold T, take the left branch; if a value on variable X is equal to or larger than T, take the right branch". The variable X and the threshold T are determined on the training set by maximizing the Gini gain. In order to explain the Gini gain, the Gini impurity needs to be introduced first. The Gini impurity is calculated based on the probability that a randomly chosen datapoint is misclassified when classifying according to the class distributions in the dataset (suppose there are four green objects and three red objects in a bowl; then the probability of misclassifying a green object is $3/7$). At a particular node[8], the Gini impurity is defined as:

$$GI = \sum_{j=1}^{C} p(j) * (1 - p(j)),$$

where *C* is the number of classes and *p(j)* is the relative frequency of class j in the data. A perfect Gini impurity has the value 0, and means that at a certain node, the probability of wrongly classifying a randomly chosen datapoint according to the class distributions is 0. The Gini gain then is the reduction in Gini impurity when a split is made, calculated by weighting the Gini impurity of each branch by how many elements there are in it. At each node, the threshold for the split of the node is determined such that the Gini gain becomes as large as possible (Oh et al., 2003). This process is followed until the last node. Based on this decision tree, predictions can be made for the testing data. Last, the predictions can be compared to the observed outcome values of the testing set, in order to determine the predictive ability of the tree. An example of a classification tree is shown in appendix A.

---

[7] The feature space of a dataset is a space that has as many dimensions as there are variables in the dataset. Each observation of a variable is represented as a point in the feature space.

[8] The point in a tree at which a split is made on a variable is called a node (for an example, see the boxes in figure A1).

Random Forest

RF has been introduced by Breiman (2001) and is a statistical learning method that builds multiple decision trees using two rules. The first rule is called bootstrap aggregation. It states that the individual trees are not built on every observation of the training dataset. For every tree, a sample of the same size as the original training set is drawn from the whole training set randomly with replacement. Thus, every tree is built on a different training set, which reduces the variance of the predictions. The second rule is that at each split of the tree, the next predictor is not chosen from all predictors available but from a random sample of *m* predictors. The reason for this is that if the next predictor is chosen from all predictors and one of them is very strong, this one will be chosen as the first predictor in most trees. As a result, the trees built will be very similar to each other, which reduces the advantageous effect of building multiple trees (Breiman, 2001). In contrast, by choosing only from a subsample *m* of predictors, the trees are more divers. After growing all trees according to these two rules, the trees can be used on the training dataset. Per observation, the category that is most often predicted by these trees is used as the prediction of RF.

Boosted Classification Trees

BCT have originally been proposed by Friedmann (2001). The idea of boosting is that multiple models that are averaged into one are more likely to yield correct predictions than one single model (Elith et al., 2008; Lawrence et al., 2004). Thus, like RF, a BCT is a combination of multiple trees. However, in contrast to RF, the trees are not grown simultaneously but sequentially in BCT: New trees are built using information from previously grown trees (G. James et al., 2013; Natekin & Knoll, 2013). This is done by assigning less weight to the accurately classified instances and more weight to the inaccurately classified instances of the current tree. Due to these assigned weights, the next tree then will focus on correctly classifying those datapoints that have been wrongly classified in the previous tree. This process continues until the last tree. Afterwards, the trees are combined: For each observation, the predicted final class is given by a weighted average of the different trees. The weights are now distributed the other way around: More weight is given to those trees that have produced less error, and less weight is given to those trees that have produced more error. The goal thereby is to minimize the distance between the observed values and the prediction of the tree.

Processing of the data

The data was available in the MySQL database of the Criminological Service of Saxony[9]. It was collected in order to evaluate the juvenile penalty system, which is prescribed by law (Hartenstein et al., 2016). The tables that were of interest for the current study originally included 663 observations of 840 variables. All analyses were carried out in R (R Core Team, 2014). In order to assure anonymity, the official identification numbers referring to the YO were never downloaded from the database. They were only used for joining the tables on the server side before the data was downloaded.

The total dataset

The dataset included male YO who have been detained in the YOI of the German federal state of Saxony in Regis-Breitingen. Female YO could not be included because they are placed in another location and different data is collected for them. Furthermore, only YO who were in the YOI for at least 183 days (half a year) could be included because only then most of the data was collected. Another inclusion criterion was applied with regard to the time frame. The year 2013 was chosen as a start date, because the central registry[10] data was available only from then on. However, due to privacy reasons, central registry data can be deleted if someone has not recidivated within five years. Thus, it could have been the case that the criminal record of someone who was released in 2013 and got a new criminal record entry until May 2014, was deleted by Mai 2019 (when the central registry data was obtained). Before carrying out the analyses, this was checked by calculating the recidivism rates for all cohorts[11]. Since the recidivism rate for the cohort of 2013 was not lower than the recidivism rates of the other cohorts, it was assumed that no (or at least not a lot of) recidivism data had been deleted from the cohort of 2013. Therefore, the cohort of 2013 was kept in the analyses. The year 2016 was chosen as the end date, because central registry data is available until then (and because it allows for an observation time of two years, see below). In total, 663 YO met these inclusion criteria.

---

[9] The Criminological Service of Saxony is a department of the penal system in Saxony that accompanies and evaluates the penal system scientifically.

[10] The central registry belongs to the Federal Office of Justice. It is accessible by the police and individuals can request their criminal record. Additionally, it can be requested for research purposes.

[11] The recidivism rates for the cohorts 2013, 2014, 2015 and 2016 were 0.57, 0.53, 0.57 and 0.54 respectively.

Data was retrieved from six tables that concerned YO sentences. The first table contained **demographic and general information (DGI)** about the YO. From the second table, the types of **crimes committed by the YO (CC)** were added to the dataset. The third source of information was the **Social Workers Release Questionnaire (SWRQ),** a questionnaire that contained professional estimations of items concerning the coping style and situation of the YO. Fourth, variables from the **Intervention Questionnaire (IQ)** were added to the dataset. This questionnaire contained information rated by a professional regarding the interventions that the YO participated in, including the degree to which the intervention goals were successfully met. The fifth source of information was the **Release Questionnaire (RQ)**. This questionnaire was filled in by the YO on a voluntary basis. It measured different aspects of the YO's own view of his life and of his release situation.

The sixth table contained the outcome criterion. It was created with information obtained from the **German Bureau of Criminal Records (CR)**, including 68 variables with different recidivism information about the YO (when/ how he recidivated). From these variables, a binary outcome variable was calculated that contained the information whether a new entry to the criminal record of the YO was made (i.e. a crime was committed after the YO was released from the YOI) within two years after his release. This variable was added to the dataset. The reason for choosing a time frame of two years was that it allowed for the cohort released in 2016 to be included in the analysis. Literature on recidivism in YO has consistently shown that the longer someone has not recidivated, the smaller the probability is that he or she will recidivate in the future. In an observation period of twenty years, the vast majority recidivates within five years (Stelly & Thomas, 2005), and that in an observation period of five years, the vast majority recidivates within three years (Jehle et al., 2016). Therefore, it was expected that an observation time of two years was reasonable.

Cleaning of the dataset

Respondents for which whole questionnaires were missing by fault (thus, all except for the voluntary RQ) were removed from the dataset. For one case, both SWRQ and IQ had not been filled in. For nine cases, only the information assessed by SWRQ was lacking. For seven cases, only the IQ had not been filled in. Furthermore, for three cases, the CR information whether the YO recidivated within two years was lacking. In total, 20 respondents were eliminated from the dataset, such that the number of YO included in the analysis amounted to 643. The voluntary RQ had not been filled in by 170 YOs. Data imputation was used for these, see below. Additionally, the information whether a YO took part in an intervention was missing by fault three times. For these, the variable whether they began participation in the intervention was filled in, because it was assumed that they were more likely to start the intervention if they had a need to.

Table 2 shows the number of variables included in the current study after different steps of transformation. Originally, these tables contained 840 variables in total. Out of these, 285 variables were selected as relevant for the current dataset. Variables not selected were for instance text variables, comments, identification numbers, entry date and variables referring to the person who entered the data. Furthermore, some questionnaires had been updated in the end of 2014. Only the variables that were assessed in both versions were kept. Last, some variables were originally included in two tables and these were only once selected for the current dataset.

From the selected variables, some were combined with each other (see table 2). For instance, from the variables that contained the date of birth[12], date of entry and date of release from the YOI, two new variables were calculated: The amount of days spent in the YOI and age on the day of release. Furthermore, some information was documented in two different tables throughout the years. Thus, some variables were combined into one. The RQ contained groups of continuous variables that assessed different aspects of the YO's lives (for instance the evaluation of the release situation, see appendix F). After rescaling all variables into the same direction (1 = low and 5 = high), for every area of life one variable was calculated that averaged all variables measuring that life aspect. This was done without considering missing values. Thus, if a YO filled in only some of the questions assessing one life area, the mean of the variables that he filled in was taken as the value for that life aspect. The reason for this was that the combination of these variables was assumed to bring less noise to the data than data imputation. Likewise, the different areas of what a YO reached in the YOI were added up.

Recoding of the variables

For the categorical variables, the goal of recoding was to reduce the number of coefficients for which prediction was needed while losing as little information as possible. First, the categories “not relevant” and “not known” were combined into one. Second,

[12] A random number between -10 and 10 was added to the dates, in order to ensure anonymity.

variable categories were combined as in Jehle et al. (2016): “no” and “rudimentary” were combined into one category, as well as “almost” and “fully”. Third, two categories were made for the different reasons for not beginning a treatment: One if the YO was “not willing or able to”, since poor treatment motivation, low intellectual capabilities and aggression during treatment have been proposed to be predictive of recidivism (Cottle et al., 2001; Farrington et al., 2016; Harder et al., 2015; Mulder et al., 2011). The other category was made for responses that have not yet been proposed to be predictive of recidivism, for instance if the treatment was currently not offered in the YOI. Since only a small number of YO were not born in Germany, they were combined into one codification. Furthermore, the question whether they were in contact with their probation assistant had three categories: “No”, “yes in writing” and “yes in person”. The two “yes” categories were combined, because only 25 YO were in contact with their probation assistant in person. Additionally, for the variable whether a YO “had work or vocational training after release”, work and vocational training were combined. This was done for both a practical and a theoretical reason. The practical reason was that missing values were imputed for this variable, and imputation of five distinct categories is likely to introduce a high degree of noise. The theoretical reason was that work and vocational training have been proposed to be protective factors due to the provision of stability and structured daily routines (Wößner & Wienhausen-Knezevic, 2013). These properties are present in both work and vocational training.

In order to make the coefficients of LR more easily understandable, the same codification was used for all variables. For “no information available”, a -1 was used. “No”-answers were coded with a 0. “Yes”-answers were coded with a 1 (or above, if there were different degrees of strength).

New variables

As a next step, new variables were calculated in order to reduce the information lost when variables with near zero variances were excluded (see table 2). With regard to CC, one variable was calculated that added up all categories of crimes committed by one YO. Regarding IQ, the dataset contained the information whether a YO had a need to participate in, began participation in, dropped out of, reached the goals of and after release needs to participate in 19 different types of interventions. Since only a subset of YO took part in a certain treatment, new variables were created that combined the total amount of intervention categories applicable per YO. Furthermore, a new variable was created that added up the self-assessment of what they reached during their time in the YOI. These seven new variables were added to the dataset.

Table 2

Number of variables included in the current study from the six tables after selecting and combining variables, adding new variables and eliminating variables with near zero variance

	DGI	CC	SWRQ	IQ	RQ	CR	total
original	64	92	79	144	393	68	840
selected	4	13	35	133	99	1	285
combined	3	10	29	95	20	1	158
new variables	3	11	29	100	21	1	165
near zero variance	3	8	26	79	21	1	138

Near zero variance

A variable has a near zero variance if the vast majority of its observations share the same value (Kuhn, 2008). This variable contains little information but increases the model complexity. Therefore, variables with near zero variances are often excluded when the goal is to make predictions. In this thesis, the default settings of the R-package "caret" were used to exclude variables with near zero variance: A variable was defined to have near zero variance if two conditions were present (Kuhn et al., 2020). The first condition was present if the ratio of the most common value to the second most common value that was larger than 95 to five. The second condition was present if less than ten percent of the values were distinct from the total sample. 27 variables were eliminated due to variances near zero. In total, the dataset consisted of 138 variables (137 predictors and one outcome variable).

Variable imputation

170 YO (26.44%) did not fill in the RQ at all, and those who did fill it in did not answer all questions. In order to impute unbiased values based on the variables that are available, the data needs to be completely missing at random[13] or missing at random[14]

[13] Data is completely missing at random if observations that contain missing values are a random subset of observations that contain no missing values.

[14] Data is missing at random if the probability that an observation contains missing values depends on other variables available in the dataset.

(Donders et al., 2006). A two-sample *t*-test revealed that at an alpha level of 0.05 there was no difference in recidivism between those YO who filled in the RQ and those who did not ($t(295.99) = -1.05$, $p = .30$). It was possible that whether a YO filled in the RQ was dependent on unobserved characteristics like motivation or social desirability, and there was no way to find out if this was in fact the case (Jakobsen et al., 2017). However, the possibility that some noise was added to the data was preferred over the possibility of not considering the RQ variables at all[15].

Before imputing missing values, the following procedure was carried out to determine the best imputation method for every individual variable M that contained missing data. First, the proportion of datapoints that was missing for M was eliminated from the observed values of M in order to obtain a testing dataset for the imputation methods. Second, these eliminated datapoints were imputed by applying the methods most commonly used for data imputation. For the continuous variables, pmm[16], sample[17], cart[18], RF, norm[19] and hotdeck[20] were used. For the categorical variables, LR, sample, cart and RF were used. For the variable with three categories, polyreg[21], sample, cart and RF were used. In order to ensure reproducibility, a fixed seed[22] was set. Ten iterations were carried out five times. Third, the imputations of the different methods were compared to the datapoints that had been eliminated in order to find the imputation method that yielded the best predictions for variable M. For continuous variables, the best prediction was defined as the one that showed the highest correlation with the observed values. For categorical variables, the best prediction was defined as the one that showed the highest accuracy when compared to the observed values. The calculations were carried out using the R-packages VIM (Templ et al., 2020) and mice (van Buuren et al., 2020).

---

[15] An elimination of all YO who did not fill in the RQ would only be appropriate if the data was missing completely at random.

[16] predictive mean matching: produces a set of possible imputation values and subsequently imputes the value that fits best based on other variables

[17] sample: imputes a random sample from the observed values

[18] Classification And Regression Tree: uses a single tree model for imputation

[19] norm: uses Bayesian linear regression

[20] hotdeck: imputes the observed value of a respondent similar on other variables

[21] Polytomous logistic regression: uses multinomial logistic regression

[22] Throughout the study, a seed of 1902 was set for every calculation that depended on a random number generator

Subsequently, the imputation method that yielded the best prediction for variable M was used to impute the real missing data for variable M. Again, ten iterations were carried out five times. For the continuous variables, the mean of the five prediction results was used as imputed data. For the categorical variables, the majority vote of the five predictions was used as imputed data. For an overview of the chosen methods and the accuracy/ correlation with the observed values see appendix B.

Class imbalance

The overall recidivism rate within two years in the current sample was 0.55. Thus, a problem of class imbalance occurred, which means that the classification categories are not approximately equally represented (Chawla, 2005). An imbalance of 0.55/ 0.45 is considered small, especially compared to other real-world problems as for instance fraud detection (Wei et al., 2013). However, it was possible that this imbalance would have an effect on how well the minority class was predicted (Japkowicz, 2000). This was due to the fact that if the methods predicted “recidivism” for all YO, the accuracy was already 55%. Therefore, the methods might have been biased towards predicting the majority class (recidivism; Akbani et al., 2004).

The problem of class imbalance is often addressed by using a resampling method in order to balance the class distribution of the training set. There are two different resampling approaches: oversampling and undersampling (Kotsiantis et al., 2006). In oversampling, the number of observations in the minority class are increased. This is either done by simply replicating the minority class observations or by creating new, synthetic minority class observations that follow the structure of the observed ones (Chawla et al., 2002). The major disadvantage of oversampling is that it changes the data structure of the minority class. In undersampling, observations from the majority class are randomly eliminated (Kotsiantis et al., 2006). The major disadvantage of undersampling is that valuable information of the majority class might be lost. However, since the imbalance was small in this case, only a small number of datapoints had to be eliminated. This option was preferred over changing the data structure of the minority class. Therefore, undersampling was considered a better solution than oversampling.

In order to find out if the methods indeed profit from an undersampled training dataset, the analyses were carried out twice: once with the original training dataset and once with a randomly undersampled training dataset. The undersampled dataset was created by randomly eliminating 44 observations from the original training dataset, so that both recidivism and nonrecidivism were observed 191 times in the undersampled training dataset. Due to the fact that the testing dataset does not have an influence on the predictions, the same testing dataset was used for both analyses.

Analyzing the data

Testing and training dataset

In order to compare LR, RF and BCT in terms of how well they can predict recidivism, the dataset was randomly divided: 66% of the data was used as the training dataset and 34% of the data was used as the testing dataset. The training and testing datasets were the same for all classification methods in order to reduce randomness. The imbalanced training dataset contained 426 YO. The balanced dataset contained 382 YO.

Model tuning

LR, RF and BCT all make use of different hyperparameters, which are explained in appendix C. Since model performance depends on the chosen values for the hyperparameters (Huang & Boutros, 2016; Tantithamthavorn et al., 2016), all models were tuned. Two approaches of model tuning were applied. First, the 20 default values calculated by the R-package “caret” were used, since its default values generally work well (Kuhn et al., 2020). Second, a manual grid search was carried out. All hyperparameters were compared using fivefold cross validation on the training dataset, repeated three times. The optimal hyperparameters were defined to be those that maximize the average area under the Receiver Operating Characteristic Curve. These analyses were carried out using the R-package “caret” (Kuhn et al., 2020). All hyperparameters that were tried out and the values that were chosen can be found in the appendix C.

Testing the predictions

Which method makes the best predictions depends on the performance measure used (Flach & Kull, 2015). Accuracy is an overall performance measure, taking into account true positives as well as true negatives, but neither false positives nor false negatives (see table 3). For that reason, it is often a misleading evaluator (Jordaney et al., 2016). Instead, precision, recall and F1 were used for model evaluation in this paper. Precision is the proportion of predicted positives that were observed as positives. Recall is the proportion of observed positives that were correctly predicted as positives (= sensitivity). F1 combines precision and recall and is a popular method in model testing. These performance measures show different results depending on which class is the one of interest (which is the positive class, see table 3). Therefore, performance measures were calculated both for recidivism and for non-recidivism as the positive class.

