Utrecht University

Predicting social competence of children

using the YOUth cohort study

A master's thesis

Name: Costian Arissen
Student number: 6644961
Programme: Applied Data Science
Supervisor: Dr. Paulina Pankowska
Second reader: Prof. dr. Albert Salah
Date: 16-7-2023

Table of contents

Section 1: Introduction 3
Section 2: Data 6
Section 2.1: About the data 6
Section 2.2: Data preparation 8
Section 2.3: Correlations 10
Section 3: Methods 12
Section 3.1: Method selection 12
Section 3.2: Further feature selection for each method 15
Section 4: Results & analysis 18
Section 4.1: Prosocial behaviour 18
Section 4.2: Peer problems 21
Section 4.3: Considerations about interpreting the results 24
Section 5: Conclusion & discussion 26
Section 5.1: Conclusion 26
Section 5.2: Discussion 27
Reference list 28
Appendix A: Datasets chosen for analysis 33
Appendix B: Minor considerations during pre-processing phase 34

Section 1: Introduction

An important human characteristic is the capacity to have meaningful interactions with other people (Onland-Moret et al., 2020). This skill is called social competence, and an alternative way to describe it is as the behavioural expression of one's emotional and regulatory competences when interacting with others around them (Junge et al., 2020). There is a variety of reasons why social competence is so important. First of all, possessing a decent level of social competence is necessary in order to become a proper participant of society, as well as to reduce the risk of developing emotional and behavioural problems in the future (Onland-Moret et al., 2020). Humans are social animals, and it is necessary to attain a set of social and communication skills in order to fulfil this need to socialize with others. One's social behaviour in adulthood also originates from the socialization process during their childhood, so a lack of proper social competence during childhood can distort the desire of socialization both during childhood and later in life (Taleipour & Motlaq, 2021). Secondly, previous literature has shown that differing levels of social competence are related to expertise in other areas, in both the present and the future. An example given is that children that find it easy to develop relationships with peers, have a better chance of becoming healthy adults, which in turn is associated with better mental health and being a well-functioning member of society (Junge et al., 2020). Also, social competence tends to be a solid predictor of social and academic success. Furthermore, research has shown that indices of social competence are predictors of this competence level later in life. For example, social competence aspects such as peer acceptance remain mostly the same across childhood. Hence, understanding a child's peer acceptance situation can aid in predicting the child's competence level at a later point during childhood (Amrei, Shafiri & Taheri, 2020; Junge et al., 2020; Blandon, Calkins & Keane, 2010). So, it is clear that there are advantages to having a solid level of social competence. On the other hand, worse levels of social competence are associated with a range of problems. For example, aggressive behaviour is usually observed in such a situation, as well as peer rejection. Interestingly, these factors also seem to be linked, as attempting to end conflicts by means of anger or aggressive behaviour is linked to peer rejection. On top of that, lacking social competence is correlated with social anxiety and bullying (Junge et al., 2020).

Given the importance of social competence as a human characteristic and the far-reaching implications it can have, many researchers have attempted to find out what leads to or affects social competence. Most studies on this topic seem to avoid using causal terms to describe the relationship between predictors and the outcome. A possible reason for this is that

confounding[1] is an important consideration in observational studies, and inferring causality from such studies with certainty is not an easy task. It requires a specific set of assumptions to be met, and there is still a possibility that unobserved confounders influence the relationship (Ananth & Schisterman, 2017; Imai et al., 2011). Most studies focus on predicting social competence using a predictor or set of predictors. Despite not aiming to identify causal relationships, such prediction studies still provide valuable information about the association between (a set of) variables and social competence. This can serve as an initial step in the process of understanding the mechanisms that lead to a certain level of social competence. Also, it is possible to shape someone's social competence level using interventions (Junge et al., 2020). Hence, being able to predict whether someone is at risk of having a lower social competence level could allow those interventions to be used to help them boost their competence level.

From the available studies, several predictors have been identified. In a preschool-focused study by Diener & Kim (2004), child-related characteristics such as age, temperament[2], and level of self-regulation[3], as well as maternal characteristics such as affection and separation anxiety, were identified as predictors of social competence. Furthermore, the interaction between a child's level of self-regulation and their susceptibility to anger is a predictor of prosocial behaviour, which is an important aspect of social competence in the YOUth cohort study (Diener & Kim, 2004; Onland-Moret et al., 2020). Other studies have similarly found that a child's characteristics and their parents' parenting styles are predictive of the child's level of social competence (Blandon, Calkins & Keane, 2010). However, this does not seem to always be the case, as for example the implications of a mother's level of parental control on the child's social competence level may differ depending on the child's externalizing behaviour[4] and weak ability to manage their emotions during the toddler phase (Blandon, Calkins & Keane, 2010). A different study puts more emphasis on the social environment and peer relations as predictors of social competence. In particular, those are the personal relationships between the child and their peers, the child and their teacher and the classroom climate. Peer relationships enable the child to practice social behaviour and develop social skills, while also satisfying their sense of belonging, and a good relationship with the teacher

[1] A confounder is defined as a variable associated with the predictor and outcome, that also occurs prior to the predictor (Assimon, 2021).
[2] Temperament is often described as the occurrence and intensity of negative emotions (Diener & Kim, 2004).
[3] Self-regulation refers to managing one's emotions, behaviour and impulses (Onland-Moret et al., 2020).
[4] Externalizing behaviour refers to anti-social behaviour that violates the social norms or behaviour that tends to be harmful towards other people (Kauten & Barry, 2020).

can encourage the development of social competence. The teacher is also responsible for the classroom climate, as they can regulate the structure and atmosphere in the group (De Swart et al., 2022). Finally, participating in after-school programs has been identified as a possible predictor of social competence, due to those programs providing an environment wherein students can engage with each other (Shernoff, 2010).

Overall, the prediction studies seem heavily focused on young children, which is understandable due to the far-reaching implications of social competence on the present and the future, as well as indices of social competence being predictive of one's social competence level in the future (Amrei, Shafiri & Taheri, 2020; Junge et al., 2020; Blandon, Calkins & Keane, 2010). In this thesis, a similar research will be performed using the YOUth cohort study data, which has been conducted in Utrecht and surrounding areas. The motivation behind that cohort is finding out how the development of social competence and self-regulation in children is shaped by the interactions between biological, psychological and environmental processes. Not much is known about that yet, as studies generally focus on one of these aspects at a time (Onland-Moret et al., 2020). Participants in the YOUth Baby & Child cohort are followed from pregnancy until around eight years old, and different types of data are obtained over the years (e.g., MRI scans, questionnaires, computer tasks, tests performed by a professional at the test centre). The data collection process in this study resulted in the accumulation of a large variety of data, which makes it a proper data source for the development of prediction models using these different biological, environmental and general child factors (Onland-Moret et al., 2020).

The overall goal of this thesis is to get a better understanding of what leads to different levels of social competence, as it is an important characteristic, and the implications can be far-reaching. From a data science perspective, an attempt will be made to find a subset of predictors in the YOUth cohort data that can best predict social competence of six-year-old children. As said before, it is hard to make accurate causal statements using observational data, but prediction studies by themselves still provide valuable information about the associations between (a set of) variables and social competence, which could be used in future studies to create causal hypotheses and test them empirically.

Section 2: Data

Section 2.1: About the data

The data made available for this thesis consist of 104 questionnaires, 216 supplements and the Peabody-Picture vocabulary task, which is a common computer task that is used for evaluating a child's vocabulary size with respect to their age (Onland-Moret et al., 2020). YOUth also has other types of data, such as MRI scans, but those are not available for this project. Regardless, the available data is still quite varied, as the questionnaires cover many different topics. A few examples would be the child's behaviour, social skills, and hobbies, as well as the child's biological information and parents' characteristics. The questionnaires are split into multiple age groups from birth up to around six years old. After the first year the age measurements will not be exact, and will be depicted as "around three years old" and "around six years old". The six-year-olds in particular are the focus of this thesis. The dataset that is used to measure the social competence level of that group is the Dutch version of the *Strengths and Difficulties Questionnaire (SDQ)* (Onland-Moret et al., 2020). Screening instruments such as this one have an important place in child mental health care and research, as they try to measure the types and severities of psychosocial problems and strengths (Stone et al., 2015). Even though a common SDQ questionnaire consists of five sub-questions, only the subscales "prosocial behaviour" and "peer problems" have been used in the YOUth cohort study (Onland-Moret et al., 2020; Youth in Mind, n.d.). For each sub-question, the possible answers are "not true" (1), "somewhat true" (2) and "definitely true" (3). Table 1 shows the exact sub-questions. Depending on the positive or negative connotation of the question, the results of each sub-question within a subscale are added up or subtracted to form final subscale scores. Whether the two separate subscales are meant to be combined is unsure; commonly the scores of the "prosocial behaviour" scale are not combined with the other subscales (Bøe et al., 2016). For that reason, these two subscales will be analysed separately from each other.

Prosocial behaviour	**Peer problems**
Considerate of other people's feelings	Rather solitary and tends to play alone
Shares easily with other children	Has at least one good friend
Helpful when someone is hurt, upset or sick	Usually liked by other children
Kind to younger children	Picked on or bullied by other children
Commonly volunteers to help others	Can get along better with adults than peers

Table 1: sub-questions of the SDQ questionnaire

A group of 17 different datasets focused on six-year-old children has been selected for prediction purposes. This collection contains a variety of data, and just like in the overall YOUth data, it contains child-specific, environmental and biological data. For the description of each set, please see Appendix A. The key consideration for choosing these datasets specifically is data missingness, which is a problem that may result in an array of issues, particularly the reduction in performance, data analysis issues and obtaining biased model outcomes (Emmanuel et al., 2021). Of the 320 available datasets, 264 (82.5%) have zero comparable subjects with the outcome dataset, rendering them unusable. Of the remaining sets, the average percentage of intersection is 96% for the ones that focus on six-year-olds, outside of a few outlier supplements. This selection process serves as a way to at least avoid having too many completely missing rows for questionnaires and supplements at the start of the study already, and especially because datasets may have various amounts of missingness already. Regardless of this choice, there is still a variety of datasets to be used for prediction.

It is important to note that the YOUth data is sensitive, and access to the data requires explicit permission by the owners of the data, as well as a login code to an external environment named SANE. For ethical reasons, the individual subjects will not be mentioned or shown at all in this document. The figures and tables show aggregate information, and cannot be traced back to the individual.

Section 2.2: Data preparation

Each dataset will go through a series of pre-processing steps. The most important considerations are discussed here, and the less important ones (e.g., dataset format, dataset-specific metadata) can be found in Appendix B. Note that these three major considerations will be applied after each other in the pre-processing phase, as the first two focus on retrieving usable variables, whereas the third one focuses on filling in the missing values of variables obtained in the previous two steps.

The first major consideration is missingness, which was already discussed in the previous section as a problem that may result in different problems (Emmanuel et al., 2021). As simple imputation techniques like median and mode imputation will be used to tackle this missingness problem, the mean and deviation of the variables may be biased, and correlations can be affected (Zhang, 2016). Especially once the amount of missingness for a variable gets higher, these imputation methods will heavily affect the variable's usefulness. Hence, the decision has been made to discard variables with more than 25% missingness. This number is a bit arbitrary, but it seems already quite generous when some studies suggest to only use it when the percentage of missingness is below 10% for a column (Tsikriktsis, 2005). As will be discussed soon, almost every resulting variable from the pre-processing process has less than 10% missingness, so the imputation techniques will not be too invasive.

The second major consideration is zero-variance and near-zero-variance variables. A zero-variance variable is by definition not helpful for prediction tasks whatsoever, and can actually cause many models to outright fail. On the other hand, having near-zero-variance variables may result in issues when using sampling techniques, such as cross-validation, as an unlucky sample may result in such a variable ending up with zero variation in one of the samples (Kuhn, 2008). Variables that show zero- or near-zero variance will be discarded, and Kuhn's criteria will be used to determine whether a variable belongs to the latter. He suggested that a predictor may be near-zero variance if the percentage of unique values in the column is less than 20%, and the ratio of the most to second most frequent value is more than 20 (Kuhn, 2008). An example would be a feature with 100 rows, of which 98 have the value X, and 2 have the value Y. There are almost no unique values, and the ratio of the most to second most frequent value is 98 / 2 = 49. There is a real likelihood that a sample of this variable will show zero variance, as it may include only the value X.

The final essential consideration is imputation, which was already briefly discussed in the first major pre-processing step. The imputation process involves replacing the missing values of the categorical variables with the mode, and the numeric variables with the median. An advantage of these methods is that they can be applied quite easily, but a notable downside is that they can bias the mean and deviation of the variable, and affect correlations with other variables (Zhang, 2016). However, the data selection process above ensured that the majority of the variables remaining has a low percentage of missingness. The previous two steps result in a final dataset of 297 variables, of which 233 (78%) have 5% missingness or less, and 286 (96%) have 10% missingness or less. Hence, these imputation techniques should not be too invasive. Note that the imputation process may generate some near-zero-variance variables, so the variable pool shrinks slightly as those are discarded.

An intermediate step between the second and third consideration is setting aside a test set consisting of 20% of the data. This allows the estimated generalization performances obtained by the 5-fold cross-validation process to then be compared to this unseen test set (Xu & Goodacre, 2018). Hence, every decision from this point onward is only based on this 80% of the dataset, to ensure that the test set has not been used for any decision making, and is only used to find out how well the models obtained by cross-validation predict truly unseen data. The reason for doing this splitting prior to the third step is to avoid the test set getting contaminated by other data during the imputation process.

Section 2.3: Correlations

The final data consists of 20 factor variables and 271 numeric variables, of which the latter can be split further into 263 ordinal and 8 continuous columns. The factor variables are mainly binary questions, such as whether the child uses medicines or whether they participate in sports. The numeric variables are generally Likert scales with different ranges. An example would be whether the child is able to understand social cues, where the options vary between "completely true" and "completely untrue". Due to this large amount of variables, it is not practical to show correlation matrices and plots. Therefore, the statistics will be discussed without figures, which should still give a clear idea about the correlations.

Using the common interpretation of Spearman's correlation coefficient (Akoglu, 2018), the first thing to point out is that almost all predictor variables have a weak correlation ($\rho < 0.4$) with the two outcome variables. Only 10 variables have a moderate correlation ($\rho \geq 0.4$) with the prosocial behaviour outcome, and only one has a moderate correlation with the peer problems outcome. Even though the correlation values between the predictors and outcomes are low on average, many are still significant. The prosocial behaviour outcome variable is significantly correlated with 102 predictors, and that number is 49 for the peer problems outcome.

The associations between the categorical predictors and outcome variables are computed using the Chi-square test of independence. This test is useful when dealing with categorical data (McHugh, 2013). Here it will be used to find out whether the outcome variable differs significantly for the different categories of the categorical predictors. For the prosocial behaviour outcome, only two Chi-square tests turn out to be significant, and even zero for the peer problems outcome.

The features that have a significant correlation with an outcome variable will form the basis for the next step in the variable selection process, which will be discussed in the next section. However, some variables will be manually removed from this feature list, as they are directly included in the outcomes (e.g., the question "my kid is being bullied" should not be used as a predictor of the peer problems outcome, as it is one of the five sub-questions). This could have been done in the previous section already, but the correlation process has already shrunk the sets of data considerably, and shines a light on these variables. Hence, it is making it easier to notice such variables and remove them.