In order to compare the performance measures for LR, RF and BCT, a data frame was created for each of the three methods containing the pairs of observed and predicted values. Subsequently, 1000 bootstrapped[23] samples were drawn from this data frame (Hothorn et al., 2005), that were of the same size as the original testing dataset[24]. Based on these samples, 95% confidence intervals were calculated.

Table 3

Confusion matrix

	observed positive	observed negative
predicted positive	A	B
predicted negative	C	D

accuracy = (A+D) / (A+B+C+D)

precision = A / (A+B)

recall = A / (A+C)

F1 = 2 * precision * recall / (precision + recall)

Additionally, the Area Under the Receiver Operating Characteristic curve (AUC-ROC) was used in order to compare the performance of the methods. An AUC-ROC is a comparison of the true positive rate against the false positive rate (Powers, 2011). A perfect classifier has an AUC-ROC value of 1 (100% true positive and 0% false positive). When classes are equally balanced, a random classifier has an AUC-ROC value of 0.5 (50% true positive and 50% false positive). In order to compare the AUC-ROC values, 16 samples of the data frame of observed and predicted values were bootstrapped (the samples were again of the same size as the training dataset). Subsequently, standard deviations were calculated and shown in a boxplot.

[23] Bootstrapping means repeatedly sampling observations from the original dataset randomly with replacement.

[24] All bootstrapped samples used for evaluation were of the same size than the original testing dataset.

Variable importance

In order to find predictors of recidivism, variable importances needed to be calculated for the three models. For LR, the absolute values of the model coefficients were used as variable importances. As explained above, even if a variable has more categories one coefficient in LR only compares two of them. Consequently, variable importances for variables with more than two categories in LR also compare only two categories. Since LR in this paper was used as a penalized model, interpretation of the actual values of the coefficients was not meaningful (Goeman, 2010). However, variable importances can be assessed by comparing these coefficients. The predictor variables were standardized in order to make the variable importances scale independent. They were calculated using the R-package “caret” (Kuhn et al., 2020).

For RF and BCT, one standard procedure is to average the mean decrease in Gini impurity over all trees for every predictor (Oh et al., 2003). However, this approach has been found to result in variable importances that are biased towards collinear variables, continuous variables and categorical variables with a high number of categories (Strobl et al., 2008). Independent variables are collinear if they are linearly related with each other (D’Ambra & Sarnacchiaro, 2010). While collinearity does not affect model fit or prediction for LR, RF or BCT, it can become an issue with variable selection (O’Brien, 2017). Instead, for RF a method that calculates variable importance and that can cope with collinear variables is **conditional permutation** (Strobl et al., 2008). In this approach, permutation stands for randomly shuffling the responses of the predictor variable X_j, whereby the original relationship between the two variables is broken. Prediction accuracy of the model is then calculated twice: Once including all original predictor variables, and once including all predictor variables but instead of the original variable X_j including the permuted variable X_j. The variable importance then is the difference between these two model accuracies averaged over all trees (Strobl et al., 2009). This procedure is carried out for every variable. However, this approach yet leads to higher variable importances for variables that are collinear. Here, the conditional part of conditional permutation comes in: The predictor variable is only shuffled within the same splits on the tree. Thereby, the correlational structures between X_j and the predictor variables are kept (Strobl et al., 2009). Conditional permutation importance was calculated using the R-package “party” (Hothorn et al., 2020).

For BCT, conditional permutation was not yet available when this paper was written. Therefore, permutation importance was calculated without keeping the original correlational structures. In order to allow a discussion on variable importance, variance inflation factors were calculated for the important variables selected by BCT. Variance inflation factors are a calculation of the degree to which an independent variable is collinear with the other independent variables. A cut-off of ten was applied (O'Brien, 2007).

As it is the case for penalized LR, the absolute values of both permutation and conditional permutation are not interpretable either (Strobl et al., 2008). Therefore, the cut-off values for the variable importances were determined based on the variable importance plot: Those variables that had a substantially higher variable importance than the other variables were selected (Smith et al., 2015).

Partial dependence plots

In order to understand the relationship between the predictors and recidivism, partial dependence plots were created for the most important variables of each method. A partial dependence plot visualizes the relationship between a predictor and the outcome, marginalized over the other predictors (Greenwell, 2017). For instance, in order to create a partial dependence plot for a binary predictor variable B containing the values "B1" and "B2", the following process is carried out (Friedman, 2001). First, the value "B1" is inserted in the dataset for every observation of the variable B. Second, the trained model is used to make predictions on the whole dataset, including the modified B variable. Third, the average prediction over all observations is plotted as the partial dependence for the value "B1". Subsequently, the same is done for "B2": The value "B2" is inserted for all observations of B, predictions are made on the modified dataset, and the average prediction for that dataset is plotted as partial dependence for the value "B2". For variables with more than two categories, the same process is followed for every category. For continuous variables, a grid is determined within which the predictions are made.

Descriptive statistics

The mean age of the YO on the release day in the training dataset was 21.67 years (standard deviation (SD) = 1.91). On average, they spent 450.00 days in the YOI (SD = 224.16). 388 of the YO were born in Germany and 38 of the YO were born in another country. Table 4 shows the sorts of crimes committed by the YO from the training dataset. The recidivism rate within two years in this sample was 0.55.

Table 4

Crimes committed by the YO before they went to the YOI (training dataset)

violation of	number
chapter 17 GCC (offences against physical integrity)	179
chapter 18 GCC (offences against personal liberty)	50
chapter 19 GCC (theft and misappropriation)	240
chapter 20 GCC (robbery and extortion)	108
chapter 22 GCC (fraud and embezzlement)	96
Road Traffic Regulations	51
Narcotics Act	60

Note. Some YO have committed crimes of multiple crime categories, therefore the total number of crimes committed exceeds the total number of YO included in the training dataset (N = 426). GCC: German Criminal Code.

Results

Evaluation of the models

Imbalanced dataset

With regard to the prediction of recidivism, the mean prediction performances of LR and BCT yielded slightly higher results than that of RF (see table 5). Concerning the prediction of non-recidivism, the mean prediction performance of BCT yielded slightly higher results than that of LR and RF. Taking into account the 95% confidence intervals, model differences were not significant. This is also shown by the overlapping AUC-ROC values for the prediction of recidivism of the models (see figure 1). This was not according to the expectations.

The F1-values for the prediction of recidivism were significantly higher than the base rate of 0.55 (see table 5). The F1-values for the prediction of non-recidivism did not significantly differ from the base rate of 0.45. All models had AUC-ROC values below 0.7, which is considered poor in the scientific community (Mandrekar, 2010).

Table 5

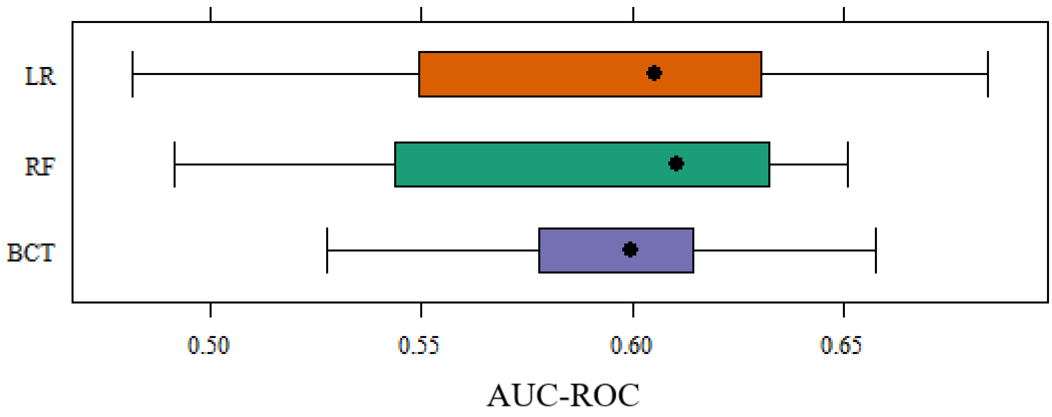
Model evaluation on the imbalanced dataset

	recidivism			non-recidivism		
	precision	recall	F1	precision	recall	F1
LR	0.60 ± 0.08	0.81 ± 0.07	0.69 ± 0.06	0.57 ± 0.13	0.32 ± 0.09	0.41 ± 0.10
RF	0.60 ± 0.08	0.77 ± 0.08	0.67 ± 0.06	0.56 ± 0.13	0.36 ± 0.10	0.44 ± 0.10
BCT	0.61 ± 0.08	0.79 ± 0.07	0.69 ± 0.06	0.59 ± 0.12	0.37 ± 0.10	0.46 ± 0.10

Note. The 95% confidence intervals were calculated by bootstrapping pairs of observed and predicted values 1000 times (see method section).

Figure 1

Comparison of the AUC-ROC of the different methods (bootstrapped imbalanced dataset)



Note. The standard deviation was calculated by bootstrapping pairs of observed and predicted values 16 times (see method section).

For the prediction of recidivism, recall was higher than precision for all models. For the prediction of non-recidivism, precision was higher than recall for all models. Thus, the models predicted the recidivism class more often than that they predicted the non-recidivism class. An examination of the predictions of the models showed that 73% of YO were predicted to recidivate with LR. Thus, compared to the recidivism rate of 55% in this sample, 18% too many YO were predicted to belong to the recidivism class. With RF and BCT, 68% and 66% of YO were predicted to belong to the recidivism class respectively. Thus, the models were biased towards predicting the majority class. Hereinafter, the results of the balanced dataset are addressed.

Balanced dataset

With regard to the prediction of recidivism, the mean prediction performances of LR and BCT were slightly higher than that of RF on the balanced dataset (see table 6). Regarding the prediction of non-recidivism, the mean prediction performance of RF was slightly higher than that of LR and BCT. Taking into account the 95% confidence intervals, model differences were not significant. The AUC-ROC values for the prediction of recidivism also show non-significant model differences (see figure 2). This was not according to the expectations. Additionally, the 95% confidence intervals of all F1-values included 0.5, which was the base rate for the balanced dataset. Thus, on the balanced dataset the models did not perform significantly better than the chance level.

Table 6

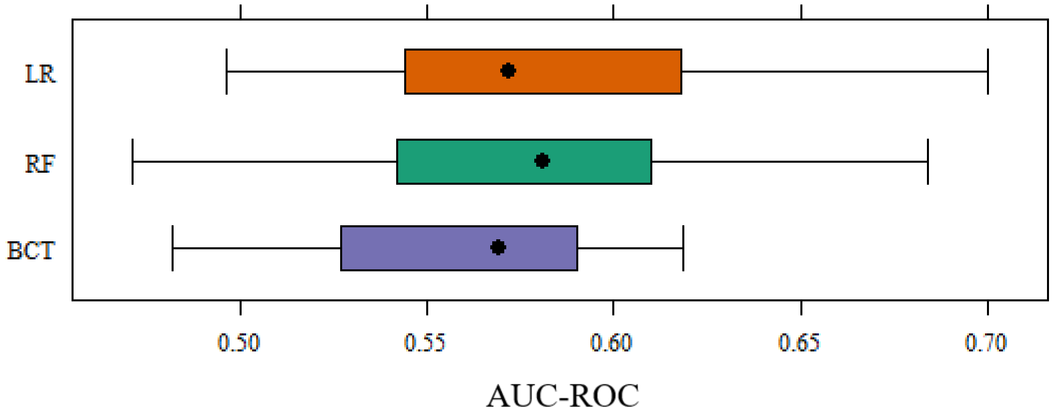
Model evaluation on the balanced dataset

	recidivism			non-recidivism		
	precision	recall	F1	precision	recall	F1
LR	0.62 ± 0.10	0.53 ± 0.10	0.57 ± 0.08	0.50 ± 0.10	0.59 ± 0.10	0.54 ± 0.08
RF	0.60 ± 0.10	0.43 ± 0.08	0.50 ± 0.08	0.48 ± 0.09	0.65 ± 0.10	0.55 ± 0.08
BCT	0.60 ± 0.09	0.54 ± 0.09	0.57 ± 0.08	0.50 ± 0.09	0.56 ± 0.10	0.52 ± 0.08

Note. The 95% confidence intervals are calculated by bootstrapping pairs of observed and predicted values 1000 times (see method section).

Figure 2

Comparison of the AUC-ROC of the different methods (bootstrapped balanced dataset)



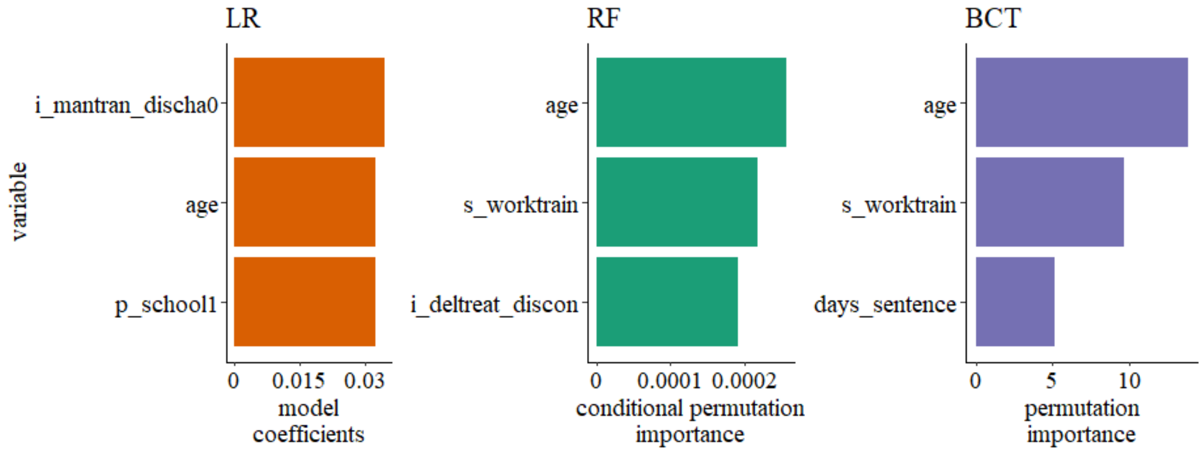
Note. The standard deviation was calculated by bootstrapping pairs of observed and predicted values 16 times (see method section).

Variables relevant for the prediction of recidivism

Below, the three most important variables selected by each method on the imbalanced dataset are discussed. The most important variables were determined by selecting those that had a substantially higher variable importance than the other variables, which for all models were three variables (see appendix D). The proportions of YO who recidivated and YO who did not recidivate on the training dataset are shown in table 7. A description of all variables is found in appendix F.

Figure 3

The three most important variables for all models on the imbalanced dataset

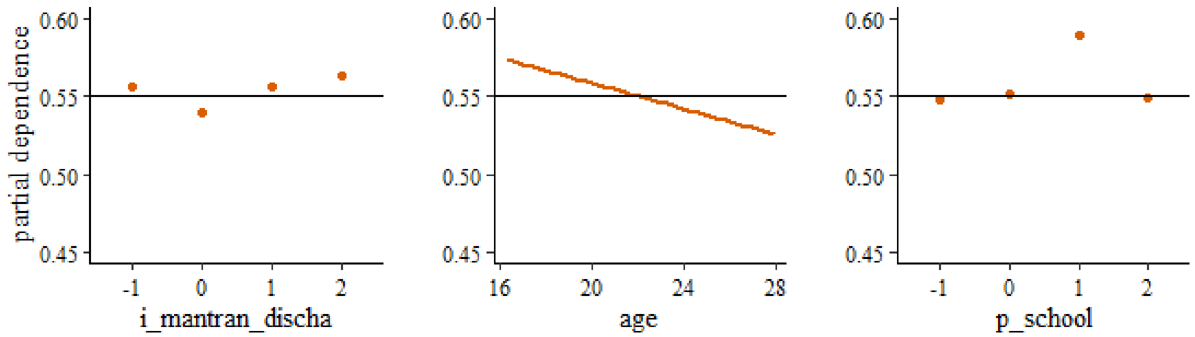


Logistic Regression

The most important variable for LR was i_mantran_discha0 (see figure 3). This variable measured whether a YO was in need to continue structured transition management after release. As explained above, the importance of LR compares only two categories against each other. The 0 behind the variable i_mantran_discha tells that category “no need” (0) was compared against the category “not relevant or known” (-1), which was the reference category. LR more often predicted recidivism for YO for whom the information was “not relevant or known” than for YO who had “no need” in structured transition management (see figure 4). The second most important variable for LR was age: Older age was negatively related with recidivism. The third most important variable for LR was p_school1. P_school was an estimation by a professional regarding the question whether the YO had a place in school after release. LR more often predicted recidivism for those YO who “probably” had a place in school (1) than for those for whom the information was “not relevant or known” (-1; reference category).

Figure 4

Partial dependences for the prediction of recidivism for the three most important variables selected by LR



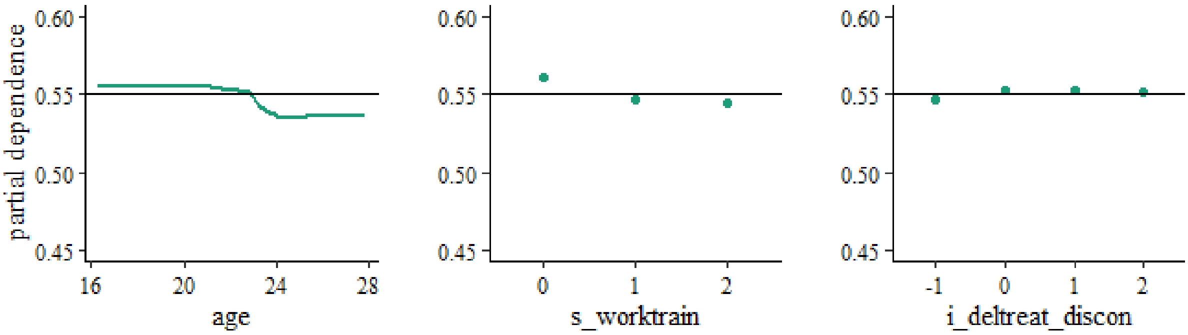
Note. The partial dependences are plotted on the probability scale. Thus, the y-axis represents the probability that the recidivism class is predicted. The base rate for recidivism is 0.55.