Disregarding the outcome variables, the correlation matrices can also be used to check whether the possible predictor variables have a strong correlation with each other, where strong is being defined as 0.7 and above (Akoglu, 2018). Even though this is a prediction study and multicollinearity does not have an impact on that like it does in inference studies (Paul, 2006), it is still useful to remove one feature from every highly correlated predictor pair for redundancy reasons. Moreover, such feature redundancy can result in a dilution of the importance scores assigned to every variable in the recursive feature elimination process. Hence, Kuhn & Johnson (2019) advice to get rid of highly correlated variables before starting the recursive feature elimination procedure, which is a technique that is discussed in the next section.

Section 3: Methods

Section 3.1: Method selection

Three different model types will be used for analysis, which will be linear regression, random forest and gradient boosting. For each model, it is useful to give a general overview and how it may be useful for this project, after which some more specific details and considerations are discussed.

Linear regression aims to find a set of weights that together with the inputs results in the best estimation of the relationship between the inputs and output (Chen & Gu, 2019). Some advantages are that it is easy to apply and understand, as well as having interpretable model coefficients. It does assume linear relationships between the inputs and output, so if a linear combination of variables turns out to be the best performing model in the best subset selection process, then the coefficient interpretability aspect is very beneficial for discussion purposes (Ray, 2019; Chen & Gu, 2019).

Whether a model is the best is commonly defined by it resulting in the lowest possible mean squared error (MSE), which is a metric that determines how close the predictions on average are from the real values. The weights associated with that best model can then be found by using optimization techniques like stochastic gradient descent (Chen & Gu, 2019). Gradient descent methods attempt to find the minimum of a cost function iteratively, by moving down the slope of the cost function derivative until the minimum is reached (Haji & Abdulazeez, 2021). Another common but different option in linear modelling is to use ordinary least squares (OLS) to estimate the optimal model coefficients, in which case the cost function is minimized by taking the partial derivatives of that cost function and making them equal to zero (Foley, 2022; Emerick, 2011). As was stated before, linear regression assumes that there is a linear relationship between the outcome and the predictors (Ray, 2019; Chen & Gu, 2019). Hence, this ought to be checked. However, as there are a lot of variables, checking linearity assumptions becomes a time-consuming process. Therefore, the decision has been made to use linear regression, as it is still possible that a few variables showing a linear relationship with the outcome end up as the best performing model. Different studies have shown that the more complex machine learning prediction models generate more accurate predictions on average, but in general it is still a useful idea to start off with linear regression before moving on to more complex techniques (Ray, 2019; Chen & Gu, 2019).

The other two models are based on decision trees, which are called that way because of the tree-like structure. You start from the root of the tree, and move downward by going through a series of split decisions, that lead you into certain branches of the tree. Eventually you reach a leaf node, which represents a certain predicted classification or value (Ali et al., 2012).

Random forests for regression are a collection of many of those trees, where each tree is created using bootstrapping (i.e., generating random samples of subjects with replacement) and a random sample of features, after which the predicted outcome values of the individual trees are averaged (Livingston, 2005; Breiman, 2001). This technique has quite some advantages. First of all, due to combining many different decision trees, and using bootstrapping to randomly select subjects for each tree, the forests are able to largely avoid overfitting (Ali et al., 2012). Also, the random selection of features for each tree makes it quite robust to outliers and noise (Breiman, 2001). Random forests results are comparable with other tree-based techniques (e.g., boosting), and turn out to be robust and highly accurate given that there are enough trees such that all the predictor variables have a chance to be included in the model (Breiman, 2001; Hegelich, 2016). Also, random forests are able to find non-linear relationships and interactions, although it is not exactly clear yet whether it is possible to separate marginal effects and interaction effects in the model results (Hatami et al., 2023; Wright, Ziegler & König, 2016). It is a well performing machine learning model, despite having almost no hyperparameters, and the default values of those hyperparameters seem to be performing remarkably well already (Bentéjac, Csörgö & Martínez-Muñoz, 2020). The model is useful for this thesis, as it provides a range of advantages with little hyperparameter tuning required, and does not require manually dealing with outliers in the variables.

However, it is important to note that the outcome of a random forest model is quite dependent on the amount of variables that are randomly sampled at each tree split (Hegelich, 2016). Hence, during the model training process in this thesis, a few different values of that hyperparameter will be attempted and 5-fold cross-validation will find out which one gives the best overall predictions.

An alternative popular tree-based method is gradient boosting. This sequential technique relies on iteratively building strong learner decision trees based on very simple weak learner decision trees, which can be defined as a trees that only perform marginally better than random guessing (Bentéjac, Csörgö & Martínez-Muñoz, 2020). The literal translation of the term gradient boosting is improving the error, which becomes clearer after seeing the process.

In the first tree the outcome is predicted using all the intended inputs, which can be seen as Y = f(X) + error. Then the next tree will predict the error from the first tree using the same inputs, which can be written as error = g(X) + error2. The initial prediction then becomes Y = f(X) + g(X) + error2, so the second tree improves the first. This process continues until you have reached the amount of specified trees in the model (Ayyadevara, 2018). Using gradient boosting with regression trees tends to result in models that are interpretable, competitive with other machine learning models and very robust (Friedman, 2001). That statement is backed by its success in the machine learning world and Kaggle competitions (Natekin & Knoll, 2013; Bentéjac, Csörgö & Martínez-Muñoz, 2020). Hence, it is useful to perform gradient boosting for this project as well.

An important distinction between gradient boosting and random forests is that the former has a lot more parameters to tune, which may make it a more extensive process compared to random forests with just a few hyperparameters to tune (Golden, Rothrock & Mishra, 2019). A particularly important one is the learning rate, which refers to how fast you intend to improve the model. A low learning rate means the model takes many small steps to learn, which ensures that the negative effect of a faulty tree on the prediction can still be corrected by future trees. However, a low learning rate results in a longer process, so it is a trade-off between computational cost and generalization capabilities. While gradient boosting has more hyperparameters to tune, this flexibility gives researchers a lot of freedom to decide what direction they want to take with respect to model tuning (Natekin & Knoll, 2013).

Section 3.2: Further feature selection for each method

Even though the correlation analysis already did some feature selection by selecting variables that have a significant correlation with an outcome variable, as well as removing a possible predictor if it is highly correlated with other predictors, there is still a large variety of variables that could serve as potential predictors in the models. The best subset selection approach that iterates over every possible combination of features is virtually impossible to perform with this many features. There is a variety of feature selection methods that can efficiently reduce the data. Even though there is no such thing as a perfect method, there are some promising techniques available (Jović, Brkić & Bogunović, 2015). For linear regression, the Least Absolute Shrinkage and Selection Operation (LASSO) technique will be used, which is a regularized linear regression method that shrinks coefficients of less important variables to zero, effectively removing them, so it has a built-in feature selection algorithm (Muthukrishnan & Rohini, 2016). LASSO works by minimizing the penalized residual sum of squares (RSS), which can be defined as the RSS plus the sum of absolute regression coefficients multiplied by the penalty tuning parameter lambda. Values of lambda of higher than one will increase the penalty, resulting in more variables shrinking to zero. The ideal value of that value is usually determined by means of cross-validation (Ranstam & Cook, 2018; Columbia University Irving Medical Center, n.d.). Some advantages of LASSO are that it is able improve the prediction accuracy and that it is easily interpretable. Less relevant for this thesis is that it can also tackle highly correlated predictors by shrinking one of the coefficients of the pair to zero, as this was already manually done in the correlation section. In general, LASSO tends to be a solid alternative to other feature selection techniques (Muthukrishnan & Rohini, 2016).

As can be seen in Figure 1, LASSO using 5-fold cross-validation shrinks many coefficients of both models to zero, so it has indeed managed to considerably reduce the amount of features that will be entered into the best subset selection algorithm. Respectively 17 and 18 variables are selected for the prosocial behaviour and peer problems models.