Random Forest

The most important variable in RF was age. RF more often predicted recidivism for YO who were younger than 23 than for YO who were older than 24 (see figure 5). The second most important variable was s_worktrain, which was a self-report variable for the question whether the YO had work or vocational training after release. RF predicted recidivism more often for those YO who stated to "not" (0) have work or vocational training after release than for those who stated to have work or vocational training "probably" (1) or "for sure" (2). Third, i_deltreat_discon was selected. RF predicted recidivism more often for YO who "did" (1, 2) or "did not" (0) discontinue other delict- or problem specific measures. RF less often predicted recidivism for those for whom the information was "not relevant or known" (-1).

Figure 5

Partial dependences for the prediction of recidivism for the three most important variables selected by RF



Note. The partial dependences are plotted on the probability scale. Thus, the y-axis represents the probability that the recidivism class is predicted. The base rate for recidivism is 0.55. The small deviances between the line/ points and the base rate are partly due to the small mtry hyperparameter selected for RF. At each node, the next variable is only selected out of 2 randomly chosen variables. Therefore, even the most important variables were not included in every tree and differences between variable importances are smaller (the finding that the variable importances in RF are more similar to each other than the variable importances in BCT can also be observed in appendix D).

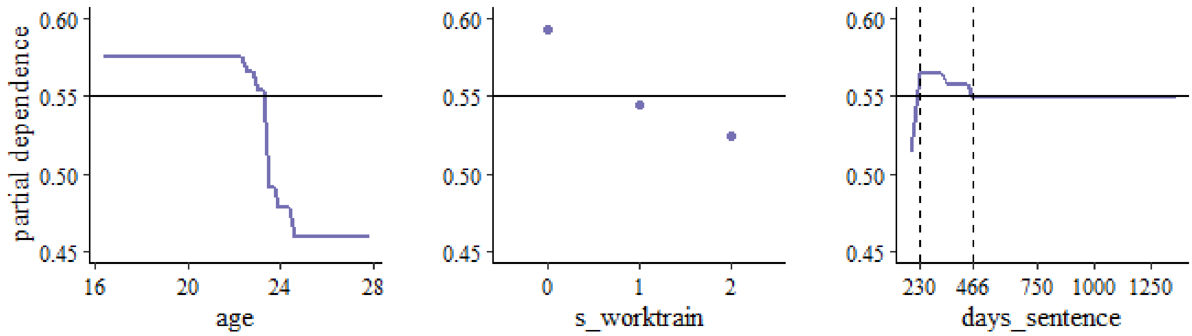
Boosted Classification Trees

For BCT, as explained earlier, the variable selection method is known to be biased towards multicollinear variables. Therefore, the variance inflation factor has to be taken into account for these variables. The variance inflation factor for age, s_worktrain1, s_worktrain2 and days_sentence were 4.16, 4.69, 6.78 and 6.04 respectively. Thus, all values were below the cut-off of 10 and therefore the variables were interpretable (O'Brien, 2007).

For BCT, the most important variable was age. As in RF, recidivism was more often predicted for those YO who were younger than 23 than for those who were older than 24 (see figure 6). The second most important variable was s_worktrain. As in RF, recidivism was more often predicted for those who stated to "not" (0) have work or vocational training after release than for those who stated to have it "probably" (1) or "for sure" (2). The third most important variable was days_sentence. YO who spent between 230 and 466 days in the YOI were more often predicted to belong to the recidivism class than YO who spent a shorter or a longer time in the YOI. The majority of YO spent this range of days in the YOI (see figure 7).

Figure 6

Partial dependences for the prediction of recidivism for the three most important variables selected by BCT



Note. The partial dependences are plotted on the probability scale. Thus, the y-axis represents the probability that the recidivism class is predicted. The base rate for recidivism is 0.55.

Figure 7

Density plot of the variable days_sentence (observed)

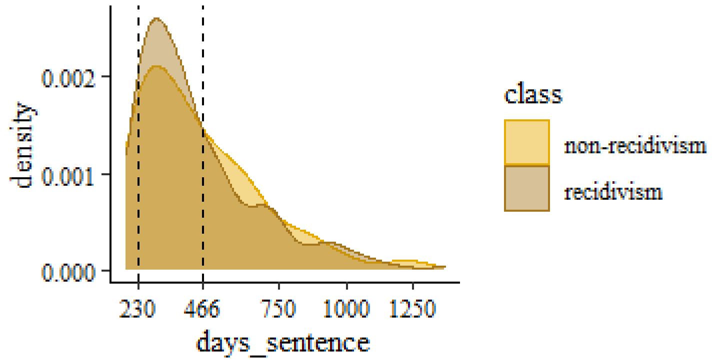


Table 7

Proportions of YO who recidivated and who did not recidivate (the training dataset)

variable	non-recidivism	recidivism	total
i_mantran_discha	PA: after release, planned continuation of structured transition management		
-1 = not relevant or known	18 (9.4%)	30 (12.8%)	48 (11.3%)
0 = not necessary	96 (50.3%)	75 (31.9%)	171 (40.1%)
1 = necessary but not planned	35 (18.3%)	56 (23.8%)	91 (21.4%)
2 = necessary and planned	42 (22.0%)	74 (31.5%)	116 (27.2%)
age	age on day of release in years		
mean (SD)	22.003 (1.978)	21.397 (1.815)	21.669 (1.911)
p_school	PA: after release, has a place in school		
-1 = not relevant or known	95 (49.7%)	81 (34.5%)	176 (41.3%)
0 = no	89 (46.6%)	130 (55.3%)	219 (51.4%)
1 = probably	2 (1.0%)	14 (6.0%)	16 (3.8%)
2 = yes	5 (2.6%)	10 (4.3%)	15 (3.5%)
s_worktrain	SE: after release, has work or vocational training		
0 = no	42 (22.0%)	93 (39.6%)	135 (31.7%)
1 = probably	85 (44.5%)	89 (37.9%)	174 (40.8%)
2 = yes	64 (33.5%)	53 (22.6%)	117 (27.5%)
days_sentence	number of days spent in young offender institution		
mean (SD)	458.084 (229.308)	443.434 (220.163)	450.002 (224.162)
i_deltreat_discon	PA: discontinued other delict- or problem specific measures		
-1 = not relevant or known	76 (39.8%)	74 (31.5%)	150 (35.2%)
0 = no	105 (55.0%)	138 (58.7%)	243 (57.0%)
1 = yes due to YOI	4 (2.1%)	12 (5.1%)	16 (3.8%)
2 = yes due to YO (not willing or able)	6 (3.1%)	11 (4.7%)	17 (4.0%)

Note: N = 191 in the non-recidivism class, N = 235 in the recidivism class, N = 426 in total. PA: professional assessment, SE: self-estimated.

Discussion

Model comparison

The first goal of this paper was to compare LR, RF and BCT in terms of how well they can predict recidivism. The results showed that method performances did not differ from each other. Therefore, the hypotheses were not confirmed: RF did not yield substantially better predictions than LR, and BCT did not yield better predictions than RF. This finding is not in concordance with earlier research showing that RF often outperforms LR in terms of prediction (Couronné et al., 2018). Neither is this finding in concordance with earlier research showing that BCT often outperforms LR (Roe et al., 2006). Instead, the results seem to support research stating that modern machine learning methods do not show an advantage over LR when predicting recidivism (Tollenaar & van der Heiden, 2013). However, the fact that all models performed poorly suggests that the data contained little information about the outcome criterion. Therefore, a reconsideration of the study is needed before any hard conclusions can be drawn.

Poor model performance

One possible reason for poor model performances could be an imbalanced dataset. If the models are biased towards predicting the majority class, prediction of the minority class becomes poor (Japkowicz, 2000). The results in the current study showed that on the imbalanced dataset, all models were biased towards predicting the majority class. On the balanced dataset, predictions of all models improved for the non-recidivism class, thereby reducing prediction performance for the recidivism class. Overall model performances were poor on both datasets. Therefore, the possibility that low model performance was due to the imbalanced dataset can be ruled out.

Another possible reason for poor model performances is that important predictors for recidivism were lacking in the dataset. One indication for this is that the most important variables were quite different for the three methods: If some variables had been much more important than others, it would have been expected that they would have been chosen by all models (see appendix D). When looking back at table 1, out of the risk factors that are widely accepted in the scientific community only one was included in the current dataset (age, which turned out to be the only variable selected as important by all methods). Therefore, the current dataset probably lacked information that is important for the prediction of recidivism in YO.

Furthermore, it is probable that the YO included in this study were highly heterogenous. The idea that there are different types of YO is widely accepted in the scientific community. For instance, Moffitt (1993) suggested that there are two types of YO: Those who already show antisocial behavior from childhood on and persist in displaying antisocial behavior over the course of their lives, and those who only show antisocial behavior during their adolescence. Her research has been replicated in numerous studies, and differences in risk factors for offending and recidivism have been found between these two groups (Kjelsberg, 1999; Lussier et al., 2012; Moffitt, 2003, 2018; Moore et al., 2017). The same applies to the difference between YO who committed sexual and YO who committed violent offences: Vast differences have been found between these groups with regard to characteristics and recidivism risk factors (Caldwell, 2007; Fox, 2017; Långström, 2002; McCann & Lussier, 2008; Mulder et al., 2012; Olver et al., 2009). Thus, another reason for poor model performance in this study probably was that different groups of YO were included who had various risk factors for recidivism.

Last, predictive performance is limited on noisy datasets (Haughton et al., 2016), and there are a number of reasons why it is probable that the current dataset contained noise. First, no scoring criteria were provided for the Social Workers Release Questionnaire. It is likely that the different professionals who filled in the questionnaire had different ideas about scoring the items, for example regarding the question whether a YO deals with his deed seriously. Second, the Intervention Questionnaire is collected within multiple federal states of Germany. Since different YOI offer different types of interventions, it is not always clearly defined which intervention belongs to which intervention category. Thus, one intervention might have been assigned to different intervention categories depending on the professional who filled in the questionnaire. For this questionnaire, it could also be that information was lost when the variables were recoded: It is possible that the difference between “not relevant” and “not known” would in fact have been predictive of recidivism. Third, multiple issues emerged regarding the Release Questionnaire. Data imputation is likely to have brought noise to the data, which is supported by the relatively small accuracies and correlations with available testing data (see appendix B). Furthermore, the validity of self-report data has often been discussed in the field of psychology (del Boca & Noll, 2000; Woolley et al., 2004). Amongst other things, it depends on the degree of introspection, social desirability and motivation of the respondent (Mills & Kroner, 2006; Mook & Scott, 2001), which might vary in the current sample. For instance, a YO who has consumed drugs during his stay in YOI potentially has a higher chance of recidivism than a YO who has not consumed drugs during his stay (Aebi et al., 2020; Huebner et al., 2007). However, at the same time it is possible that a YO who admits that he has consumed drugs has a lower chance of recidivism than a YO who does not admit that he has consumed drugs (Aleixo & Norris, 2000). Thus, regarding the self-estimated variables, there are contradicting effects that can have relativized each other. Finally, one source of error lies in the outcome criterion. The recidivism class does not include all YO who recidivated but only those who recidivated and were detected. Therefore it is possible that recidivism was predicted correctly in some cases, but that the outcome criterion was wrongly observed (Heinz, 2004). In summary, it is assumable that the poor prediction performance of the methods was partly due to limited data quality.

Differences between the models

The models showed no difference in prediction performance, but they differed in terms of what variables they often selected for prediction. This is due to the fact that the way a method works has an effect on how well a variable predicts the outcome. The largest differences were found between LR versus RF and BCT.

The question that now emerges is which method is the most useful in praxis. The first advantage of LR is that main effects can be calculated, while in RF and BCT main effects and interaction effects are automatically combined. Even though there is never only one factor present in the real world, main effects are well-interpretable and can contribute to the understanding of a variable. Furthermore, the scientific community has agreed on a cut-off value to determine whether a difference is significant (even though this agreement is arbitrary; James et al., 2013). RF and BCT have the advantages that the distribution of the predictors is less relevant for their meaningfulness and that they can fit the data more precisely than LR (Stoltzfus, 2011). For instance, RF and BCT could fit the distribution of the variable age better than LR in the current study (see appendix E). Moreover, they do not only include interactions between two variables but also higher order interactions which can yield new insights. For RF and BCT, no dummy variables need to be calculated and interpretation of variable importance does not depend on the reference category (G. James et al., 2013). Another difference between the methods was that RF and BCT seemed to cope with imbalance better, because LR was more strongly biased towards predicting the majority class. Earlier research has indeed proposed that RF and BCT show better performances on imbalanced datasets that LR (Brown & Mues, 2012). If this idea is confirmed in future research, this would add to the advantages of tree-based models over LR when analyzing recidivism data.

In summary, the predictive performance of the methods did not differ in the current study, which was possibly due to characteristics of the dataset. Nonetheless, the methods had different properties that might be more or less advantageous depending on the research question at hand. Thus, prediction performance is not the only factor that determines methods suitability. Researchers should be aware of the different properties of the methods and depending on the situation choose the one that fits their goal best.

Risk factors of recidivism

The second goal of this paper was to evaluate the variables included in this study in terms of relevance for the prediction of recidivism. Before interpreting the variables, it should be noted that the predictors are affected by the limitations of the dataset discussed above. Furthermore, the variables were only predictive of recidivism when the methods were biased towards predicting the majority class. Therefore, the risk factors that emerged in this study should be interpreted as indications rather than firm findings.

Static risk factors

Overall, age was the most important risk factor for recidivism, being the most important predictor in RF and BCT, and the second most important predictor in LR. The decreased offence risk of an adolecent developing into adulthood is widely accepted in the scientific community (Moffitt, 2018; Sweeten et al., 2013). Furthermore, it is well-explainable by biological, cognitive and social factors, as for instance a higher decision making ability, an increased impulse control and a decreased exposure to antisocial peers (Chen et al., 2015; K. Monahan et al., 2015; Steinberg & Scott, 2003). Thus, young age emerged as a risk factor for recidivism, which is well supported by previous literature.

A shorter time spent in the YOI was the second most important static risk factor in this study. In BCT, it was the third most important variable and in LR it was selected as important as well (see appendix D). Earlier research has found no relationship between the amount of time spent in the YOI and recidivism (Loughran et al., 2009; Walker & Bishop, 2016; Winokur et al., 2008). The reason that this variable emerged in this study possibly is that more than half of the YO included in this study had committed a theft or misappropriation (see table 4), for which YO sentences are comparably low on average, while the recidivism rate is comparably high (Jehle & Heinz, 2003; Naplava, 2012). This would also explain why this variable was not included in RF, because the variable whether the YO had committed a theft was included in RF instead. Therefore, the crime category whether a YO committed a theft possibly was a confounding factor for the prediction of recidivism in this study.

Dynamic risk factors

The most important dynamic risk factor in this study was if the YO himself said he will not have work or vocational training after release. This risk factor was included as relevant in all methods and in RF and BCT, it was the second most relevant risk factor. Even though this risk factor is not yet well-established in the scientific literature, low commitment to work or school has been proposed to be predictive of recidivism in earlier research (Mulder et al., 2010). Furthermore, it has been proposed that specific plans after release increase the self-efficacy of YO, which might be negatively related to recidivism (Forste et al., 2011). Besides, going to work or school after YOI release is regarded as an elementary part of reintegration into society (Bouffard & Bergseth, 2008). Hence, the finding that a YO who says he has school or vocational training after release has a smaller chance to recidivate falls into line with earlier research.

The same applies to the elevated recidivism probability for YO who only “probably” had a place in school, and the slightly elevated recidivism probability for YO who had “no” place in school after release (see figure 4). This variable was the third most important variable in LR and emerged as important in RF and BCT as well. The finding that recidivism was more likely for YO who “probably” had a place in school than for those who had “no” place in school cannot be explained by earlier research and possibly occurred due to the small sample size in the “probably” category[25]. Thus, having no or only probably a place in school was a risk factor for recidivism in this study[26].

---

[25] In the training dataset, only 16 YO “probably” had a place in school after release (see table 7).

[26] Table 7 shows that on the training dataset, even those YO who “had” a place in school after release on average were more likely to recidivate than to not recidivate. The difference between table 7 and the partial dependence plot (see figure 4) is that in the table, the effect of the other variables is not included while in the partial dependence plot, it is. For example, it is likely that in table 7, those YO who “had” a place in school were younger on average than those YO who had “no” place in school (since older YO are more likely to go to work than to go to school). As just explained, younger YO were more likely to recidivate than older YO. On the partial dependence plot, the effect of age is averaged out (see section “partial dependence plots”). This is a probable reason for the difference between the table and the partial dependence plot: When the effect of age was ignored (table), going to school emerged as a risk factor for recidivism. However, when the effect of age was averaged out (partial dependence plot), going to school was no risk factor for recidivism.

With regard to the intervention variables, two dynamic risk factors for recidivism emerged. The first was if the YO "had a need" to continue structured transition management after release or if it was "not relevant or known" whether he had that need. This risk factor was the most important in LR and was also selected by RF and BCT. Structured transition management in the YOI of Regis-Breitingen entails a reintegration planning that exceeds common sentence planning (Hartenstein, 2014). It can for example include arrangements to get communal help and/ or to organize a place in school and/ or work. The risk factor of a need to continue structured transition management after release is in accordance with literature stating that well-implemented transition management reduces recidivism (C. James et al., 2013). Furthermore, it is in accordance with literature stating that an unmet need of a YO to participate in an intervention is associated with recidivism (Luong & Wormith, 2011). "Not relevant or known" in this context can mean three things: Either the professional could not estimate whether a need is present, the professional did not know whether a continuation after release was planned, or the professional did not fill in this question. Arguably, this is a diverse group, which should not be interpreted before more precise information is gained. Furthermore, table 7 shows that group differences were smallest for the not relevant or known criterion, thus the methods possibly relied on the other categories more strongly. In summary, another risk factor for recidivism that emerged in this study was if the YO had a need to continue structured transition management after release, which is supported by previous literature.

The second risk factor concerning the interventions was if the YO "did" or "did not" discontinue other delict- or problem specific measures. This risk factor was the third most important in RF. In LR it was included, in BCT it was not included. In the YOI of Regis-Breitingen, other delict- or problem specific measures entail for YO to deal with the deeds committed and their consequences (Hartenstein, 2014). A training in moral reasoning and a feeling of guilt have been proposed to be associated with lower recidivism rates (Gibbs et al., 1996; van Vugt et al., 2011). A lack of treatment adherence has been proposed to increase recidivism risk (Mulder et al., 2011). However, due to the finding that those who did not drop out of the treatment were more likely to recidivate as well, this risk factor in fact proposes that those who did participate in delict- or problem specific measures were more likely to recidivate than those who did not. Accordingly, all methods included the variable whether this intervention was begun as important. This does not mean that participation in the intervention increases the recidivism risk, because YO were not selected to participate at random. Thus, there was a selection bias (Mook & Scott, 2001). It is possible that participation in delict- or problem specific measures reduces the risk of recidivism, but that those YO who are likely to recidivate are selected to participate in the intervention. Consequently, a YO who participated in delict- or problem specific measures had a higher recidivism risk than a YO who did not participate. Therefore, the current study cannot be interpreted in terms of intervention effectiveness. In summary, participation in delict- or problem specific measures was a risk factor in this study, possibly due to selection bias.

As already explained, the risk factors discussed above are to be interpreted as ideas rather than firm research findings due to poor model performances. In summary, the most important static risk factors for recidivism were a younger age and less time spent in the YOI. The most important dynamic risk factors were having no place in work, vocational training or school after release, having a need to continue structured transition management after release and having participated in delict- or problem specific measures.

General discussion

This was one of the first studies to examine machine learning methods on a recidivism dataset. Thereby, this study contributes to closing the gap between methodology and research (Sharpe, 2013). While the field of methodology has experienced vast developments in the past decades, psychological research has barely profited from these (Borsboom, 2006; Jungmann et al., 2015). The use of tree-based models for the prediction of recidivism in YO contributes to bringing innovation to recidivism data analysis.

Furthermore, this study contributes to the literature stating that prediction of recidivism is difficult (Cooke & Michie, 2010). One reason for this is a well-known problem even beyond the field of psychology: Individuals differ in their cognitions, emotions and behavior, and therefore differences found between groups are not necessarily applicable to an individual (Borsboom et al., 2003). Human beings have the ability of both equifinality and multifinality (Richters, 1997). Equifinality refers to the capacity to reach the same outcomes from a different starting point and/ or development. Multifinality refers to the capacity to reach different outcomes from a similar starting point and/ or development. Thus, at the moment a YO is released, it cannot be known for sure how the YO will adjust to and perform in home and school. These factors have been proposed to be important risk factors for recidivism (Heilbrun et al., 2000), which again does not mean that they predict the same outcome for different individuals. This emphasizes the importance of supporting and mentoring the YO after release, thereby noticing the development of their criminogenic needs (Barry, 2000; Cunneen & Luke, 2007).

The strengths of this study include the prospective dataset collected over more than six years. A high number of variables was observed, including both professional estimations as well as self-estimations. Furthermore, all YO were in the same YOI, which reduces heterogeneity of the data. Regarding the methodological decisions, strengths of this study include that prediction performance of the non-recidivism class was taken into consideration and that the model performance was not tested on the same entries as it was trained on. This neat and necessary practice reduces the prediction performance of models, which might be a reason why it is not always followed (J. Monahan et al., 2000).

The main limitation of this study refers to the fact that recidivism was not well predictable from the dataset. As discussed, this was probably partly due to a lack of important predictors in the dataset, heterogenous groups of YO and noisiness of some variables. Consequently, the model predictions were too poor in order to draw any firm conclusions. Additionally, results of this study are not well generalizable to other contexts, since only male YO from the same YOI have been included.

For future research, one recommendation is to include the discussed risk factors for recidivism that have been widely accepted in the scientific community. This most probably would improve model performance and therefore allow for a better analysis of model differences. Furthermore, it is advisable to carry out the research for different types of YO separately, even though a large dataset would be needed in that case. Moreover, it would be informative to gather pre-measures of interventions in order to make an evaluation possible. Last, it would be interesting to use self-reported recidivism as the outcome variable in order to address the issue that not all crimes are detected (van Vugt et al., 2011).

Knowledge concerning the risk factors for recidivism in YO is crucial in order to help professionals make informed statements about interventions and recidivism risk for a YO. In order to gain that knowledge, a careful consideration of what statistical method best fits the goals of the research at hand can potentially yield new insights. However, predicting recidivism in YO remains a challenge.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aebi, M., Bessler, C., & Steinhausen, H. C. (2020). A Cumulative Substance Use Score as a Novel Measure to Predict Risk of Criminal Recidivism in Forensic Juvenile Male Outpatients. *Child Psychiatry and Human Development*. https://doi.org/10.1007/s10578-020-00986-7

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying Support Vector Machines to Imbalanced Datasets. In *European Conference on Machine Learning* (pp. 39–50). Springer Verlag.

Aleixo, P. A., & Norris, C. E. (2000). Personality and moral reasoning in young offenders. *Personality and Individual Differences*, *28*(3), 609–623. https://doi.org/10.1016/S0191-8869(99)00124-5

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, *9*(5), 272–278.

Alonso, J. M., Castiello, C., & Mencar, C. (2015). Interpretability of fuzzy systems: Current research trends and prospects. *Springer Handbook of Computational Intelligence*, 219–237. https://doi.org/10.1007/978-3-662-43505-2_14

Andersson, R. (2017). *Classification of Video Traffic An Evaluation of Video Traffic Classification using Random Forests and Gradient Boosted Trees*. Karlstads Universiteit.

Barry, M. (2000). The Mentor / Monitor Debate in Criminal Justice: ‘ What Works ’ for Offenders. *British Journal of Social Work, 30,* 575–595. https://doi.org/doi.org/10.1093/bjsw/30.5.575

Benda, B. B., Corwyn, R. F., & Toombs, N. J. (2001). Recidivism among adolescent serious offenders prediction of entry into the correctional system for adults. *Criminal Justice and Behavior*, *28*(5), 588–613.

Borsboom, D. (2006). The Attack of the Psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Bouffard, J. A., & Bergseth, K. J. (2008). The impact of reentry services on juvenile offenders' recidivism. *Youth Violence and Juvenile Justice*, *6*(3), 295–318. https://doi.org/10.1177/1541204007313384

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brillante, L., Gaiotti, F., Lovat, L., Vincenzi, S., Giacosa, S., Torchio, F., Segade, S. R., Rolle, L., & Tomasi, D. (2015). Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical-mechanical characteristics in wine grapes. *Computers and Electronics in Agriculture*, *117*, 186–193. https://doi.org/10.1016/j.compag.2015.07.017

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Buskirk, T. D., & Kolenikov, S. (2015). Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification. *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach*, 1–17. http://surveyinsights.org/?p=5108

Caldwell, M. F. (2007). Sexual offense adjudication and sexual recidivism among juvenile offenders. *Sexual Abuse: Journal of Research and Treatment*, *19*(2), 107–113. https://doi.org/10.1177/107906320701900203

Chawla, N., Bowyer, K. W., Lawrence, O. H., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In *Data mining and knowledge discovery handbook* (pp. 853–867). Springer. https://doi.org/10.1007/0-387-25465-X

Chen, D., Drabick, D. A. G., & Burgers, D. E. (2015). A Developmental Perspective on Peer Rejection, Deviant Peer Affiliation, and Conduct Problems Among Youth. *Child Psychiatry and Human Development*, *46*(6), 823–838. https://doi.org/10.1007/s10578-014-0522-y

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior*, *34*(4), 259–274. https://doi.org/10.1007/s10979-009-9176-x

Cottle, C., Lee, R., & Heilbrun, K. (2001). The Prediction of Criminal Recidivism in Juveniles. *Criminal Justice and Behaviour*, *28*(3), 367–394. https://doi.org/https://doi.org/10.1177/0093854801028003005

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2264-5

Cunneen, C., & Luke, G. (2007). Criminal Justice Interventions: Juvenile Offenders and Post Release Support. *Current Issues in Criminal Justice*, *19*(2), 197–210. https://doi.org/10.1080/10345329.2007.12036426

D'Ambra, A., & Sarnacchiaro, P. (2010). Some data reduction methods to analyze the dependence with highly collinear variables: A simulation study. *Asian Journal of Mathematics and Statistics*, *3*(2), 68–81. https://doi.org/10.3923/ajms.2010.69.81

del Boca, F. K., & Noll, J. A. (2000). Truth or consequences: the validity of self-report data in health services research on addictions. *Addiction*, *95*(3), 347–360. https://doi.org/10.1080/09652140020004278

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Farrington, D. P., Ttofi, M. M., & Piquero, A. R. (2016). Risk, promotive, and protective factors in youth offending: Results from the Cambridge study in delinquent development. *Journal of Criminal Justice*, *45*(May), 63–70. https://doi.org/10.1016/j.jcrimjus.2016.02.014

Feng, Y., & Wang, S. (2017). A forecast for bicycle rental demand based on random forests and multiple linear regression. *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, 101–105. https://doi.org/10.1109/ICIS.2017.7959977

Flach, P. A., & Kull, M. (2015). Precision-Recall-Gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems* (pp. 838–846).

Forste, R., Clarke, L., & Bahr, S. (2011). Staying out of trouble: Intentions of young male offenders. *International Journal of Offender Therapy and Comparative Criminology*, *55*(3), 430–444. https://doi.org/10.1177/0306624X09359649

Fox, B. H. (2017). What Makes a Difference?: Evaluating the Key Distinctions and Predictors of Sexual and Non-Sexual Offending Among Male and Female Juvenile Offenders. *Journal of Criminal Psychology*, *7*(2), 134–150. https://doi.org/10.1108/JCP-12-2016-0047

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.2307/2699986

Gao, J., Liu, N., Lawley, M., & Hu, X. (2017). An Interpretable Classification Framework for Information Extraction from Online Healthcare Forums. *Journal of Healthcare Engineering*.

Gibbs, J. C., Potter, G. B., Barriga, A. Q., & Liau, A. K. (1996). Developing the helping skills and prosocial motivation of aggressive adolescents in peer group programs. *Aggression and Violent Behavior*, *1*(3), 283–305. https://doi.org/10.1016/1359-1789(95)00018-6

Gill, N., Hall, P., & Montgomery, K. (2020). A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. *Information*, *11*(3), 1–32. https://doi.org/10.3390/info11030137

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, *52*(1), 70–84. https://doi.org/10.1002/bimj.200900028

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *R Journal*, *9*(1), 421–436. https://doi.org/10.32614/rj-2017-016

Guo, L., Ma, Y., Cukic, B., & Singh, H. (2004). Robust prediction of fault-proneness by random forests. *Proceedings - International Symposium on Software Reliability Engineering, ISSRE*, 417–428. https://doi.org/10.1109/ISSRE.2004.35

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The Accuracy of Recidivism Risk Assessments for Sexual Offenders: A Meta-Analysis of 118 Prediction Studies. *Psychological Assessment*, *21*(1), 1–21. https://doi.org/10.1037/a0014421

Harder, A. T., Knorth, E. J., & Kalverboer, M. E. (2015). Risky or needy? Dynamic risk factors and delinquent behavior of adolescents in secure residential youth care. *International Journal of Offender Therapy and Comparative Criminology*, *59*(10), 1047–1065. https://doi.org/10.1177/0306624X14531036

Hartenstein, S. (2014). Behandlungsmaßnahmen: Bedarf und Versorgung. *Daten & Dialog*, *2*, 1–8. https://www.justiz.sachsen.de/kd/download/daten-dialog-02_2014-03_bedarf-und-versorgung.pdf

Hartenstein, S., Hinz, S., & Meischner-Al-Mousawi, M. (2016). Transparenz, Reflexion, Veränderung: Vom Beitrag der Evaluation des sächsischen Jugendstrafvollzugs für die Praxis. *Forensische Psychiatrie Und Psychotherapie - Werkstattschriften*, *23*(2).

Haughton, N., Abramowitz, G., Pitman, A. J., Or, D., Best, M. J., Johnson, H. R., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Santanello, J. A., … Vuichard, N. (2016). The plumbing of land surface models: Is poor performance a result of methodology or data quality? *Journal of Hydrometeorology*, *17*(6), 1705–1723. https://doi.org/10.1175/JHM-D-15-0171.1

Heilbrun, K., Brock, W., Waite, D., Lanier, A., Schmid, M., Witte, G., Keeney, M., Westendorf, M., Buinavert, L., & Shumate, M. (2000). Risk Factors for Juvenile Criminal Recidivism: The Postrelease Community Adjustment of Juvenile Offenders. *Criminal Justice and Behavior*, *27*(3), 275–291. https://doi.org/10.1177/0093854800027003001

Heinz, W. (2004). Die neue Rückfallstatistik – Legalbewährung junger Straftäter. *Infor*, *96*, 35–48.

Hong, J. S., Ryan, J. P., Chiu, Y. L., & Sabri, B. (2013). Re-arrest among juvenile justice-involved youth: An examination of the static and dynamic risk factors. *Residential Treatment for Children and Youth*, *30*(2), 131–148. https://doi.org/10.1080/0886571X.2013.785230

Hosmer, D. W., Taber, S., & Lemeshow, S. (1991). The importance of assessing the fit of logistic regression models: A case study. *American Journal of Public Health*, *81*(12), 1630–1635. https://doi.org/10.2105/AJPH.81.12.1630

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2020). *Package ‘party .’* R Package Version 1.3-4. https://doi.org/10.1198/106186006X133933>

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, *14*(3), 675–699. https://doi.org/10.1198/106186005X59630

Hsieh, C. H., Lu, R. H., Lee, N. H., Chiu, W. T., Hsu, M. H., & Li, Y. C. (2011). Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. *Surgery*, *149*(1), 87–93. https://doi.org/10.1016/j.surg.2010.03.023

Huang, B. F. F., & Boutros, P. C. (2016). The parameter sensitivity of random forests. *BMC*

Bioinformatics, *17*(1), 1–13. https://doi.org/10.1186/s12859-016-1228-x

Huebner, B. M., Varano, S. P., & Bynum, T. S. (2007). Gangs, Guns, and Drugs: Recidivism Among Serious, Young Offenders. *Criminology & Public Policy*, *6*(2), 187–221. https://doi.org/10.1111/j.1745-9133.2007.00429.x

Hutchinson, R. A., Liu, L. P., & Dietterich, T. G. (2011). Incorporating boosted regression trees into ecological latent variable models. *Proceedings of the National Conference on Artificial Intelligence*, *2*, 1343–1348.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1), 1–10. https://doi.org/10.1186/s12874-017-0442-1

James, C., Stams, G. J. J. M., Asscher, J., de Roo, A. K., & van der Laan, P. H. (2013). Aftercare programs for reducing recidivism among juvenile and young adult offenders: A meta-analytic review. *Clinical Psychology Review*, *33*(2), 263–274. https://doi.org/10.1016/j.cpr.2012.10.013

James, G., Witten, D., Trevor, H., & Tibshirani, R. (2013). An Introduction to Statistical Learning. In *Springer* (Vol. 102). https://doi.org/10.1016/j.peva.2007.06.006

Japkowicz, N. (2000). Learning from Imbalanced Data Sets: A Comparison of Various Strategies. *AAAI Technical Report WS-00-05*, 10–15. https://doi.org/10.1007/978-3-319-98074-4_11

Jehle, J. M., Albrecht, H. J., Hohmann-Fricke, S., & Tetal, C. (2016). *Legalbewährung nach strafrechtlichen Sanktionen: eine bundesweite Rückfalluntersuchung 2010 bis 2013 und 2004 bis 2013*. Forum Verlag.

Jehle, J. M., & Heinz, W. (2003). Legalbewährung nach strafrechtlichen Sanktionen Eine kommentierte Rückfallstatistik. In *Bundesministerium der Justiz*. Bundesministerium der Justiz. https://publikationen.uni-tuebingen.de/xmlui/bitstream/handle/10900/62657/Legalbwaehrung_nach_strafrechtlichen_Sanktionen_Rückfallstatistik.pdf?sequence=1&isAllowed=y

Jimerson, S. R., Sharkey, J. D., O'Brien, K. M., & Furlong, M. J. (2004). The Santa Barbara Assets and Risks Assessment to predict recidivism among male and female juveniles: An investigation of inter-rater reliability and predictive validity. *Education & Treatment of Children*, *27*(4), 353–365. http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2015-29392-

009&site=ehost-live%5Cnhttp://Jimerson@education.ucsb.edu

Jordaney, R., Wang, Z., Papini, D., Nouretdinov, I., & Cavallaro, L. (2016). *Misleading Metrics : On Evaluating Machine Learning for Malware with Confidence*.