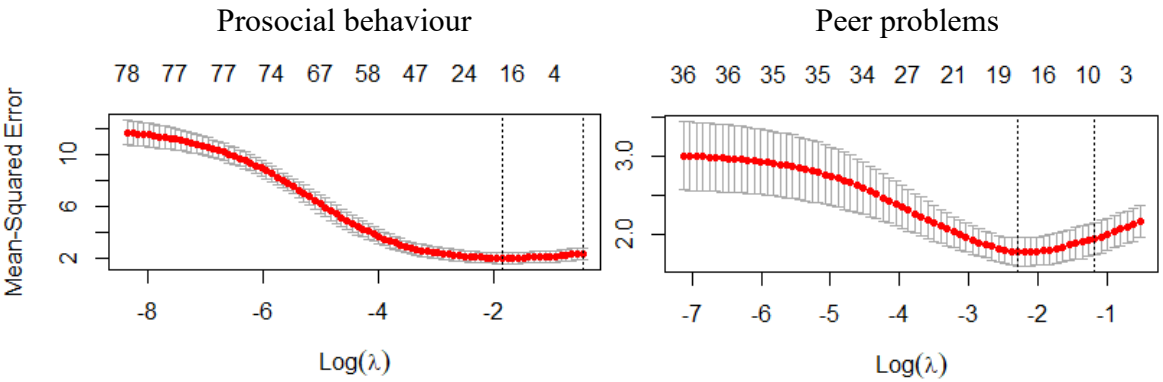

Figure 1: LASSO process results

For some more complex models (e.g., tree models) that can capture non-linearity and interactions, it might be better to use a method that is able to capture those things, as opposed to the quite restrictive LASSO. Some of the best performing overall subset selection techniques are greedy stepwise wrappers (Jović, Brkić & Bogunović, 2015). One such technique is recursive feature elimination (RFE), which is a commonly used technique that has proven to be effective and efficient at performing feature selection (Chen et al., 2018; Kuhn & Johnson, 2019). It is a relatively straightforward iterative process, as it starts by building a model using the entire predictor set, and then moves on to calculating the importance of each of those predictors using an importance score. The variable with the lowest score is then dropped, and the same process is repeated until the end (Granitto et al., 2006). Such methods are often used with random forest models, as they have a popular built-in feature importance method (Kuhn & Johnson, 2019). The variable importance of a variable as defined by Breiman (2001) describes the difference in prediction error of the model with the full set of variables and prediction error of the model without the variable. The key implication is that the removal of an informative variable will increase the prediction error (Lu & Ishwaran, 2020).

Table 2 shows the exact results of the 5-fold cross-validated RFE process. It is useful to read it from bottom to top, as the process initially starts using all variables, and then iteratively removes variables. Interestingly, for both the prosocial behaviour and peer problems outcomes, the RFE process will select all the variables to be included in the model, as those models are associated with the lowest RMSE. Why this exactly happens is unclear, but a possible hypothesis could be that almost all correlations between an outcome and the predictors are weak, and therefore the RFE model cannot pinpoint a small subset that best predicts the outcome. In an alternative scenario where we would have not removed the

predictors that were already included in the outcome variable (e.g., being bullied), which had moderate Spearman correlation values, the RFE process would select only a small handful of variables. The scenario right now is impractical, as the goal is to find a small subset of variables to be used in the best subset selection process. Hence, the 15 most influential features for each outcome variable will be saved to be used in the best subset selection part. This is somewhat arbitrary, but the main reason for this is that increasing this number even slightly will increase the necessary computation time in the next step by a lot.

Prosocial behaviour		**Peer problems**	
Number of variables	**RMSE**	**Number of variables**	**RMSE**
4	1.449	4	1.383
8	1.411	8	1.362
16	1.409	16	1.284
78	1.311	37	1.239

Table 2: RFE process results

By now, the number of possible predictors has shrunk considerably. A best subset algorithm will be applied for the three different model types, in order to find a subset with the best estimation of the out-of-sample prediction error. This process will do a 5-fold cross-validated model train loop over every combination of one, two, three and four features in the subsets obtained by the previous steps. The motivation for not using more than four features is mainly for time and computational considerations. The process is already quite intensive, as it has to perform 5-fold cross-validated model training, using a different range of hyperparameters, for three different model types, and using four different subset sizes. Increasing the subset size would be very time-consuming. However, as will be seen in the next section, the speed at which the cross-validated RMSE reduces will level off quite fast at about three to four variables used. The best models from the best subset process will be analysed.

Section 4: Results & analysis

Section 4.1: Prosocial behaviour

Table 3 shows the cross-validated RMSE values for the best model of each combination of model type and subset size. Each model type will also be evaluated using just the prosocial behaviour variable mean as predictor, of which the RMSE values will be used as baseline. This is useful, as you can see how the model performs using the most basic specification, and it serves as a baseline to which more complex models can be compared (Fugard, 2022). Each cross-validated RMSE value also has a number within parentheses behind it, indicating the proportion of the baseline RMSE. For the prosocial behaviour models, the random forests outperform the linear regression models overall. It is possible that there are some non-linear relationships or interactions that are not captured by the linear model, whereas random forests are able to capture non-linearity and interactions (Hatami et al., 2023; Wright, Ziegler & König, 2016). The random forest models also outperform the gradient boosting models with respect to the predictive capacity. This could be due to using the default range of hyperparameters as proposed by the writers of the *caret* package, as gradient boosting may require some more tuning than random forests (Golden, Rothrock & Mishra, 2019). However, *caret* still tries a range of hyperparameters, but it may not have been enough. An alternative possibility is that unlike random forests, gradient boosting can be affected by outliers (Breiman, 2001; Li & Bradic, 2018). Interesting to note is that the speed at which the RMSE decreases starts stabilizing for each model type at around three and four variables used, although we have no information about what happens for larger subsets. The best performing model is a random forest with four variables, which will now be discussed in depth.

Linear regression	
Count	**RMSE**
Mean	1.50 (1)
1	1.38 (0.92)
2	1.30 (0.87)
3	1.27 (0.85)
4	1.24 (0.83)

Random forest	
Count	**RMSE**
Mean	1.50 (1)
1	1.36 (0.91)
2	1.32 (0.88)
3	1.23 (0.82)
4	1.19 (0.79)

Gradient boosting	
Count	**RMSE**
Mean	1.50 (1)
1	1.41 (0.94)
2	1.35 (0.90)
3	1.28 (0.85)
4	1.28 (0.85)

Table 3: Best subset selection results for each model type and number of variables (the numbers within the parentheses indicate the proportion of the baseline RMSE value)

The table on this page and the figures on the next page provide details about the model with the best predictive capacity from the best subset selection process. Table 4 shows the descriptions of the variables, after which Figure 2 and Figure 3 offer details about the variable importance[5] and partial dependence[6] respectively. The most important prediction variables are the "LS5" questions, which are sub-questions of the self-regulation questionnaire, and higher levels of self-regulation have been found to be associated with higher levels of prosocial behaviour in children (Diener & Kim, 2004). The partial dependence plots for these self-regulation variables also depict a positive marginal association, which is in line with this article. The other two variables are a bit harder to describe, as it is challenging to find articles that specifically relate prosocial behaviour to these concepts. An older article by Eisenberg et al. (1996) claims that there tends to be a positive association between prosocial behaviour and socially accepted behaviour for children. One could make a case that saying bye to others and responding to others saying bye, as well as having a versatile character, are aspects of socially accepted behaviour. However, this explanation is not particularly strong, and is based on old literature and an assumption. Just like in Eisenberg et al. (1996), the respective conditional dependence plots show a positive marginal association. Looking at the variable importance plot, it seems that these variables do not account for a large part of the predictive quality, as it is mainly the self-regulation questions that give the model its predictive capacity.