Jungmann, R., Baur, N., & Ametowobla, D. (2015). Grasping processes of innovation empirically. A call for expanding the methodological toolkit. An introduction. *Historical Social Research*, *40*(3), 7–29. https://doi.org/10.12759/hsr.40.2015.3.7-29

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(1). https://doi.org/10.1186/1472-6947-11-51

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *Recommended Citation Kirasich*, *1*(3), 9. https://scholar.smu.edu/datasciencereview/vol1/iss3/9

Kjelsberg, E. (1999). Adolescence-limited versus life-course-persistent criminal behaviour in adolescent psychiatric inpatients. *European Child and Adolescent Psychiatry*, *8*(4), 276–282. https://doi.org/10.1007/s007870050102

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *Science*, *30*(1), 25–36. https://doi.org/10.1007/978-0-387-09823-4_45

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lesearbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2020). *Package 'caret.'* R Package Version 6.0-86. https://cran.r-project.org/web/packages/caret/caret.pdf

Långström, N. (2002). Long-term follow-up of criminal recidivism in young sex offenders: Temporal patterns and risk factors. *Psychology, Crime and Law*, *8*(1), 41–58. https://doi.org/10.1080/10683160208401808

Lawrence, R., Bunn, A., Powell, S., & Zambon, M. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, *90*(3), 331–336. https://doi.org/10.1016/j.rse.2004.01.007

Lee, D., Mulrow, J., Haboucha, C. J., Derrible, S., & Shiftan, Y. (2019). Attitudes on Autonomous Vehicle Adoption using Interpretable Gradient Boosting Machine. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(11), 865–878. https://doi.org/10.1177/0361198119857953

Liu, Y., Gu, Y., Nguyen, J. C., Li, H., Zhang, J., Gao, Y., & Huang, Y. (2017). Symptom severity classification with gradient tree boosting. *Journal of Biomedical Informatics*, *75*, 105–111. https://doi.org/10.1016/j.jbi.2017.05.015

Lodewijks, H. P. B., de Ruiter, C., & Doreleijers, T. A. H. (2010). The impact of protective factors in desistance from violent reoffending: A study in three samples of adolescent offenders. *Journal of Interpersonal Violence*, *25*(3), 568–587. https://doi.org/10.1177/0886260509334403

Loughran, T. A., Mulvey, E. P., Schubert, C. A., Fagan, J., Piquero, A. R., & Losoya, S. H. (2009). Estimating a dose-response relationship between length of stay and future recidivism in serious juvenile offenders. *Criminology*, *47*(3), 699–740. https://doi.org/10.1111/j.1745-9125.2009.00165.x

Luong, D., & Wormith, J. S. (2011). Applying risk/need assessment to probation practice and its impact on the recidivism of young offenders. *Criminal Justice and Behavior*, *38*(12), 1177–1199. https://doi.org/10.1177/0093854811421596

Lussier, P., Van Den Berg, C., Bijleveld, C., & Hendriks, J. (2012). A Developmental Taxonomy of Juvenile Sex Offenders for Theory, Research, and Prevention: The Adolescent-Limited and the High-Rate Slow Desister. *Criminal Justice and Behavior*, *39*(12), 1559–1581. https://doi.org/10.1177/0093854812455739

MacRae, L. D., Bertrand, L. D., Paetsch, J. J., & Hornick, J. P. (2011). *A Study of Youth Reoffending in Calgary*. http://hdl.handle.net/1880/107574

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & De Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, *4*, 1–14. https://doi.org/10.1186/1756-0500-4-299

McCann, K., & Lussier, P. (2008). Antisociality, Sexual Deviance,and Sexual Reoffending in Juvenile Sex Offenders: A Meta-Analytical Investigation. *Youth Violence and Juvenile Justice*, *6*(4), 363–385. https://doi.org/https://doi.org/10.1177/1541204008320260

McReynolds, L. S., Schwalbe, C. S., & Wasserman, G. A. (2010). The contribution of psychiatric disorder to juvenile recidivism. *Criminal Justice and Behavior*, *37*(2), 204–216. https://doi.org/10.1177/0093854809354961

Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors Associated With Persistence in Science and Engineering Majors: An Exploratory Study Using Classification Trees and Random Forests. *Journal of Engineering Education*, *97*(1), 57–70. https://doi.org/10.1002/j.2168-9830.2008.tb00954.x

Mills, J. F., & Kroner, D. G. (2006). Impression management and self-report among violent offenders. *Journal of Interpersonal Violence*, *21*(2), 178–192. https://doi.org/10.1177/0886260505282288

Moffitt, T. E. (1993). Adolescence-Limited and Life-Course-Persistent Antisocial Behavior: A Developmental Taxonomy. *Psychological Review*, *100*(4), 674–701. https://doi.org/doi.org/10.1037/0033-295X.100.4.674

Moffitt, T. E. (2003). Life-course-persistent and adolescence-limited antisocial behavior: A 10-year research review and a research agenda. *Causes of Conduct Disorder and Juvenile Delinquency*, 49–75.

Moffitt, T. E. (2018). Male antisocial behaviour in adolescence and beyond. *Nature Human Behaviour*, *2*(3), 177–186. https://doi.org/10.1038/s41562-018-0309-4

Monahan, J., Steadman, H. J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Silver, E., Roth, L. H., & Grisso, T. (2000). Developing a clinically useful actuarial tool for assessing violence risk. *British Journal of Psychiatry*, *176*(4). https://doi.org/doi.org/10.1192/bjp.176.4.312

Monahan, K., Steinberg, L., & Piquero, A. R. (2015). Juvenile justice policy and practice: A developmental perspective. *Crime and Justice*, *44*(1), 577–619. https://doi.org/10.1086/681553

Mook, D. G., & Scott, P. (2001). *Psychological Research: The Ideas Behind the Methods*. Norton.

Moore, A. A., Silberg, J. L., Roberson-Nay, R., & Mezuk, B. (2017). Life course persistent and adolescence limited conduct disorder in a nationally representative US sample: prevalence, predictors, and outcomes. *Social Psychiatry and Psychiatric Epidemiology*, *52*(4), 435–443. https://doi.org/10.1007/s00127-017-1337-5

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, *24*(1), 87–103. https://doi.org/10.1093/pan/mpv024

Mulder, E., Brand, E., Bullens, R., & van Marle, H. (2010). A classification of risk factors in serious juvenile offenders and the relation between patterns of risk factors and recidivism. *Criminal Behaviour and Mental Health, 20,* 23–38. https://doi.org/10.1002/cbm

Mulder, E., Brand, E., Bullens, R., & van Marle, H. (2011). Risk factors for overall recidivism and severity of recidivism in serious juvenile offenders. *International Journal of Offender Therapy and Comparative Criminology*, *55*(1), 118–135. https://doi.org/10.1177/0306624X09356683

Mulder, E., Vermunt, J., Brand, E., Bullens, R., & van Marle, H. (2012). Recidivism in subgroups of serious juvenile offenders: Different profiles, different risks? *Criminal Behaviour and Mental Health*, *22*(2), 122–135. https://doi.org/https://doi.org/10.1002/cbm.1819

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment*, *188*(1), 1–27. https://doi.org/10.1007/s10661-015-5049-6

Naplava, T. (2012). Kriterien zur Auswahl jugendlicher Intensivtäter auf der Basis von Rückfallanalysen. *SIAK-Journal - Zeitschrift Für Polizeiwissenschaft Und Polizeiliche Praxis, 2012*, 40–45. https://doi.org/10.7396/2012

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(21). https://doi.org/10.3389/fnbot.2013.00021

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

O'Brien, R. M. (2017). Dropping Highly Collinear Variables from a Model: Why it Typically is Not a Good Idea*. *Social Science Quarterly*, *98*(1), 360–375. https://doi.org/10.1111/ssqu.12273

Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, *5*(3), 3–7. https://doi.org/10.1186/1753-6561-5-S3-S11

Oh, J., Laubach, M., & Luczak, A. (2003). Estimating neuronal variable importance with Random Forest. In *Proceedings of the IEEE Annual Northeast Bioengineering Conference, NEBEC*. https://doi.org/10.1109/NEBC.2003.1215978

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice and Behavior*, *36*(4), 329–353. https://doi.org/10.1177/0093854809331457

Ottenbacher, K. J., Ottenbacher, H. R., Tooth, L., & Ostir, G. V. (2004). A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. *Journal of Clinical Epidemiology*, *57*(11), 1147–1152. https://doi.org/10.1016/j.jclinepi.2003.05.003

Pittman, S. J., & Brown, K. A. (2011). Multi-scale approach for predicting fish species distributions across coral reef seascapes. *PLoS ONE*, *6*(5), 1–12. https://doi.org/10.1371/journal.pone.0020583

Powers, D. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Qi, Y. (2012). Random forest for bioinformatics. *Ensemble Machine Learning: Methods and ApplicatiOns*, 307–323. https://doi.org/10.1007/9781441993267_10

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*.

Richters, J. E. (1997). The Hubble hypothesis and the developmentalist's dilemma. *Development and Psychopathology*, *9*(2), 193–229. https://doi.org/10.1017/s0954579497002022

Robinson, R. L. M., Palczewska, A., Palczewski, J., & Kidley, N. (2017). Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *Journal of Chemical Information and Modeling*, *57*(8), 1–84. https://doi.org/10.1021/acs.jcim.6b00753

Roe, B. P., Yang, H. J., & Zhu, J. (2006). Boosted decision trees, a powerful event classifier. In *Statistical Problems in Particle Physics, Astrophysics and Cosmology - Proceedings of PHYSTAT 2005*. Imperial College Press.

Rudin, C., Wang, C., & Coker, B. (2018). *The age of secrecy and unfairness in recidivism prediction*. http://arxiv.org/abs/1811.00731

Ruiz, A., & Villa, N. (2008). *Storms prediction : Logistic regression vs random forest for unbalanced data*. http://arxiv.org/abs/0804.0650

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*(4), 572–582. https://doi.org/10.1037/a0034177

Smith, A. D. A. C., Heron, J., Mishra, G., Gilthorpe, M. S., Ben-Shlomo, Y., & Tilling, K. (2015). Model Selection of the Effect of Binary Exposures over the Life Course. *Epidemiology*, *26*(5), 719–726. https://doi.org/10.1097/EDE.0000000000000348

Steinberg, L., & Scott, E. S. (2003). Less Guilty by Reason of Adolescence: Developmental Immaturity, Diminished Responsibility and the Juvenile Death Penalty. *American Psychologist*, *58*(12), 1009–1018. https://doi.org/10.1037/0003-066X.58.12.1009

Stelly, W., & Thomas, J. (2005). *Kriminalität im Lebenslauf*. Institut für Kriminologie der Universität Tübingen.

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 1–11. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Hothorn, T., & Zeileis, A. (2009). Party on! A new, conditional variable importance measure available in the party package. *The R Journal*, *2*, 14–17.

Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., Lu, C., & Zhao, N. (2016). Assessing the potential of random forest method for estimating solar radiation using air pollution index. *Energy Conversion and Management*, *119*, 121–129. https://doi.org/10.1016/j.enconman.2016.04.051

Svetnik, V., Liaw, A., & Tong, C. (2000). Variable Selection in Random Forest with Application to Quantitative Structure-Activity Relationship. In *Proceedings of the 7th Course on Ensemble Methods for Learning Machines*. https://www.csie.ntu.edu.tw/~b88052/tmp/vietri.pdf

Sweeten, G., Piquero, A. R., & Steinberg, L. (2013). Age and the Explanation of Crime, Revisited. *Journal of Youth and Adolescence*, *42*(6), 921–938. https://doi.org/10.1007/s10964-013-9926-4

Tantithamthavorn, C., McIntosh, S., Hassan, A. E., & Matsumoto, K. (2016). Automated parameter optimization of classification techniques for defect prediction models. *Proceedings - International Conference on Software Engineering*, *14-22-May-*, 321–332. https://doi.org/10.1145/2884781.2884857

Templ, M., Kowarik, A., Alfons, A., de Cillia, G., & Prantner, B. (2020). *Package 'VIM.'* R Package Version 6.0.0. https://cran.r-project.org/web/packages/VIM/VIM.pdf

Tollenaar, N., & van der Heiden, P. G. M. (2013). Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society*, *176*(2), 565–584. https://doi.org/doi.org/10.1111/j.1467-985x.2012.01056.x

Tonry, M. (2013). Legal and Ethical Issues in the Prediction of Recidivism. *SSRN Electronic Journal*, *167*. https://doi.org/10.2139/ssrn.2329849

van Buuren, S., Groothuis-Oudshoorn, K., Vink, G., Schouten, R., Robitzsch, A., Doove, L., Jolani, S., Moreno-Betancur, M., White, I., Gaffert, P., Meinfelder, F., & Gray, B. (2020). *Package 'mice.'* R Package Version 3.8.0. https://cran.r-project.org/web/packages/mice/mice.pdf

van der Put, C. E., Stams, G. J. J. M., Hoeve, M., Deković, M., Spanjaard, H. J. M., Van Der Laan, P. H., & Barnoski, R. P. (2012). Changes in the relative importance of dynamic risk factors for recidivism during adolescence. *International Journal of Offender Therapy and Comparative Criminology*, *56*(2), 296–316. https://doi.org/10.1177/0306624X11398462

van Vugt, E., Gibbs, J., Stams, G. J., Bijleveld, C., Hendriks, J., & Van Der Laan, P. (2011). Moral development and recidivism: A meta-analysis. *International Journal of Offender Therapy and Comparative Criminology*, *55*(8), 1234–1250. https://doi.org/10.1177/0306624X11396441

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. *Oxford Handbooks Online*. https://doi.org/10.1093/oxfordhb/9780199957996.013.14

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroisement. *Correspondance Mathématique et Physique*, *10*, 113–120.

Wainer, J. (2016). *Comparison of 14 different families of classification algorithms on 115 binary datasets* (Issue 2014). http://arxiv.org/abs/1606.00930

Walker, S. C., & Bishop, A. S. (2016). Length of stay, therapeutic change, and recidivism for incarcerated juvenile offenders. *Journal of Offender Rehabilitation*, *55*(6), 355–376. https://doi.org/10.1080/10509674.2016.1194946

Walter, J. (2006). Optimale Förderung oder was sollte Jugendstrafvollzug leisten? *Neue Kriminalpolitik*, *18*(3), 93–98. https://doi.org/10.5771/0934-9200-2006-3-93

Wei, W., Li, J., Cao, L., Ou, Y., & Chen, J. (2013). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, *16*(4), 449–475. https://doi.org/10.1007/s11280-012-0178-0

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2015). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *Medical Care*, *51*(3), 1–14. https://doi.org/10.1371/JOURNAL.PONE.0174944

Westreich, D. J., Lessler, J., & Jonsson Funk, M. (2011). Classification Methods As Alternatives To Logistic Regression. *Journal of Clinical Epidemiology*, *63*(8), 826–833. https://doi.org/10.1016/j.jclinepi.2009.11.020.Propensity

Winokur, K. P., Smith, A., Bontrager, S. R., & Blankenship, J. L. (2008). Juvenile recidivism and length of stay. *Journal of Criminal Justice*, *36*(2), 126–137. https://doi.org/10.1016/j.jcrimjus.2008.02.001

Woolley, M. E., Bowen, G. L., & Bowen, N. K. (2004). Cognitive pretesting and the developmental validity of child self-report instruments: theory and applications. *Research on Social Work Practice*, *14*(3), 191–200. https://doi.org/10.1177/1049731503257882

Wößner, G., & Wienhausen-Knezevic, E. (2013). No country for young men–Ausbildung und Beruf vor, während und nach der Inhaftierung im Jugendstrafvollzug. *Monatsschrift Für Kriminologie Und Strafrechtsreform*, *96*(6), 477–495. https://doi.org/doi.org/10.1515/mks-2013-960604

Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2016). Predicting 30-day all-cause readmissions from hospital inpatient discharge data. *IEEE 18th International Conference on E-Health Networking, Applications and Services (Healthcom)*, 1–6.

Appendices

Appendix A: Example of a tree model

Figure A1

Example tree on the imbalanced training dataset

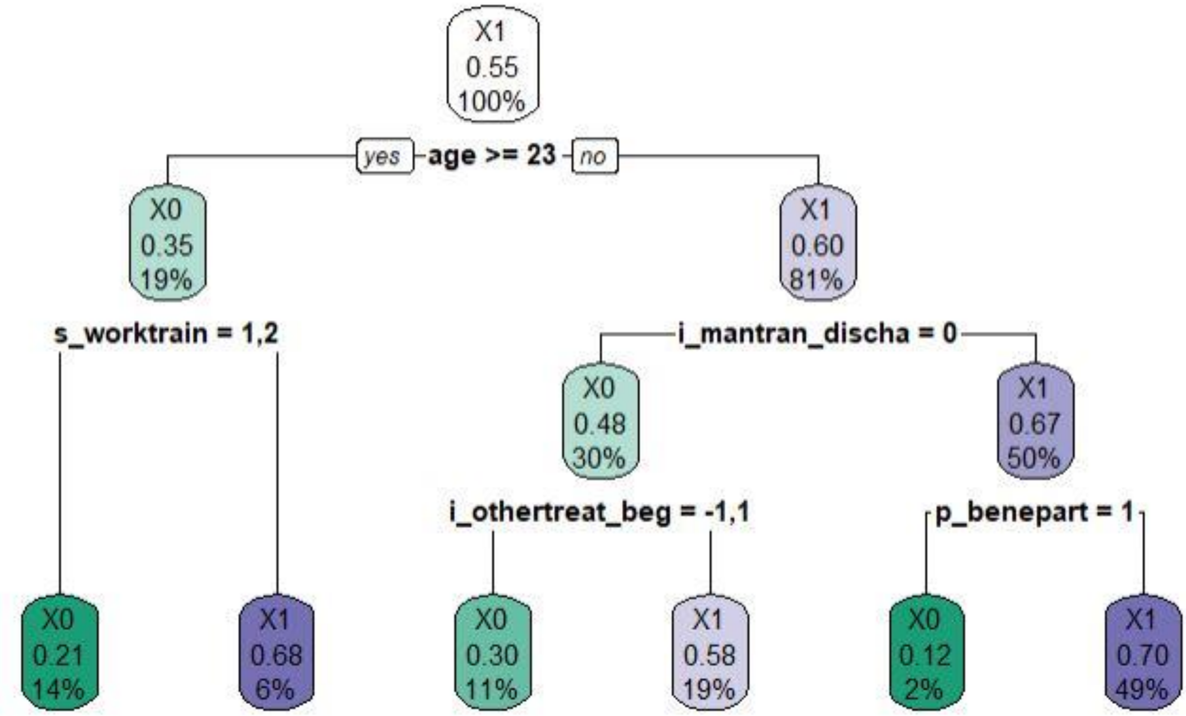


Figure A1 shows an example tree built on the imbalanced dataset. Inside the boxes, the first row refers to the class predicted at that stage: X1 stands for recidivism and X0 stands for non-recidivism. The second row inside the boxes contains the chance of recidivism. The third row inside the box refers to the number of YO observed in that node of the tree.

At the top of the tree, the probability that a YO belongs to the recidivism class is 0.55 (base rate). The first split is made on the variable age at the time of release. If a YO was 23 or more years old at the time of release, the left branch of the tree is followed. In that case, non-recidivism is predicted, and chances of recidivism are 0.35. 19% of the YO in the current sample were 23 or more years old. Following the left branch of the tree, the next split is made on the self-reported information whether the YO had work or vocational training after release (s_worktrain). 14% of the current sample were older than 23 and answered 1 or 2 on s_worktrain. If the YO's answer was "probably" or "yes" (1 or 2), non-recidivism is predicted, and chances of recidivism are 0.21. 14% of the current sample were 23 or more years old and answered 1 or 2 on the variable s_worktrain. If the YO's answer was not 1 or 2, which in this case means that the answer was "no" (0), the right branch of the tree is taken.

Recidivism is predicted and chances of recidivism are 0.68. 6% of the YO were 23 or more years old and answered not 1 or 2 on the variable s_worktrain.

Likewise, if a YO was younger than 23, the right branch is taken at the top of the tree and chances of recidivism are 0.60. If the YO then had “no need” of structured transition management after release (i_mantran_discha), non-recidivism is predicted, and chances of recidivism are 0.48. Afterwards, if a YO “participated in” other treatment interventions (i_othertreat_beg) or the information was “not known” (-1, 1), non-recidivism is predicted, and chances of recidivism are 0.30. If a YO “did not participate” in other treatment interventions, recidivism is predicted, and chances of recidivism are 0.58.