Variable code	**Meaning**	**RMSE$_{Test}$**
LS5_6_Q20_SC	"Can easily stop when he/she is told 'no'"	1.28
CONV_SKILL_GREEL_LEAV_SC	"Says bye to others and reacts to others saying bye"	
LS5_6_Q10_SC	"Prepares for trips by thinking what is needed"	
SC1_14_SC	"Versatile"	

Table 4: Variable meaning and test set RMSE of the best performing random forest model

[5] This is Breiman's (2001) previously discussed variable importance, which is the difference in prediction error of the model with the full set of variables and prediction error of the model without the variable (Lu & Ishwaran, 2020). The R package *randomForest* also uses this method (Breiman et al., 2022).
[6] Partial dependence plots depict the marginal effect of a variable (i.e., holding other features constant) on the prediction, which is useful for showing the direction and type of relationship (Molnar, 2023).

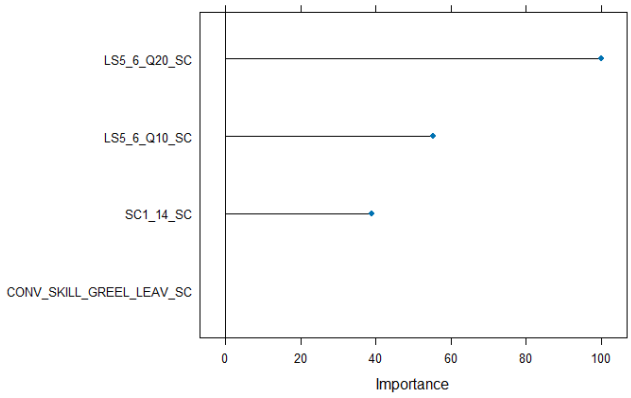

Figure 2: Variable importance of the best performing random forest model

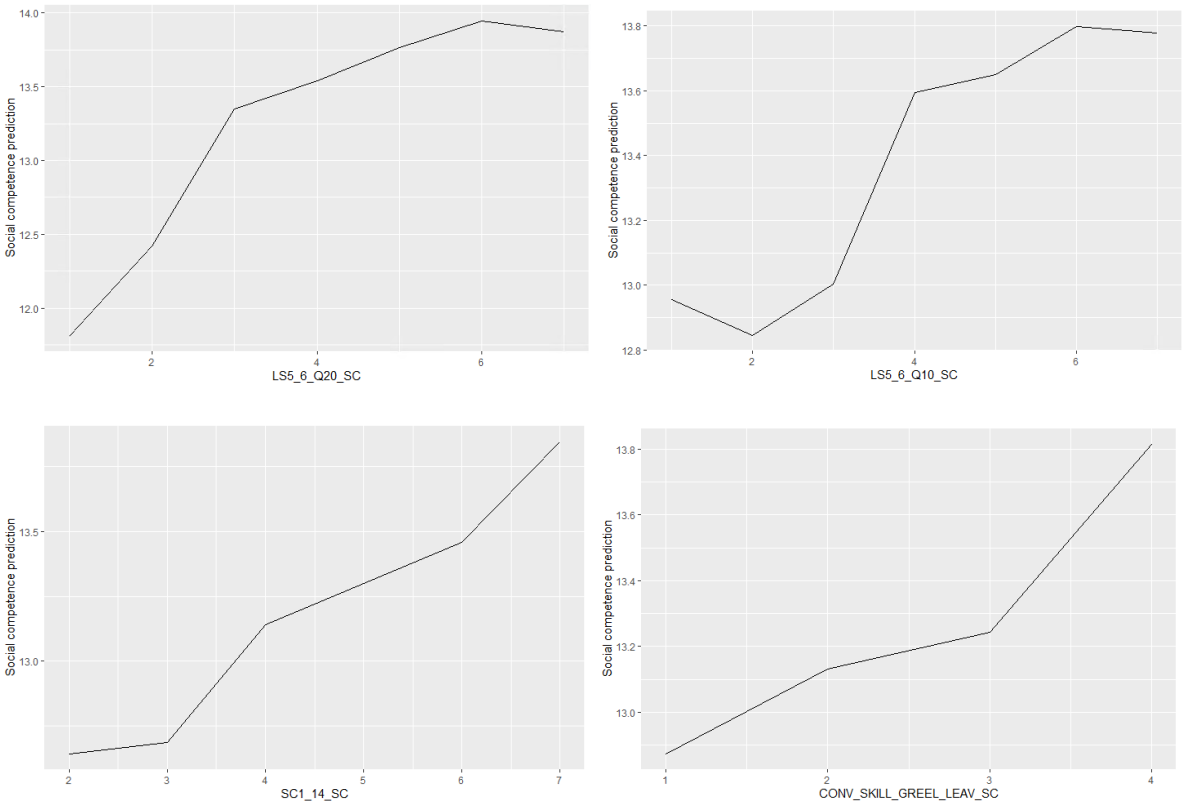

Figure 3: Partial dependence plots of the best performing random forest model

Looking back at Table 3, it shows that the best performing random forest model has a cross-validated RMSE value of 1.19, which is only 21% lower than just using the mean as predictor, so four different variables only resulted in a small increase in predictive capacity. The model does have roughly the same performance on the unused test set as it did on the validation sets, as is shown in Table 4. This RMSE of 1.28 is not unexpected, as the folds differ a bit from each other due to the resampling process, and the resampled validation RMSE values varied between 1.04 and 1.42.

Section 4.2: Peer problems

This section is structured the same way as the previous one, and the figures show similar concepts, except here the best subset selection process and specifics of the peer problems outcome models are discussed. Table 5 shows the results of the best subset selection process for the peer problems outcome models, and interestingly some observations are similar to the prosocial behaviour models from Table 3. The random forest models outperform the other two, and therefore the reasons could be similar to the ones mentioned in the previous chapter. It may outperform the linear regression models due to its capability of incorporating non-linearity and interactions (Hatami et al., 2023; Wright, Ziegler & König, 2016). As to why it outperforms the gradient boosting models could be due to a lack of thorough hyperparameter tuning for this hyperparameter-sensitive model, or due to being more affected by outliers than random forests (Golden, Rothrock & Mishra, 2019; Breiman, 2001; Li & Bradic, 2018). Just like for the prosocial behaviour models, the speed at which the RMSE seems to level off as more variables are added, with only marginal decreases in RMSE between three-variable and four-variable subsets. The best performing model is once again a random forest with four variables, which will be discussed in depth.

Linear regression	
Count	**RMSE**
Mean	1.46 (1)
1	1.36 (0.93)
2	1.28 (0.88)
3	1.23 (0.84)
4	1.19 (0.82)

Random forest	
Count	**RMSE**
Mean	1.46 (1)
1	1.27 (0.87)
2	1.13 (0.77)
3	1.11 (0.76)
4	1.10 (0.75)

Gradient boosting	
Count	**RMSE**
Mean	1.46 (1)
1	1.36 (0.93)
2	1.23 (0.84)
3	1.16 (0.79)
4	1.14 (0.78)

Table 5: Best subset selection results for each model type and number of variables (the numbers within the parentheses indicate the proportion of the baseline RMSE value)

The table and figures on the next two pages provide details about the model with the best predictive capacity from the best subset selection process. Table 6 shows the descriptions of the variables, after which Figure 4 and Figure 5 offer details about the variable importance and partial dependence respectively. The most influential predictor in this model by far is irritability, which is understandable, as children with a higher level of irritability tend to show mood and social adaptation disorders. They tend to be aggressive and avoidant, and may ignore social order (Zhang et al., 2020; Vidal-Ribas et al., 2016). Aggression in particular is

also positively associated with peer rejection (Yue & Zhang, 2023). The positive direction of the marginal dependence plot for irritability (i.e., higher irritability associated with higher level of peer problems) also seems to be in line with the articles that were just mentioned. The second most influential variable is the "LS5" question, which was surprisingly also a predictor in the prosocial behaviour model. As was discussed in the previous section, these "LS5" questions are part of the self-regulation dataset, and having a lower level of self-regulation during early childhood may result in peer problems (Saraç, Abanoz & Gülay Ogelman, 2021). Interestingly, the marginal dependence plot shows that children with higher levels of self-regulation have higher levels of peer problems, holding the other variables constant. This is remarkable, and there is no clear explanation for it. As for understanding explicit and implicit social rules, it has been shown that children that are unable to follow such rules, tend to behave unpredictably and may act unpleasantly during activities (Fabiano, Vujnovic & Pariseau, 2010). The conditional dependence plot is in line with this. The final variable about bullying other children seems harder to motivate, as literature linking the predictor and outcome has not been identified. However, this particular variable seems to have very low predictive capacity in this model.

Variable code	**Meaning**	**$RMSE_{Test}$**
SC1_2_SC	"Irritable"	
LS5_6_Q10_SC	"Prepares for trips by thinking what is needed"	
NON_VERB_USE_EXPL_IMPLIC	"Can understand explicit and implicit rules at school and in the social environment"	1.16
CHILD_DID_2_SC	"Bullied other kids by offending or laughing at them"	

Table 6: Variable meaning and test set RMSE of the best performing random forest model

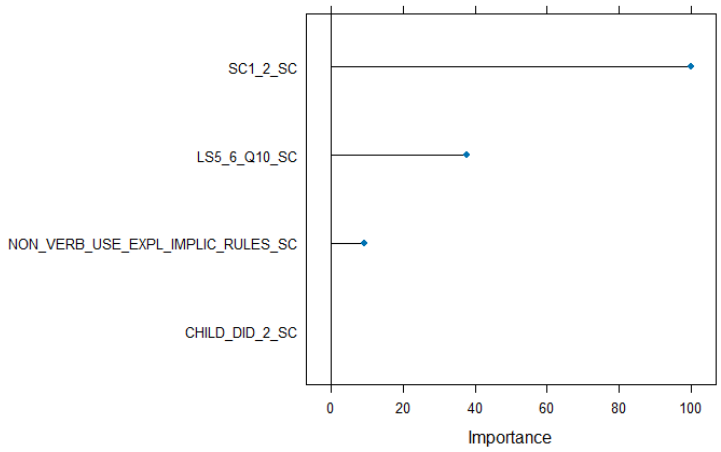

Figure 4: Variable importance of the best performing random forest model

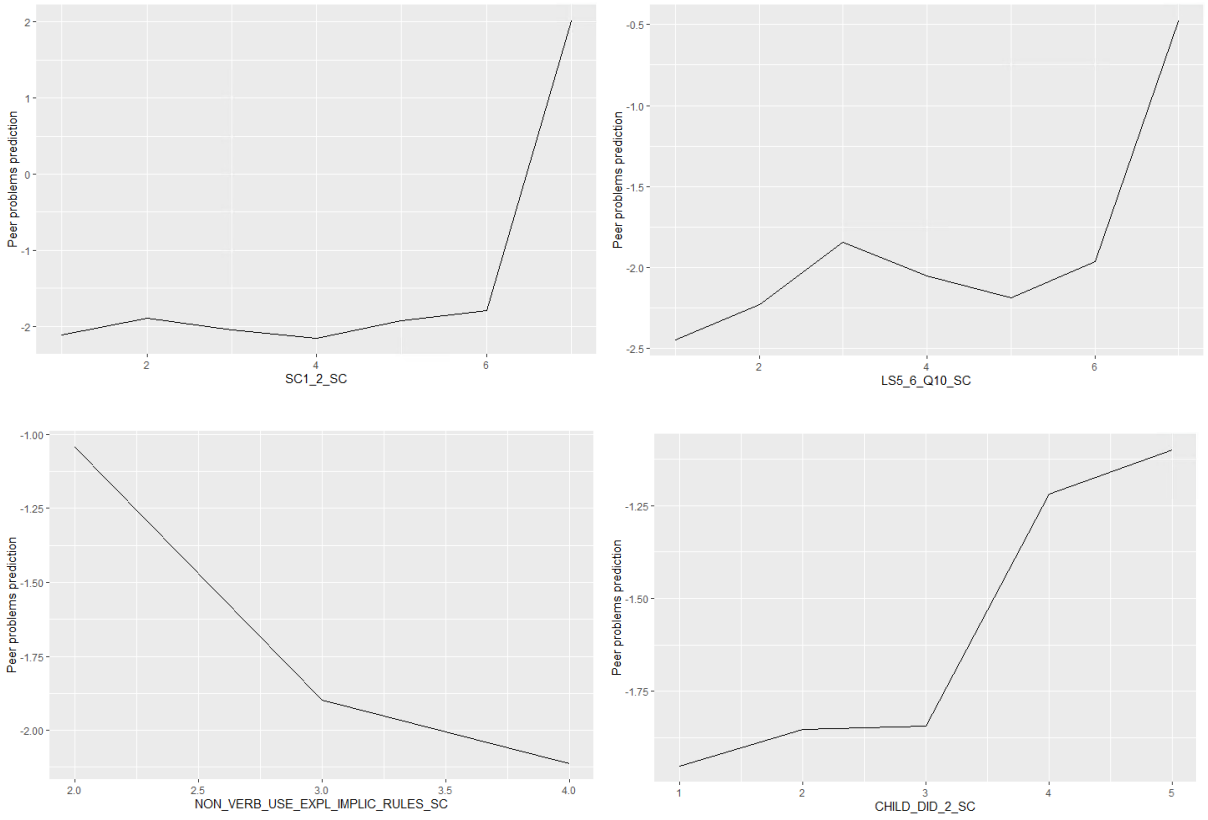

Figure 5: Partial dependence plots of the best performing random forest model

Looking back at Table 5, it shows that the model we have been discussing so far has a cross-validated RMSE value of 1.10, which is 25% lower than just using the mean as predictor. So again, four variables only result in a small increase in predictive capacity. The performance on the test set of 1.16 as shown in Table 6 is similar to the cross-validated RMSE value of 1.10. Again, this is not unexpected, as the folds differ a bit from each other due to the resampling process, which in this case results in resampled RMSE values between 0.98 and 1.30.

Section 4.3: Considerations about interpreting the results

This section will discuss some aspects of the data and the thesis decisions that may have led to the results from the previous sections, with the intention to show that the outcomes may be based on a certain collection of data specifications and research choices.

It is important to consider the distribution of the outcome variables, as shown in Figure 6. Prosocial behaviour can vary between 0 and 15, but the boxplot shows that the majority of the values is situated between 13 and 15. A similar observation happens within the peer problems outcome, as it can vary between -6 and 9, but most are found between -3 and -1. This means that on average, the children in this study show a high level of prosocial behaviour and low level of peer problems. A question could be raised whether this distribution resembles the real population of Dutch six-year-old children, as this dataset only consists of 152 participants, that specifically participated in the cohort study, and are from the province of Utrecht. Hence, there are doubts about the generalizability of the results. There may even be doubts about the generalizability to the YOUth cohort, as only 152 out of the around 7000 participants in the YOUth Baby & Child cohort actually filled in the outcome questionnaire (SDQ). It is no surprise that the participant pool shrinks over time in cohort studies, but there may be other reasons why there are so little participants in the SDQ questionnaire. Perhaps the questions are sensitive and parents refuse to let the child participate. If that is the case, the data could be missing not at random.