If a YO “did have a need” of structured transition management (i_mantran_discha) or the information was “not known” (-1, 1, 2), the right branch of the tree is taken. Recidivism is then predicted, and chances of recidivism are 0.67. Afterwards, if a YO “was” in a beneficial partnership (p_benepart), non-recidivism is predicted, and chances of recidivism are 0.12. If a YO “was not” in a beneficial partnership, recidivism is predicted, and chances of recidivism are 0.70.

Appendix B: Imputation methods

Table B1

Chosen imputation methods and their correlation/ accuracy with observed values

variable	number of missing cases	selected method	accuracy/ correlation with observed values[27]	variable type
s_evalYOI	170	RF	0.42	continous
s_evalsit	170	RF	0.52	continous
s_selfp	170	RF	0.47	continous
s_chanrela	189	cart	0.15	continous
s_supsatis	202	sample	0.15	continous
s_reached	170	cart	0.42	continous
s_eval	180	norm	0.28	continous
s_feel	170	cart	0.37	continous
s_recidivism	170	RF	0.45	continous
s_change	170	RF	0.24	continous
s_knowprobass	170	RF	0.72	categorical
s_contprobass	170	sample	0.77	categorical
s_dept	170	RF	0.66	categorical
s_worktrain	171	polyreg	0.48	categorical
s_partner	170	RF	0.67	categorical
s_viovic	170	RF	0.74	categorical
s_vioseen	170	sample	0.75	categorical
s_viodone	171	cart	0.72	categorical
s_consalc	253	RF	0.84	categorical
s_conscann	232	cart	0.82	categorical

[27] Accuracy for categorical variables, correlation for continuous variables

Appendix C: Optimization hyperparameters

In the following, the hyperparameters tried out for each model are listed. The hyperparameters that resulted in the best cross-validated prediction performances were used for the predictions on the testing dataset. They are printed in bold type.

Optimization hyperparameters for the imbalanced dataset

LR: caret

alpha[28]: 0.1, 0.1473684, 0.1947368, 0.2421053, 0.2894737, 0.3368421, 0.3842105, 0.4315789, 0.4789474, 0.5263158, 0.5736842, 0.6210526, 0.6684211, 0.7157895, 0.7631579, 0.8105263, 0.8578947, 0.9052632, 0.9526316, 1.0

lambda[29]: $2.870728 * 10^{-5}$, $4.451108 * 10^{-5}$, $6.901514 * 10^{-5}$, $1.070091 * 10^{-4}$, $1.659192 * 10^{-4}$, $2.572604 * 10^{-4}$, $3.988863 * 10^{-4}$, $6.184795 * 10^{-4}$, $9.589623 * 10^{-4}$, $1.486886 * 10^{-3}$, $2.305440 * 10^{-3}$, $3.574622 * 10^{-3}$, $5.542508 * 10^{-3}$, $8.593746 * 10^{-3}$, $1.332474 * 10^{-2}$, $2.066022 * 10^{-2}$, $3.203399 * 10^{-2}$, $4.966921 * 10^{-2}$, $7.701290 * 10^{-2}$, $1.194097 * 10^{-1}$

LR: manual

Alpha: From 0 to 1 in steps of 0.005 (selected: **0.02**)

Lambda: From 0 to 1 in steps of 0.005 (selected: **0.99**)

RF: caret

mtry[30]: **2**, 15, 29, 43, 57, 71, 85, 99, 113, 127, 141, 155, 169, 183, 197, 211, 225, 239, 253, 267

ntree[31]: **500**

RF: manual

mtry: from 1 to 24 in steps of 1

ntree: 1500

BCT: caret

interaction.depth[32]: from 1 to 10 in steps of 1

ntree: from 50 to 1000 in steps of 50

---

[28] alpha determines the degree to which lasso regression versus ridge regression is performed (when alpha = 0, lasso is performed; when alpha = 1, ridge is performed).

[29] lambda is a regularization hyperparameter that penalizes model complexity (the larger it is, the more it penalizes complexity).

[30] Mtry is the number of variables from which the next predictor is selected at each split.

[31] Ntree determines the number of trees grown in total.

[32] Interaction.depth is the maximum number of splits performed at each tree.

shrinkage[33]: 0.1

n.minobsinnode[34]: 10

BCT: manual

interaction.depth: **1**, 2

ntree: from 20 to 140 in steps of 20 (selected: **80**)

shrinkage: from 0.02 to 0.2 in steps of 0.02 (selected: **0.04**)

n.minobsinnode: from 3 to 10 in steps of 1 (selected: **6**)

Optimization hyperparameters for the balanced dataset

LR: caret

Alpha: 0.1, 0.1473684, 0.1947368, 0.2421053, 0.2894737, 0.3368421, 0.3842105, 0.4315789, 0.4789474, 0.5263158, 0.5736842, 0.6210526, 0.6684211, 0.7157895, 0.7631579, 0.8105263, 0.8578947, 0.9052632, 0.9526316, 1.0

Lambda: $2.802222 * 10^{-5}$, $4.344889 * 10^{-5}$, $6.736819 * 10^{-5}$, $1.044554 * 10^{-4}$, $1.619598 * 10^{-4}$, $2.511212 * 10^{-4}$, $3.893675 * 10^{-4}$, $6.037204 * 10^{-4}$, $9.360780 * 10^{-4}$, $1.451404 * 10^{-3}$, $2.250424 * 10^{-3}$, $3.489319 * 10^{-3}$, $5.410244 * 10^{-3}$, $8.388668 * 10^{-3}$, $1.300676 * 10^{-2}$, $2.016719 * 10^{-2}$, $3.126955 * 10^{-2}$, $4.848393 * 10^{-2}$, $7.517509 * 10^{-2}$, $1.165602 * 10^{-1}$

LR: manual

Alpha: From 0 to 1 in steps of 0.005 (selected: **0.025**)

Lambda: From 0 to 1 in steps of 0.005 (selected: **0.955**)

RF: caret

mtry: **2**, 15, 29, 43, 57, 71, 85, 99, 113, 127, 141, 155, 169, 183, 197, 211, 225, 239, 253, 267

ntree: **500**

RF: manual

mtry: from 1 to 24 in steps of 1

ntree: 1500

BCT: caret

interaction.depth: from 1 to 10 in steps of 1 (selected: **4**)

ntree: from 50 to 1000 in steps of 50 (selected: **50**)

shrinkage: **0.1**

---

[33] Shrinkage determines how quickly the algorithm adapts.

[34] N.minobsinnode determines the minimum number of training set samples in a terminal node.

n.minobsinnode: **10**

BCT: manual

interaction.depth: 1, 2

ntree: from 20 to 140 in steps of 20

shrinkage: from 0.02 to 0.2 in steps of 0.02

n.minobsinnode: from 3 to 10 in steps of 1

Appendix D: Variable importances on the imbalanced dataset

Figure D1

Variable importances on the imbalanced dataset

Appendix E: Partial dependencies of the predictor variable age

Figure E.1

Partial dependences of age on recidivism by LR, RF, and BCT, and the prediction of recidivism dependent on age on the training dataset

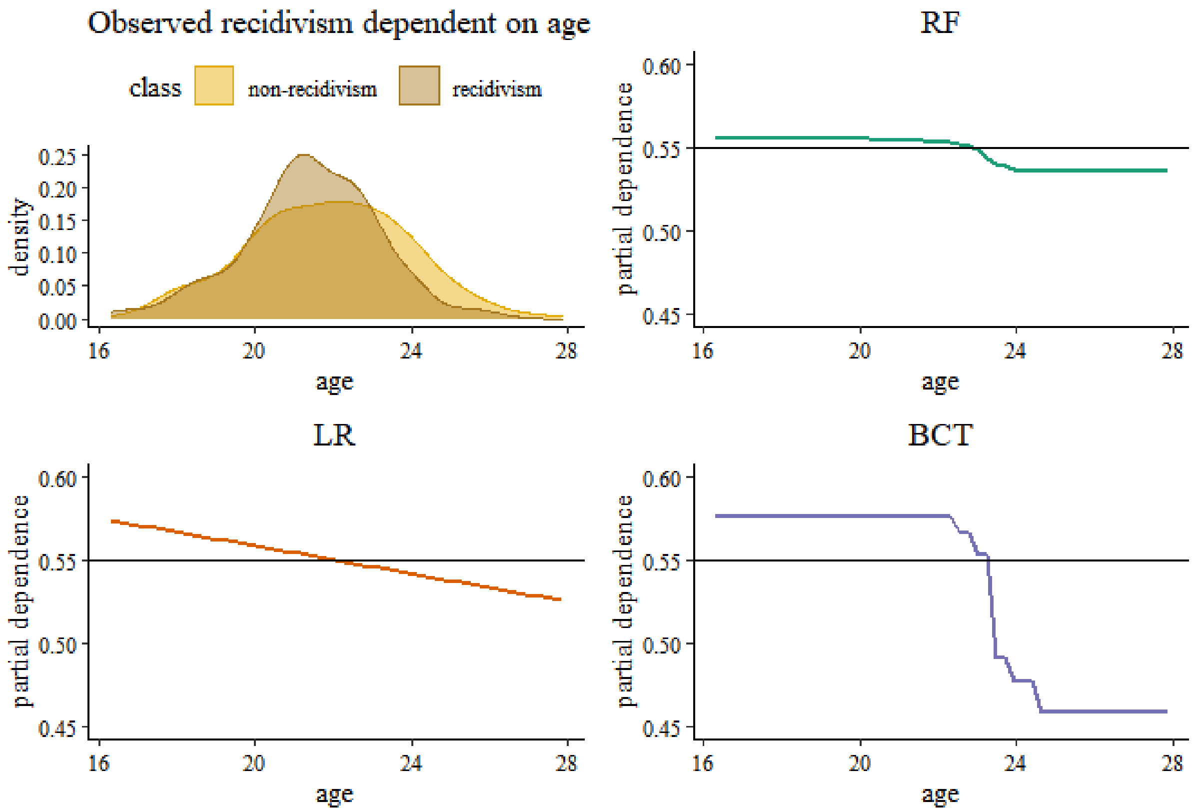


Figure E1 shows that RF and GBM can more flexibly fit the age variable than LR. The plot on the top left contains the observed distributions of non-recidivism and recidivism on the training dataset. It shows that for YO who are approximately 23 or less years old, recidivism is more likely than non-recidivism. Furthermore, it shows that for YO who are older than 23, non-recidivism is more likely than recidivism. The plot on the bottom left shows that in LR, the prediction of recidivism is linearly dependent on age. Contrarily, in RF and BCT the prediction of recidivism declines around the age of 23. Thus, RF and BCT fit the observed distribution more closely than LR.

Appendix F: Variable description

Table F1

Variable description and proportions of (non-)recidivism on the whole dataset

	variable	non-recidivism (N=288)	recidivism (N=355)	total (N=643)
1	age	age on day of release in years		
	mean (SD)	21.966 (1.999)	21.399 (1.850)	21.653 (1.938)
2	country_birth	country of birth		
	0 = Germany	253 (87.8%)	336 (94.6%)	589 (91.6%)
	1 = else	35 (12.2%)	19 (5.4%)	54 (8.4%)
3	days_sentence	number of days spent in young offender institution		
	mean (SD)	455.785 (229.069)	438.755 (209.320)	446.383 (218.378)
4	com_robbery	violation of chapter 20 GCC (robbery and extortion)		
	0 = no	220 (76.4%)	257 (72.4%)	477 (74.2%)
	1 = yes	68 (23.6%)	98 (27.6%)	166 (25.8%)
5	com_bodilyharm	violation of chapter 17 GCC (offences against physical integrity)		
	0 = no	161 (55.9%)	203 (57.2%)	364 (56.6%)
	1 = yes	127 (44.1%)	152 (42.8%)	279 (43.4%)
6	com_drugs	violation of the Narcotics Act		
	0 = no	239 (83.0%)	311 (87.6%)	550 (85.5%)
	1 = yes	49 (17.0%)	44 (12.4%)	93 (14.5%)
7	com_theft	violation of chapter 19 GCC (theft and misappropriation)		
	0 = no	154 (53.5%)	140 (39.4%)	294 (45.7%)
	1 = yes	134 (46.5%)	215 (60.6%)	349 (54.3%)
8	com_fraud	violation of chapter 22 GCC (fraud and embezzlement)		
	0 = no	236 (81.9%)	270 (76.1%)	506 (78.7%)
	1 = yes	52 (18.1%)	85 (23.9%)	137 (21.3%)
9	com_liberty	violation of chapter 18 GCC (offences against personal liberty)		
	0 = no	256 (88.9%)	317 (89.3%)	573 (89.1%)
	1 = yes	32 (11.1%)	38 (10.7%)	70 (10.9%)
10	com_traffic	violation of Road Traffic Regulations		
	0 = no	261 (90.6%)	314 (88.5%)	575 (89.4%)
	1 = yes	27 (9.4%)	41 (11.5%)	68 (10.6%)
11	com.total	number of applied crime categories (including GCC chapter 13, 16 and 28)		
	mean (SD)	1.774 (1.073)	1.946 (1.079)	1.869 (1.079)
12	p_dealswith	PA: deals with his deed seriously		
	-1 = not relevant or known	21 (7.3%)	13 (3.7%)	34 (5.3%)

	0 = no/ rudimentary	105 (36.5%)	161 (45.4%)	266 (41.4%)
	1 = almost/ fully	162 (56.2%)	181 (51.0%)	343 (53.3%)
13	p_workson	PA: works actively to reach the goal of the young offender institution		
	-1 = not relevant or known	8 (2.8%)	6 (1.7%)	14 (2.2%)
	0 = no/ rudimentary	92 (31.9%)	135 (38.0%)	227 (35.3%)
	1 = almost/ fully	188 (65.3%)	214 (60.3%)	402 (62.5%)
14	p_willing	PA: is willing to follow a well-structured educational training or work after release		
	-1 = not relevant or known	15 (5.2%)	12 (3.4%)	27 (4.2%)
	0 = no/ rudimentary	60 (20.8%)	94 (26.5%)	154 (24.0%)
	1 = almost/ fully	213 (74.0%)	249 (70.1%)	462 (71.9%)
15	p_able	PA: is able to follow a well-structured educational training or work after release		
	-1 = not relevant or known	16 (5.6%)	12 (3.4%)	28 (4.4%)
	0 = no/ rudimentary	59 (20.5%)	94 (26.5%)	153 (23.8%)
	1 = almost/ fully	213 (74.0%)	249 (70.1%)	462 (71.9%)
16	p_benefam	PA: has beneficial family relationships		
	-1 = not relevant or known	94 (32.6%)	103 (29.0%)	197 (30.6%)
	0 = no/ rudimentary	72 (25.0%)	119 (33.5%)	191 (29.7%)
	1 = almost/ fully	122 (42.4%)	133 (37.5%)	255 (39.7%)
17	p_benepart	PA: is in a beneficial partnership		
	-1 = not relevant or known	83 (28.8%)	106 (29.9%)	189 (29.4%)
	0 = no/ rudimentary	172 (59.7%)	225 (63.4%)	397 (61.7%)
	1 = almost/ fully	33 (11.5%)	24 (6.8%)	57 (8.9%)
18	p_benefrie	PA: has beneficial friends outside of YOI		
	-1 = not relevant or known	185 (64.2%)	208 (58.6%)	393 (61.1%)
	0 = no/ rudimentary	83 (28.8%)	128 (36.1%)	211 (32.8%)
	1 = almost/ fully	20 (6.9%)	19 (5.4%)	39 (6.1%)
19	p_viopr	PA: is assumed to be highly violence-prone		
	-1 = not relevant or known	26 (9.0%)	23 (6.5%)	49 (7.6%)
	0 = no/ rudimentary	146 (50.7%)	178 (50.1%)	324 (50.4%)
	1 = almost/ fully	116 (40.3%)	154 (43.4%)	270 (42.0%)

20	p_drugs	PA: has a substantial problem with drug addiction		
	-1 = not relevant or known	26 (9.0%)	35 (9.9%)	61 (9.5%)
	0 = no/ rudimentary	121 (42.0%)	117 (33.0%)	238 (37.0%)
	1 = almost/ fully	141 (49.0%)	203 (57.2%)	344 (53.5%)
21	p_alc	PA: has a substantial problem with alcohol addiction		
	-1 = not relevant or known	69 (24.0%)	103 (29.0%)	172 (26.7%)
	0 = no/ rudimentary	145 (50.3%)	152 (42.8%)	297 (46.2%)
	1 = almost/ fully	74 (25.7%)	100 (28.2%)	174 (27.1%)
22	p_realpl	PA: has future plans that are realistic and reachable by legal means		
	-1 = not relevant or known	23 (8.0%)	17 (4.8%)	40 (6.2%)
	0 = no/ rudimentary	115 (39.9%)	179 (50.4%)	294 (45.7%)
	1 = almost/ fully	150 (52.1%)	159 (44.8%)	309 (48.1%)
23	p_general	PA: estimated risk that the prisoner will commit any crime after release		
	-1 = not relevant or known	17 (5.9%)	6 (1.7%)	23 (3.6%)
	0 = no/ rudimentary	51 (17.7%)	45 (12.7%)	96 (14.9%)
	1 = almost/ fully	220 (76.4%)	304 (85.6%)	524 (81.5%)
24	p_violence	PA: estimated risk the prisoner will commit a violent crime after release		
	-1 = not relevant or known	44 (15.3%)	46 (13.0%)	90 (14.0%)
	0 = no/ rudimentary	110 (38.2%)	140 (39.4%)	250 (38.9%)
	1 = almost/ fully	134 (46.5%)	169 (47.6%)	303 (47.1%)
25	p_sexual	PA: estimated risk that the prisoner will commit a sexual crime after release		
	-1 = not relevant or known	39 (13.5%)	43 (12.1%)	82 (12.8%)
	0 = no/ rudimentary	239 (83.0%)	299 (84.2%)	538 (83.7%)
	1 = almost/ fully	10 (3.5%)	13 (3.7%)	23 (3.6%)
26	p_leave	PA: was allowed to go on temporary YOI leaves		
	0 = no	142 (49.3%)	184 (51.8%)	326 (50.7%)
	1 = yes once	20 (6.9%)	43 (12.1%)	63 (9.8%)
	2 = yes multiple times	126 (43.8%)	128 (36.1%)	254 (39.5%)
27	p_unacc	PA: was allowed to go on temporary unaccompanied YOI leaves		