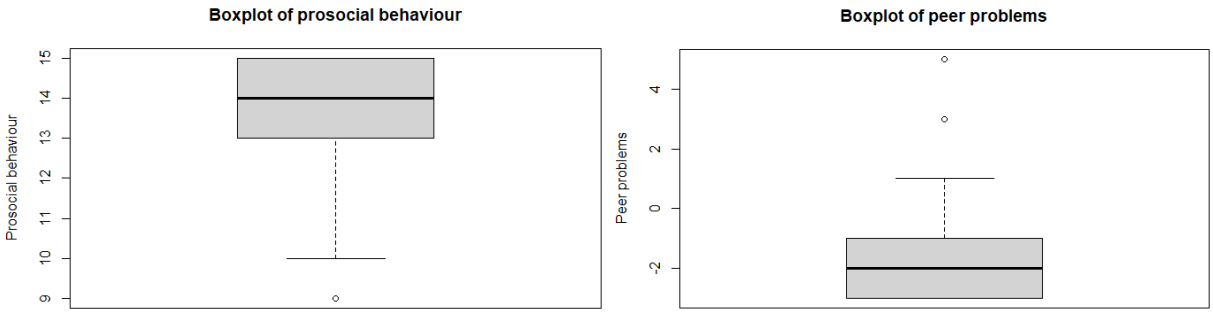

Figure 6: Boxplots of the prosocial behaviour and peer problems outcome variables

The result may also depend on the variable selection process. The missingness threshold of 25% is quite strict, and could have resulted in a drop of relevant variables. However, as was said before, the median and mode imputation techniques can heavily affect the variation and correlations of affected variables (Zhang, 2016), so there would also be a need for a different imputation technique.

Likewise, throughout the feature selection process, possibly interesting variables may have been removed. Overall, the correlation coefficients, LASSO and recursive feature elimination processes will attempt to select the most important features, but some interesting ones may have been left out. Unfortunately, it is not possible to check every subset combination, so for computational efficiency reasons there needs to be some middle ground.

Also, there are a lot of model types for machine learning, so it is possible that another model specification with a different subset combination has a better predictive capacity. Due to the amount of time involved in motivating the model type, learning how it works in some capacity and the eventual training and validating, the decision was made to focus on these three model types.

Section 5: Conclusion & discussion

Section 5.1: Conclusion

In this thesis, the goal was to predict two different aspects of social competence, which are prosocial behaviour and peer problems. Linear regression, random forest and gradient boosting models were chosen for this purpose, and for both outcome variables, the random forest models outperformed the other two model types. The non-linearity and interaction capturing capabilities were hypothesized to be the reason for outperforming the linear model, and reasons for the weak performance of gradient boosting were thought to be due to either not tuning the hyperparameters enough, or due to outliers affecting random forests less. As for the variables, the correlation analysis already showed that most predictors are weakly associated with the outcome variables. Moving forward to the model analyses, it is also shown that even the best models have a weak predictive capacity. Compared to the baseline models that use the mean as single predictor, the best models only have a cross-validated RMSE value that is 21% and 25% lower for the prosocial behaviour and peer problems models respectively. If you zoom in on the best prosocial behaviour outcome model, being able to stop when being told “no” accounts for the majority of the predictive power of the best prosocial behaviour model, whereas that was the case for irritability for the best peer problems model. The presence of most predictors, as well as their direction in the partial dependence plots, has been traced back to literature, However, it is still hard to make any concrete statements about the predictors when the predictive capacity of the models is so weak. Therefore, the conclusion of this research is that it is hard to predict the two social competence aspects of six-year-old children using the YOUth cohort study, given the data that was used and the preprocessing steps that have been taken.

Section 5.2: Discussion

The conclusion discussed the key takeaways from this thesis, as well as how the findings relate to literature. This discussion section is meant to summarize the limitations of this thesis, to make suggestions for future research, and to briefly add a note about the experience of working in the SANE environment.

Most of the limitations in this thesis were already briefly discussed in Section 4.3, which mainly described how the lack of data in the study, as well as decision making, may lead to these results. The first issue could have been tackled perhaps by focusing on a younger group of children (e.g., 3 years old, instead of 6 years old), as a larger group of children could be analysed. As for the decision making aspect, it is necessary to make decisions throughout the duration of a thesis, as it is not possible to exhaust every option. However, having a more lenient threshold for deleting variables in combination with different imputation techniques, as well as having a more thorough hyperparameter range to check for the models, may result in better results in the end. Do note that these changes may make the process more intensive. This thesis focused on the prediction side, and future studies could go more in depth by focusing on the interactions between variables, or alternatively they can focus on generating causal hypotheses and testing them using empirical studies.

To end this thesis, a request was made to add a small paragraph about the experience of working on the data in the SANE environment. The secure environment is fast, and it is easy to navigate. A suggestion for the future would be to make installing packages more straightforward, or to pre-install a large range of possible packages. This would make it easier for the researcher to quickly get their hands on a package, without having to send mails each time.

Reference list

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

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

Amrei, M. T., Shafiri, N., Taheri, A. (2020). A model for predicting social competence from resilience by interpreting the mediating role of academic adjustment of medical students of Mazandaran. *Int J Med Invest, 9*(3), 17-36. http://intjmi.com/article-1-515-en.html

Ananth, C. V., & Schisterman, E. F. (2017). Confounding, causality, and confusion: the role of intermediate variables in interpreting observational studies in obstetrics. *American Journal of Obstetrics and Gynecology*, *217*(2), 167–175. https://doi.org/10.1016/j.ajog.2017.04.016

Assimon, M. M. (2021). Confounding in observational studies evaluating the safety and effectiveness of medical treatments. *Kidney360*, *2*(7), 1156–1159. https://doi.org/10.34067/kid.0007022020

Ayyadevara, V. K. (2018). Gradient boosting machine. In *Apress eBooks* (pp. 117–134). https://doi.org/10.1007/978-1-4842-3564-5_6

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Blandon, A. Y., Calkins, S. D., & Keane, S. P. (2010). Predicting emotional and social competence during early childhood from toddler risk and maternal behavior. *Development and Psychopathology*, *22*(1), 119–132. https://doi.org/10.1017/s0954579409990307

Bøe, T., Hysing, M., Skogen, J. C., & Breivik, K. (2016). The Strengths and Difficulties Questionnaire (SDQ): Factor structure and gender equivalence in Norwegian Adolescents. *PLOS ONE*, *11*(5), e0152202. https://doi.org/10.1371/journal.pone.0152202

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

Breiman, L., Cutler, A., Liaw, A., Wiener, M. (2022). *Package 'randomForest'*. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Chen, C., & Gu, G. X. (2019). Machine learning for composite materials. *MRS Communications*, *9*(2), 556–566. https://doi.org/10.1557/mrc.2019.32

Chen, Q., Meng, Z., Liu, X., Jin, Q., & Su, R. (2018). Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes*, *9*(6), 301. https://doi.org/10.3390/genes9060301

Columbia University Irving Medical Center. (n.d.). *Least Absolute Shrinkage and Selection Operator (LASSO).* https://www.publichealth.columbia.edu/research/population-health-methods/least-absolute-shrinkage-and-selection-operator-lasso

De Swart, F., Burk, W. J., Nelen, W. B. L., Van Efferen, E., Van der Stege, H., & Scholte, R. H. J. (2022). Social competence and relationships for students with emotional and behavioral disorders. *Journal of Special Education*, *56*(4), 225–236. https://doi.org/10.1177/00224669221105838

Diener, M. L., & Kim, D. (2004). Maternal and child predictors of preschool children's social competence. *Journal of Applied Developmental Psychology*, *25*(1), 3–24. https://doi.org/10.1016/j.appdev.2003.11.006

Eisenberg, N., Fabes, R. A., Karbon, M., Murphy, B. C., Wosinski, M., Polazzi, L., Carlo, G., & Juhnke, C. (1996). The relations of children's dispositional prosocial behavior to emotionality, regulation, and social functioning. *Child Development*, *67*(3), 974–992. https://doi.org/10.1111/j.1467-8624.1996.tb01777.x

Emerick, K. (2011). *Derivation of OLS Estimator*. https://are.berkeley.edu/courses/EEP118/current/derive_ols.pdf

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00516-9

Fabiano, G. A., Vujnovic, R. K., & Pariseau, M. E. (2010). Peer problems. In *Springer eBooks* (pp. 1563–1588). https://doi.org/10.1007/978-0-387-09757-2_57

Foley, M. (2022). *Ordinary Least Squares.* https://bookdown.org/mpfoley1973/supervised-ml/ordinary-least-squares.html

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

Fugard, A. (2022). *Chapter 5: Linear regression | Using R for social reserach.* https://inductivestep.github.io/R-notes/linear-regression.html