	-1 = not known	2 (0.7%)	3 (0.8%)	5 (0.8%)
	0 = no	224 (77.8%)	301 (84.8%)	525 (81.6%)
	1 = yes once	2 (0.7%)	10 (2.8%)	12 (1.9%)
	2 = yes multiple times	60 (20.8%)	41 (11.5%)	101 (15.7%)
28	p_vac	PA: was allowed to go on vacational YOI leave		
	0 = no	256 (88.9%)	329 (92.7%)	585 (91.0%)
	1 = yes once	10 (3.5%)	13 (3.7%)	23 (3.6%)
	2 = yes multiple times	22 (7.6%)	13 (3.7%)	35 (5.4%)
29	p_school	PA: after release, has a place in school		
	-1 = not relevant or known	138 (47.9%)	125 (35.2%)	263 (40.9%)
	0 = no	137 (47.6%)	193 (54.4%)	330 (51.3%)
	1 = probably	4 (1.4%)	19 (5.4%)	23 (3.6%)
	2 = yes	9 (3.1%)	18 (5.1%)	27 (4.2%)
30	p_voctrain	PA: after release, the prisoner is enrolled in a vocational training		
	-1 = not relevant or known	78 (27.1%)	83 (23.4%)	161 (25.0%)
	0 = no	141 (49.0%)	199 (56.1%)	340 (52.9%)
	1 = probably	36 (12.5%)	55 (15.5%)	91 (14.2%)
	2 = yes	33 (11.5%)	18 (5.1%)	51 (7.9%)
31	p_work	PA: after release, the prisoner has a workplace		
	-1 = not relevant or known	80 (27.8%)	104 (29.3%)	184 (28.6%)
	0 = no	148 (51.4%)	215 (60.6%)	363 (56.5%)
	1 = probably	40 (13.9%)	27 (7.6%)	67 (10.4%)
	2 = yes	20 (6.9%)	9 (2.5%)	29 (4.5%)
32	p_housing	PA: after release, the prisoner has an accommodation		
	-1 = not relevant or known	20 (6.9%)	22 (6.2%)	42 (6.5%)
	0 = no	6 (2.1%)	23 (6.5%)	29 (4.5%)
	1 = probably	12 (4.2%)	26 (7.3%)	38 (5.9%)
	2 = yes	250 (86.8%)	284 (80.0%)	534 (83.0%)
33	p_conprobass	PA: in case the prisoner is released on parole, was he in contact with the probation assistant during his stay		
	-1 = not relevant or known	140 (48.6%)	189 (53.2%)	329 (51.2%)
	0 = no	69 (24.0%)	73 (20.6%)	142 (22.1%)
	1 = yes	68 (23.6%)	79 (22.3%)	147 (22.9%)

	writing/telephone			
	2 = yes, personal	11 (3.8%)	14 (3.9%)	25 (3.9%)
34	p_consup	PA: in case the supervision of conduct takes place, was the prisoner in contact with the probation assistant during his stay		
	-1 = not relevant or known	215 (74.7%)	244 (68.7%)	459 (71.4%)
	0 = no	53 (18.4%)	65 (18.3%)	118 (18.4%)
	1 = yes writing/telephone	14 (4.9%)	36 (10.1%)	50 (7.8%)
	2 = yes, personal	6 (2.1%)	10 (2.8%)	16 (2.5%)
35	p_need_isona	PA: necessity to participate in transition project ISONA		
	0 = no	268 (93.1%)	326 (91.8%)	594 (92.4%)
	1 = yes	20 (6.9%)	29 (8.2%)	49 (7.6%)
36	p_need_dna	PA: necessity to participate in transition project DNA		
	0 = no	207 (71.9%)	242 (68.2%)	449 (69.8%)
	1 = yes	81 (28.1%)	113 (31.8%)	194 (30.2%)
37	p.beg_isona	PA: began participation in transition project DNA		
	0 = no	262 (91.0%)	323 (91.0%)	585 (91.0%)
	1 = yes	26 (9.0%)	32 (9.0%)	58 (9.0%)
38	i_elemen_need	PA: necessity for participation in elementary course		
	0 = no	268 (93.1%)	321 (90.4%)	589 (91.6%)
	1 = yes	20 (6.9%)	34 (9.6%)	54 (8.4%)
39	i_remed_need	PA: necessity for participation in remedial course for school		
	0 = no	250 (86.8%)	287 (80.8%)	537 (83.5%)
	1 = yes	38 (13.2%)	68 (19.2%)	106 (16.5%)
40	i_remed_discha	PA: after release, planned continuation of remedial course for school		
	-1 = not relevant or known	11 (3.8%)	10 (2.8%)	21 (3.3%)
	0 = not necessary	267 (92.7%)	314 (88.5%)	581 (90.4%)
	1 = necessary but not planned	10 (3.5%)	29 (8.2%)	39 (6.1%)
	2 = necessary and planned	0 (0.0%)	2 (0.6%)	2 (0.3%)
41	i_school_need	PA: necessity for participation in measures to obtain a school leaving certificate		
	0 = no	73 (25.3%)	75 (21.1%)	148 (23.0%)
	1 = yes	215 (74.7%)	280 (78.9%)	495 (77.0%)
42	i_school_beg	PA: began participation in measures to obtain a school leaving certificate		

	-1 = not relevant or known	5 (1.7%)	3 (0.8%)	8 (1.2%)
	0 = no due to YO	25 (8.7%)	29 (8.2%)	54 (8.4%)
	1 = no due to YOI	170 (59.0%)	201 (56.6%)	371 (57.7%)
	2 = yes	88 (30.6%)	122 (34.4%)	210 (32.7%)
43	i_school_discon	PA: discontinued measures to obtain a school leaving certificate		
	-1 = not relevant or known	201 (69.8%)	231 (65.1%)	432 (67.2%)
	0 = no	73 (25.3%)	99 (27.9%)	172 (26.7%)
	1 = yes due to YOI	5 (1.7%)	7 (2.0%)	12 (1.9%)
	2 = yes due to YO	9 (3.1%)	18 (5.1%)	27 (4.2%)
44	i_school_reach	PA: reached goals of the measures to obtain a school leaving certificate		
	-1 = not relevant or known	205 (71.2%)	236 (66.5%)	441 (68.6%)
	0 = no/ rudimentary1 = almost/ fully	17 (5.9%)	26 (7.3%)	43 (6.7%)
	1 = almost/ fully	66 (22.9%)	93 (26.2%)	159 (24.7%)
45	i_school_discha	PA: after release, planned continuation of measures to obtain a school leaving certificate		
	-1 = not relevant or known	19 (6.6%)	17 (4.8%)	36 (5.6%)
	0 = not necessary	174 (60.4%)	184 (51.8%)	358 (55.7%)
	1 = necessary but not planned	83 (28.8%)	127 (35.8%)	210 (32.7%)
	2 = necessary and planned	12 (4.2%)	27 (7.6%)	39 (6.1%)
46	i_vocprep_need	PA: necessity for participation in vocational preparation training		
	0 = no	99 (34.4%)	92 (25.9%)	191 (29.7%)
	1 = yes	189 (65.6%)	263 (74.1%)	452 (70.3%)
47	i_vocprep_beg	PA: began participation in vocational preparation training		
	-1 = not relevant or known	5 (1.7%)	7 (2.0%)	12 (1.9%)
	0 = no due to YO	21 (7.3%)	25 (7.0%)	46 (7.2%)
	1 = no due to YOI	215 (74.7%)	264 (74.4%)	479 (74.5%)
	2 = yes	47 (16.3%)	59 (16.6%)	106 (16.5%)
48	i_vocprep_discon	PA: discontinued vocational preparation training		
	-1 = not relevant or known	242 (84.0%)	294 (82.8%)	536 (83.4%)
	0 = no	35 (12.2%)	43 (12.1%)	78 (12.1%)

	1 = yes due to YOI	7 (2.4%)	11 (3.1%)	18 (2.8%)
	2 = yes due to YO	4 (1.4%)	7 (2.0%)	11 (1.7%)
49	i_vocprep_reach	PA: reached goals of the vocational preparation training		
	-1 = not relevant or known	247 (85.8%)	297 (83.7%)	544 (84.6%)
	0 = no/ rudimentary1 = almost/ fully	12 (4.2%)	13 (3.7%)	25 (3.9%)
	1 = almost/ fully	29 (10.1%)	45 (12.7%)	74 (11.5%)
50	i_vocprep_discha	PA: after release, planned continuation of vocational preparation training		
	-1 = not relevant or known	16 (5.6%)	30 (8.5%)	46 (7.2%)
	0 = not necessary	169 (58.7%)	179 (50.4%)	348 (54.1%)
	1 = necessary but not planned	91 (31.6%)	127 (35.8%)	218 (33.9%)
	2 = necessary and planned	12 (4.2%)	19 (5.4%)	31 (4.8%)
51	i_vocqua_need	PA: necessity for participation in vocational qualification training		
	0 = no	41 (14.2%)	28 (7.9%)	69 (10.7%)
	1 = yes	247 (85.8%)	327 (92.1%)	574 (89.3%)
52	i_vocqua_beg	PA: began participation in vocational qualification training		
	-1 = not relevant or known	3 (1.0%)	4 (1.1%)	7 (1.1%)
	0 = no due to YO	13 (4.5%)	21 (5.9%)	34 (5.3%)
	1 = no due to YOI	127 (44.1%)	153 (43.1%)	280 (43.5%)
	2 = yes	145 (50.3%)	177 (49.9%)	322 (50.1%)
53	i_vocqua_discon	PA: discontinued vocational qualification training		
	-1 = not relevant or known	144 (50.0%)	179 (50.4%)	323 (50.2%)
	0 = no	100 (34.7%)	110 (31.0%)	210 (32.7%)
	1 = yes due to YOI	26 (9.0%)	26 (7.3%)	52 (8.1%)
	2 = yes due to YO	18 (6.2%)	40 (11.3%)	58 (9.0%)
54	i_vocqua_reach	PA: reached goals of vocational qualification training		
	-1 = not relevant or known	159 (55.2%)	203 (57.2%)	362 (56.3%)
	0 = no/ rudimentary1 = almost/ fully	51 (17.7%)	76 (21.4%)	127 (19.8%)
	1 = almost/ fully	78 (27.1%)	76 (21.4%)	154 (24.0%)
55	i_vocqua_discha	PA: after release, planned continuation of vocational		

		qualification training		
	-1 = not relevant or known	23 (8.0%)	34 (9.6%)	57 (8.9%)
	0 = not necessary	97 (33.7%)	95 (26.8%)	192 (29.9%)
	1 = necessary but not planned	148 (51.4%)	204 (57.5%)	352 (54.7%)
	2 = necessary and planned	20 (6.9%)	22 (6.2%)	42 (6.5%)
56	i_vocedu_need	PA: necessity for participation in vocational education		
	0 = no	26 (9.0%)	18 (5.1%)	44 (6.8%)
	1 = yes	262 (91.0%)	337 (94.9%)	599 (93.2%)
57	i_vocedu_beg	PA: began participation in vocational education		
	-1 = not relevant or known	2 (0.7%)	1 (0.3%)	3 (0.5%)
	0 = no due to YO	13 (4.5%)	21 (5.9%)	34 (5.3%)
	1 = no due to YOI	224 (77.8%)	283 (79.7%)	507 (78.8%)
	2 = yes	49 (17.0%)	50 (14.1%)	99 (15.4%)
58	i_vocedu_discon	PA: discontinued vocational education		
	-1 = not relevant or known	237 (82.3%)	308 (86.8%)	545 (84.8%)
	0 = no	35 (12.2%)	28 (7.9%)	63 (9.8%)
	1 = yes due to YOI	11 (3.8%)	13 (3.7%)	24 (3.7%)
	2 = yes due to YO	5 (1.7%)	6 (1.7%)	11 (1.7%)
59	i_vocedu_reach	PA: reached goals of vocational education		
	-1 = not relevant or known	246 (85.4%)	312 (87.9%)	558 (86.8%)
	0 = no/ rudimentary1 = almost/ fully	15 (5.2%)	18 (5.1%)	33 (5.1%)
	1 = almost/ fully	27 (9.4%)	25 (7.0%)	52 (8.1%)
60	i_vocedu_discha	PA: after release, planned continuation of vocational education		
	-1 = not relevant or known	20 (6.9%)	31 (8.7%)	51 (7.9%)
	0 = not necessary	47 (16.3%)	47 (13.2%)	94 (14.6%)
	1 = necessary but not planned	186 (64.6%)	248 (69.9%)	434 (67.5%)
	2 = necessary and planned	35 (12.2%)	29 (8.2%)	64 (10.0%)
61	i_occthe_need	PA: necessity for participation in occupational therapy		
	0 = no	263 (91.3%)	295 (83.1%)	558 (86.8%)
	1 = yes	25 (8.7%)	60 (16.9%)	85 (13.2%)

62	i_occthe_discha	PA: after release, planned continuation of occupational therapy		
	-1 = not relevant or known	8 (2.8%)	13 (3.7%)	21 (3.3%)
	0 = not necessary	267 (92.7%)	305 (85.9%)	572 (89.0%)
	1 = necessary but not planned	11 (3.8%)	33 (9.3%)	44 (6.8%)
	2 = necessary and planned	2 (0.7%)	4 (1.1%)	6 (0.9%)
63	i_psythe_need	PA: necessity for participation in psychotherapy		
	0 = no	258 (89.6%)	306 (86.2%)	564 (87.7%)
	1 = yes	30 (10.4%)	49 (13.8%)	79 (12.3%)
64	i_psythe_discha	PA: after release, planned continuation of psychotherapy		
	-1 = not relevant or known	9 (3.1%)	15 (4.2%)	24 (3.7%)
	0 = not necessary	257 (89.2%)	305 (85.9%)	562 (87.4%)
	1 = necessary but not planned	15 (5.2%)	18 (5.1%)	33 (5.1%)
	2 = necessary and planned	7 (2.4%)	17 (4.8%)	24 (3.7%)
65	i_antvio_need	PA: necessity for participation in anti-violence training		
	0 = no	188 (65.3%)	229 (64.5%)	417 (64.9%)
	1 = yes	100 (34.7%)	126 (35.5%)	226 (35.1%)
66	i_antvio_beg	PA: began participation in anti-violence training		
	-1 = not relevant or known	11 (3.8%)	9 (2.5%)	20 (3.1%)
	0 = no due to YO	14 (4.9%)	24 (6.8%)	38 (5.9%)
	1 = no due to YOI	217 (75.3%)	256 (72.1%)	473 (73.6%)
	2 = yes	46 (16.0%)	66 (18.6%)	112 (17.4%)
67	i_antvio_discon	PA: discontinued anti-violence training		
	-1 = not relevant or known	242 (84.0%)	291 (82.0%)	533 (82.9%)
	0 = no	42 (14.6%)	59 (16.6%)	101 (15.7%)
	1 = yes due to YOI	2 (0.7%)	1 (0.3%)	3 (0.5%)
	2 = yes due to YO	2 (0.7%)	4 (1.1%)	6 (0.9%)
68	i_antvio_reach	PA: reached goals of anti-violence training		
	-1 = not relevant or known	246 (85.4%)	293 (82.5%)	539 (83.8%)
	0 = no/ rudimentary	4 (1.4%)	14 (3.9%)	18 (2.8%)
	1 = almost/ fully	38 (13.2%)	48 (13.5%)	86 (13.4%)
69	i_antvio_discha	PA: after release, planned continuation of anti-violence training		

	-1 = not relevant or known	23 (8.0%)	19 (5.4%)	42 (6.5%)
	0 = not necessary	227 (78.8%)	275 (77.5%)	502 (78.1%)
	1 = necessary but not planned	35 (12.2%)	56 (15.8%)	91 (14.2%)
	2 = necessary and planned	3 (1.0%)	5 (1.4%)	8 (1.2%)
70	i_deltreat_need	PA: necessity for participation in other delict- or problem specific measures		
	0 = no	40 (13.9%)	32 (9.0%)	72 (11.2%)
	1 = yes	248 (86.1%)	323 (91.0%)	571 (88.8%)
71	i_deltreat_beg	PA: began participation in other delict- or problem specific measures		
	-1 = not relevant or known	13 (4.5%)	15 (4.2%)	28 (4.4%)
	0 = no due to YO	30 (10.4%)	44 (12.4%)	74 (11.5%)
	1 = no due to YOI	67 (23.3%)	52 (14.6%)	119 (18.5%)
	2 = yes	178 (61.8%)	244 (68.7%)	422 (65.6%)
72	i_deltreat_discon	PA: discontinued other delict- or problem specific measures		
	-1 = not relevant or known	119 (41.3%)	119 (33.5%)	238 (37.0%)
	0 = no	157 (54.5%)	208 (58.6%)	365 (56.8%)
	1 = yes due to YOI	6 (2.1%)	15 (4.2%)	21 (3.3%)
	2 = yes due to YO	6 (2.1%)	13 (3.7%)	19 (3.0%)
73	i_deltreat_reach	PA: reached goals of other delict- or problem specific measures		
	-1 = not relevant or known	127 (44.1%)	131 (36.9%)	258 (40.1%)
	0 = no/ rudimentary	25 (8.7%)	60 (16.9%)	85 (13.2%)
	1 = almost/ fully	136 (47.2%)	164 (46.2%)	300 (46.7%)
74	i_deltreat_discha	PA: after release, planned continuation of other delict- or problem specific measures		
	-1 = not relevant or known	40 (13.9%)	53 (14.9%)	93 (14.5%)
	0 = not necessary	172 (59.7%)	187 (52.7%)	359 (55.8%)
	1 = necessary but not planned	62 (21.5%)	92 (25.9%)	154 (24.0%)
	2 = necessary and planned	14 (4.9%)	23 (6.5%)	37 (5.8%)
75	i_addcon_need	PA: necessity for participation in addiction counselling		
	0 = no	50 (17.4%)	46 (13.0%)	96 (14.9%)

	1 = yes	238 (82.6%)	309 (87.0%)	547 (85.1%)
76	i_addcon_beg	PA: began participation in addiction counselling		
	-1 = not relevant or known	6 (2.1%)	10 (2.8%)	16 (2.5%)
	0 = no due to YO	23 (8.0%)	26 (7.3%)	49 (7.6%)
	1 = no due to YOI	57 (19.8%)	50 (14.1%)	107 (16.6%)
	2 = yes	202 (70.1%)	269 (75.8%)	471 (73.3%)
77	i_addcon_discon	PA: discontinued addiction counselling		
	-1 = not relevant or known	97 (33.7%)	93 (26.2%)	190 (29.5%)
	0 = no	165 (57.3%)	216 (60.8%)	381 (59.3%)
	1 = yes due to YOI	9 (3.1%)	13 (3.7%)	22 (3.4%)
	2 = yes due to YO	17 (5.9%)	33 (9.3%)	50 (7.8%)
78	i_addcon_reach	PA: reached goals of addiction counselling		
	- -1	121 (42.0%)	124 (34.9%)	245 (38.1%)
	- 0	36 (12.5%)	69 (19.4%)	105 (16.3%)
	- 1	131 (45.5%)	162 (45.6%)	293 (45.6%)
79	i_addcon_discha	PA: after release, planned continuation of addiction counselling		
	-1 = not relevant or known	42 (14.6%)	50 (14.1%)	92 (14.3%)
	0 = not necessary	116 (40.3%)	99 (27.9%)	215 (33.4%)
	1 = necessary but not planned	60 (20.8%)	101 (28.5%)	161 (25.0%)
	2 = necessary and planned	70 (24.3%)	105 (29.6%)	175 (27.2%)
80	i_addthe_need	PA: necessity for participation in addiction therapy		
	0 = no	185 (64.2%)	197 (55.5%)	382 (59.4%)
	1 = yes	103 (35.8%)	158 (44.5%)	261 (40.6%)
81	i_addthe_beg	PA: began participation in addiction therapy		
	-1 = not relevant or known	13 (4.5%)	14 (3.9%)	27 (4.2%)
	0 = no due to YO	12 (4.2%)	23 (6.5%)	35 (5.4%)
	1 = no due to YOI	257 (89.2%)	309 (87.0%)	566 (88.0%)
	2 = yes	6 (2.1%)	9 (2.5%)	15 (2.3%)
82	i_addthe_discha	PA: after release, planned continuation of addiction therapy		
	-1 = not relevant or known	29 (10.1%)	36 (10.1%)	65 (10.1%)
	0 = not necessary	176 (61.1%)	184 (51.8%)	360 (56.0%)
	1 = necessary but not planned	34 (11.8%)	64 (18.0%)	98 (15.2%)

	2 = necessary and planned	49 (17.0%)	71 (20.0%)	120 (18.7%)
83	i_debt_need	PA: necessity for participation in in dept counselling		
	0 = no	124 (43.1%)	142 (40.0%)	266 (41.4%)
	1 = yes	164 (56.9%)	213 (60.0%)	377 (58.6%)
84	i_debt_beg	PA: began participation in dept counselling		
	-1 = not relevant or known	16 (5.6%)	13 (3.7%)	29 (4.5%)
	0 = no due to YO	13 (4.5%)	17 (4.8%)	30 (4.7%)
	1 = no due to YOI	140 (48.6%)	169 (47.6%)	309 (48.1%)
	2 = yes	119 (41.3%)	156 (43.9%)	275 (42.8%)
85	i_debt_discon	PA: discontinued dept counselling		
	-1 = not relevant or known	181 (62.8%)	215 (60.6%)	396 (61.6%)
	0 = no	95 (33.0%)	123 (34.6%)	218 (33.9%)
	1 = yes due to YOI	11 (3.8%)	14 (3.9%)	25 (3.9%)
	2 = yes due to YO	1 (0.3%)	3 (0.8%)	4 (0.6%)
86	i_debt_reach	PA: reached goals of dept counselling		
	-1 = not relevant or known	194 (67.4%)	242 (68.2%)	436 (67.8%)
	0 = no/ rudimentary1 = almost/ fully	20 (6.9%)	35 (9.9%)	55 (8.6%)
	1 = almost/ fully	74 (25.7%)	78 (22.0%)	152 (23.6%)
87	i_debt_discha	PA: after release, planned continuation of dept counselling		
	-1 = not relevant or known	55 (19.1%)	73 (20.6%)	128 (19.9%)
	0 = not necessary	176 (61.1%)	190 (53.5%)	366 (56.9%)
	1 = necessary but not planned	38 (13.2%)	65 (18.3%)	103 (16.0%)
	2 = necessary and planned	19 (6.6%)	27 (7.6%)	46 (7.2%)
88	i_soctra_need	PA: necessity for participation in social skills training		
	0 = no	111 (38.5%)	119 (33.5%)	230 (35.8%)
	1 = yes	177 (61.5%)	236 (66.5%)	413 (64.2%)
89	i_soctra_beg	PA: began participation in social skills training		
	-1 = not relevant or known	14 (4.9%)	21 (5.9%)	35 (5.4%)
	0 = no due to YO	21 (7.3%)	37 (10.4%)	58 (9.0%)
	1 = no due to YOI	125 (43.4%)	161 (45.4%)	286 (44.5%)
	2 = yes	128 (44.4%)	136 (38.3%)	264 (41.1%)

90	i_soctra_discon	PA: discontinued social skills training		
	-1 = not relevant or known	163 (56.6%)	223 (62.8%)	386 (60.0%)
	0 = no	113 (39.2%)	119 (33.5%)	232 (36.1%)
	1 = yes due to YOI	9 (3.1%)	8 (2.3%)	17 (2.6%)
	2 = yes due to YO	3 (1.0%)	5 (1.4%)	8 (1.2%)
91	i_soctra_reach	PA: reached goals of social skills training		
	-1 = not relevant or known	171 (59.4%)	226 (63.7%)	397 (61.7%)
	0 = no/ rudimentary1 = almost/ fully	24 (8.3%)	26 (7.3%)	50 (7.8%)
	1 = almost/ fully	93 (32.3%)	103 (29.0%)	196 (30.5%)
92	i_soctra_discha	PA: after release, planned continuation of social skills training		
	-1 = not relevant or known	29 (10.1%)	33 (9.3%)	62 (9.6%)
	0 = not necessary	200 (69.4%)	225 (63.4%)	425 (66.1%)
	1 = necessary but not planned	46 (16.0%)	85 (23.9%)	131 (20.4%)
	2 = necessary and planned	13 (4.5%)	12 (3.4%)	25 (3.9%)
93	i_socthe_need	PA: necessity for participation in social therapy		
	0 = no	244 (84.7%)	303 (85.4%)	547 (85.1%)
	1 = yes	44 (15.3%)	52 (14.6%)	96 (14.9%)
94	i_socthe_beg	PA: began participation in social therapy		
	-1 = not relevant or known	5 (1.7%)	4 (1.1%)	9 (1.4%)
	0 = no due to YO	5 (1.7%)	9 (2.5%)	14 (2.2%)
	1 = no due to YOI	246 (85.4%)	310 (87.3%)	556 (86.5%)
	2 = yes	32 (11.1%)	32 (9.0%)	64 (10.0%)
95	i_socthe_discon	PA: discontinued social therapy		
	-1 = not relevant or known	256 (88.9%)	324 (91.3%)	580 (90.2%)
	0 = no	26 (9.0%)	23 (6.5%)	49 (7.6%)
	1 = yes due to YOI	4 (1.4%)	4 (1.1%)	8 (1.2%)
	2 = yes due to YO	2 (0.7%)	4 (1.1%)	6 (0.9%)
96	i_socthe_reach	PA: reached goals of social therapy		
	-1 = not relevant or known	256 (88.9%)	324 (91.3%)	580 (90.2%)
	0 = no/ rudimentary1 = almost/ fully	10 (3.5%)	9 (2.5%)	19 (3.0%)

	1 = almost/ fully	22 (7.6%)	22 (6.2%)	44 (6.8%)
97	i_leis_need	PA: necessity for participation in structured pedagogic leisure measures		
	0 = no	102 (35.4%)	114 (32.1%)	216 (33.6%)
	1 = yes	186 (64.6%)	241 (67.9%)	427 (66.4%)
98	i_leis_beg	PA: began participation in structured pedagogic leisure measures		
	-1 = not relevant or known	13 (4.5%)	24 (6.8%)	37 (5.8%)
	0 = no due to YO	20 (6.9%)	41 (11.5%)	61 (9.5%)
	1 = no due to YOI	140 (48.6%)	150 (42.3%)	290 (45.1%)
	2 = yes	115 (39.9%)	140 (39.4%)	255 (39.7%)
99	i_leis_discon	PA: discontinued structured pedagogic leisure measures		
	-1 = not relevant or known	176 (61.1%)	222 (62.5%)	398 (61.9%)
	0 = no	91 (31.6%)	105 (29.6%)	196 (30.5%)
	1 = yes due to YOI	13 (4.5%)	12 (3.4%)	25 (3.9%)
	2 = yes due to YO	8 (2.8%)	16 (4.5%)	24 (3.7%)
100	i_leis_reach	PA: reached goals of structured pedagogic leisure measures		
	-1 = not relevant or known	185 (64.2%)	240 (67.6%)	425 (66.1%)
	0 = no/ rudimentary1 = almost/ fully	28 (9.7%)	28 (7.9%)	56 (8.7%)
	1 = almost/ fully	75 (26.0%)	87 (24.5%)	162 (25.2%)
101	i_leis_discha	PA: after release, planned continuation of structured pedagogic leisure measures		
	-1 = not relevant or known	34 (11.8%)	51 (14.4%)	85 (13.2%)
	0 = not necessary	196 (68.1%)	220 (62.0%)	416 (64.7%)
	1 = necessary but not planned	50 (17.4%)	67 (18.9%)	117 (18.2%)
	2 = necessary and planned	8 (2.8%)	17 (4.8%)	25 (3.9%)
102	i_mantran_need	PA: necessity for participation in structured transition management		
	0 = no	71 (24.7%)	50 (14.1%)	121 (18.8%)
	1 = yes	217 (75.3%)	305 (85.9%)	522 (81.2%)
103	i_mantran_beg	PA: began participation in structured transition management		
	-1 = not relevant or known	7 (2.4%)	17 (4.8%)	24 (3.7%)
	0 = no due to YO	16 (5.6%)	38 (10.7%)	54 (8.4%)

	1 = no due to YOI	129 (44.8%)	127 (35.8%)	256 (39.8%)
	2 = yes	136 (47.2%)	173 (48.7%)	309 (48.1%)
104	i_mantran_discon	PA: discontinued structured transition management		
	-1 = not relevant or known	157 (54.5%)	192 (54.1%)	349 (54.3%)
	0 = no	110 (38.2%)	143 (40.3%)	253 (39.3%)
	1 = yes due to YOI	10 (3.5%)	9 (2.5%)	19 (3.0%)
	2 = yes due to YO	11 (3.8%)	11 (3.1%)	22 (3.4%)
105	i_mantran_reach	PA: reached goals of structured transition management		
	-1 = not relevant or known	169 (58.7%)	202 (56.9%)	371 (57.7%)
	0 = no/ rudimentary1 = almost/ fully	26 (9.0%)	33 (9.3%)	59 (9.2%)
	1 = almost/ fully	93 (32.3%)	120 (33.8%)	213 (33.1%)
106	i_mantran_discha	PA: after release, planned continuation of structured transition management		
	-1 = not relevant or known	23 (8.0%)	49 (13.8%)	72 (11.2%)
	0 = not necessary	140 (48.6%)	125 (35.2%)	265 (41.2%)
	1 = necessary but not planned	58 (20.1%)	82 (23.1%)	140 (21.8%)
	2 = necessary and planned	67 (23.3%)	99 (27.9%)	166 (25.8%)
107	i_othertreat_need	PA: necessity for participation in other treatment		
	0 = no	82 (28.5%)	83 (23.4%)	165 (25.7%)
	1 = yes	206 (71.5%)	272 (76.6%)	478 (74.3%)
108	i_othertreat_beg	PA: began participation in other treatment		
	-1 = not relevant or known	14 (4.9%)	25 (7.0%)	39 (6.1%)
	0 = no due to YO	27 (9.4%)	43 (12.1%)	70 (10.9%)
	1 = no due to YOI	99 (34.4%)	109 (30.7%)	208 (32.3%)
	2 = yes	148 (51.4%)	178 (50.1%)	326 (50.7%)
109	i_othertreat_discon	PA: discontinued other treatment		
	-1 = not relevant or known	148 (51.4%)	179 (50.4%)	327 (50.9%)
	0 = no	127 (44.1%)	157 (44.2%)	284 (44.2%)
	1 = yes due to YOI	9 (3.1%)	11 (3.1%)	20 (3.1%)
	2 = yes due to YO	4 (1.4%)	8 (2.3%)	12 (1.9%)
110	i_othertreat_reach	PA: reached goals of other treatment		
	-1 = not relevant or	169 (58.7%)	194 (54.6%)	363 (56.5%)

	known			
	0 = no/ rudimentary1 = almost/ fully	16 (5.6%)	39 (11.0%)	55 (8.6%)
	1 = almost/ fully	103 (35.8%)	122 (34.4%)	225 (35.0%)
111	i_othertreat_discha	PA: after release, planned continuation of other treatment		
	-1 = not relevant or known	45 (15.6%)	59 (16.6%)	104 (16.2%)
	0 = not necessary	192 (66.7%)	216 (60.8%)	408 (63.5%)
	1 = necessary but not planned	37 (12.8%)	60 (16.9%)	97 (15.1%)
	2 = necessary and planned	14 (4.9%)	20 (5.6%)	34 (5.3%)
112	i_total.need	total number of intervention categories for which participation is necessary		
	mean (SD)	9.455 (2.663)	10.301 (2.427)	9.922 (2.568)
113	i_total.beg	total number of interventions categories that were began		
	mean (SD)	5.201 (2.583)	5.335 (2.391)	5.275 (2.477)
114	i_total.discon	total number of discontinued intervention categories		
	mean (SD)	0.628 (1.200)	0.642 (1.060)	0.636 (1.124)
115	i_total.reach	total number of intervention categories with goals met		
	mean (SD)	3.618 (2.688)	3.442 (2.531)	3.521 (2.602)
116	i_total.con	total number of planned continued intervention categories after release		
	mean (SD)	1.191 (1.342)	1.414 (1.564)	1.314 (1.472)
117	s_evalYOI	SE: evaluation of the time in YOI		
	mean (SD)	3.648 (0.464)	3.581 (0.471)	3.611 (0.469)
118	s_evalsit	SE: evaluation of the release situation		
	mean (SD)	4.038 (0.494)	3.934 (0.508)	3.980 (0.504)
119	s_selfp	SE: self-image		
	mean (SD)	4.054 (0.492)	4.013 (0.517)	4.031 (0.506)
120	s_chanrela	SE: changements in relationships during his time in YOI		
	mean (SD)	3.389 (0.780)	3.284 (0.830)	3.331 (0.809)
121	s_supsatis	SE: satisfaction about support he received in YOI		
	mean (SD)	3.312 (0.617)	3.324 (0.696)	3.319 (0.661)
122	s_reached	SE: what could he reach for himself in YOI		
	mean (SD)	1.942 (1.309)	1.859 (1.212)	1.896 (1.256)
123	s_eval	SE: evaluation of treatment offers in YOI		
	mean (SD)	3.692 (0.812)	3.682 (0.799)	3.686 (0.804)
124	s_feel	SE: feeling when he thinks about his life in general		
	mean (SD)	3.907 (0.742)	3.876 (0.840)	3.890 (0.797)

125	s_recidivism	SE: will commit crimes after release		
	mean (SD)	1.691 (0.825)	1.956 (0.994)	1.837 (0.931)
126	s_change	SE: evaluation of his changements in YOI		
	mean (SD)	4.307 (0.624)	4.284 (0.655)	4.294 (0.641)
127	s_knowprobass	SE: knows his probation assistant		
	0 = no	63 (21.9%)	94 (26.5%)	157 (24.4%)
	1 = yes	225 (78.1%)	261 (73.5%)	486 (75.6%)
128	s_contprobass	SE: in contact with his probation assistant during stay in YOI		
	0 = no	244 (84.7%)	293 (82.5%)	537 (83.5%)
	1 = yes	44 (15.3%)	62 (17.5%)	106 (16.5%)
129	s_dept	SE: has debts after release		
	0 = no	104 (36.1%)	114 (32.1%)	218 (33.9%)
	1 = yes	184 (63.9%)	241 (67.9%)	425 (66.1%)
130	s_worktrain	SE: after release, has work or vocational training		
	0 = no	65 (22.6%)	141 (39.7%)	206 (32.0%)
	1 = probably	123 (42.7%)	126 (35.5%)	249 (38.7%)
	2 = yes	100 (34.7%)	88 (24.8%)	188 (29.2%)
131	s_partner	SE: has a partner		
	0 = no	211 (73.3%)	260 (73.2%)	471 (73.3%)
	1 = yes	77 (26.7%)	95 (26.8%)	172 (26.7%)
132	s_viovic	SE: was victim of physical violence during his stay in YOI		
	0 = no	249 (86.5%)	289 (81.4%)	538 (83.7%)
	1 = yes	39 (13.5%)	66 (18.6%)	105 (16.3%)
133	s_vioseen	SE: has witnessed physical violence during his stay in YOI		
	0 = no	71 (24.7%)	84 (23.7%)	155 (24.1%)
	1 = yes	217 (75.3%)	271 (76.3%)	488 (75.9%)
134	s_viodone	SE: has been physically violent during his stay in YOI		
	0 = no	213 (74.0%)	238 (67.0%)	451 (70.1%)
	1 = yes	75 (26.0%)	117 (33.0%)	192 (29.9%)
135	s_consalc	SE: has consumed alcohol during his stay in YOI		
	0 = no	266 (92.4%)	313 (88.2%)	579 (90.0%)
	1 = yes	22 (7.6%)	42 (11.8%)	64 (10.0%)
136	s_conscann	SE: has consumed cannabis during his stay in YOI		
	0 = no	223 (77.4%)	245 (69.0%)	468 (72.8%)
	1 = yes	65 (22.6%)	110 (31.0%)	175 (27.2%)
137	s_consdrug	SE: has consumed other drugs during his stay in YOI		
	0 = no	237 (82.3%)	280 (78.9%)	517 (80.4%)
	1 = yes	51 (17.7%)	75 (21.1%)	126 (19.6%)

Note. SD: Standard Deviation; GCC: German criminal code; PA: professional assessment; SE: self-estimation; YO: Young Offender; YOI: Young Offender Institution.