Golden, C. E., Rothrock, M. J., & Mishra, A. (2019). Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the

environment of pastured poultry farms. *Food Research International*, *122*, 47–55. https://doi.org/10.1016/j.foodres.2019.03.062

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, *83*(2), 83–90. https://doi.org/10.1016/j.chemolab.2006.01.007

Haji, S. H., & Abdulazeez, A. M. (2021). Comparison of optimization techniques based on Gradient Descent algorithm: a review. *PJAEE*, *18*(4). 2715–2743. https://archives.palarch.nl/index.php/jae/article/view/6705

Hatami, F., Rahman, M. M., Nikparvar, B., & Thill, J. (2023). Non-Linear Associations between the urban Built environment and Commuting modal split: a random forest approach and SHAP evaluation. *IEEE Access*, *11*, 12649–12662. https://doi.org/10.1109/access.2023.3241627

Hegelich, S. (2016). Decision Trees and Random Forests: Machine learning techniques to classify rare events. *European Policy Analysis*, *2*(1). https://doi.org/10.18278/epa.2.1.7

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review*, *105*(4), 765–789. https://doi.org/10.1017/s0003055411000414

Jovic, A., Brkić, K., & Bogunović, N. (2015). *A review of feature selection methods with applications*. https://doi.org/10.1109/mipro.2015.7160458

Junge, C., Valkenburg, P. M., Deković, M., & Branje, S. (2020). The building blocks of social competence: Contributions of the Consortium of Individual Development. *Developmental Cognitive Neuroscience*, *45*, 100861. https://doi.org/10.1016/j.dcn.2020.100861

Kauten, R. L., & Barry, C. T. (2020). Externalizing behavior. In *Springer eBooks* (pp. 1509–1512). https://doi.org/10.1007/978-3-319-24612-3_894

Kuhn, M. (2008). Building Predictive Models in *R* Using the caret Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Kuhn, M. & Johnson, K. (2019). *Recursive Feature Elimination.* https://bookdown.org/max/FES/recursive-feature-elimination.html

Livingston, F. (2005). Implementation of Breiman's Random Forest Machine Learning Algorithm. *ECE591Q Machine Learning Journal Paper*, 1–13.

https://www.researchgate.net/publication/242751368_Implementation_of_Breiman%27s_Random_Forest_Machine_Learning_Algorithm

Li, A. H., & Bradic, J. (2018). Boosting in the presence of outliers: adaptive classification with non-convex loss functions. *Journal of the American Statistical Association: Theory and Methods, 113*(522), 660-674.

Lu, M., & Ishwaran, H. (2020). Discussion on "Nonparametric variable importance assessment using machine learning techniques" by Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon. *Biometrics, 77*(1), 23–27. https://doi.org/10.1111/biom.13391

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/bm.2013.018

Molnar, C. (2023). *Partial Dependence Plot (PDP)*. https://christophm.github.io/interpretable-ml-book/pdp.html

Muthukrishnan, R., & Rohini, R. (2016). *LASSO: A feature selection technique in predictive modeling for machine learning*. https://doi.org/10.1109/icaca.2016.7887916

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

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E. W. A., Brouwer, R. M., Buimer, E. E. L., Hessels, R. S., De Heus, R., Huijding, J., Junge, C. M. M., Mandl, R. C. W., Pas, P., Vink, M., Van Der Wal, J. J. M., Pol, H. E. H., & Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience, 46*. https://doi.org/10.1016/j.dcn.2020.100868

Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. *IASRI, New Delhi, 1*(1), 58–65. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=wBWuZJgAAAAJ&citation_for_view=wBWuZJgAAAAJ:P5F9QuxV20EC

Ranstam, J., & Cook, J. (2018). LASSO regression. *British Journal of Surgery, 105*(10), 1348. https://doi.org/10.1002/bjs.10895

Ray, S. (2019). *A quick review of machine learning algorithms*. https://doi.org/10.1109/comitcon.2019.8862451

Saraç, S., Abanoz, T., & Gülay Ogelman, H. (2021). Self-regulation predicts young children's peer relations. *Turkish International Journal of Special Education and Guidance & Counselling, 10*(1), 56–65. https://www.tijseg.org/index.php/tijseg/article/view/5

Shernoff, D. J. (2010). Engagement in After-School programs as a predictor of social competence and academic performance. *American Journal of Community Psychology*, *45*(3–4), 325–337. https://doi.org/10.1007/s10464-010-9314-0

Stone, L. L., Janssens, J. M. A. M., Vermulst, A. A., Van der Maten, M., Engels, R. C. M. E., & Otten, R. (2015). The Strengths and Difficulties Questionnaire: psychometric properties of the parent and teacher version in children aged 4–7. *BMC Psychology*, *3*(1). https://doi.org/10.1186/s40359-015-0061-8

Taleipour, N., & Motlaq, M. (2021). The Role of Family Emotional Climate in Predicting Social Competence among Female High School Students in Dezful City. *Sociological Studies of Youth*, *12*(42), 123–134. https://ssyj.babol.iau.ir/article_684307_aa38c4747982c13dc9bdb95cf6394611.pdf

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management*, *24*(1), 53–62. https://doi.org/10.1016/j.jom.2005.03.001

Vidal-Ribas, P., Brotman, M. A., Valdivieso, I., Leibenluft, E., & Stringaris, A. (2016). The Status of Irritability in Psychiatry: A Conceptual and Quantitative Review. *J Am Acad Child Adolesc Psychiatry*, *55*(7), 556–570. https://doi.org/10.1016/j.jaac.2016.04.014

Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/s12859-016-0995-8

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of Cross-Validation, Bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*(3), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Youth in Mind. (n.d.). *Scoring the SDQ*. https://www.sdqinfo.org/py/sdqinfo/c0.py

Yue, X., & Zhang, Q. (2023). The association between peer rejection and aggression types: A meta-analysis. *Child Abuse & Neglect*, *135*, 105974. https://doi.org/10.1016/j.chiabu.2022.105974

Zhang, L., Yao, B., Zhang, X., & Xu, H. (2020). Effects of Irritability of the Youth on Subjective Well-Being: Mediating Effect of coping styles. *Iranian Journal of Public Health*. https://doi.org/10.18502/ijph.v49i10.4685

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *PubMed*, *4*(1), 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38

Appendix A: Datasets chosen for analysis

This appendix sub-section provides a brief description of the datasets used. The LANGUAGE dataset also has a supplement with the same description. Hence, the 17 datasets consist of 16 different questionnaires, and one supplement.

Dataset code	**Description**
CBQ	"My child's personality"
DAYCARE	"The care of my child"
FCOMP	"The parents' competence feeling with respect to raising a child"
FEELINGS	"My child's thoughts and feelings"
GENDER	"My child's gender identity"
GK	"Parenthood and upbringing"
HEALTH	"My child's health"
K_GRAM	"Data from the growth booklet"
L_BULLY	"Bullying behaviour around my child"
LANGUAGE	"Language use in my child's social situations"
MEDIA	"Apps, television, games and books"
NUTRITION	"My child's nutrition"
Q6	"My child's problematic behaviour and skills"
SLEEP	"My child's sleeping habits"
SPORT	"My child's sports and hobbies"
V_BIG5	"My child's personal characteristics"

Appendix B: Minor considerations during pre-processing phase

While the most important pre-processing considerations were discussed in the main text, there are still some minor steps that have been taken to ensure having data that is usable. The actions below have been performed before any of the major considerations from the main text:

- Make sure that the datasets are in the same format, which consists of checking whether the set has only one time variable (e.g., for six-year-olds) and that none of the sets are in the long format.
- Remove dataset-specific metadata, which are simply variables describing some common study characteristics such as the location, date and study name. These are deemed not interesting for prediction purposes.
- Take a look at possibly problematic variables (e.g., a lot of missingness, or very low variation) and see if it is possible to create a new variable out of it.

The following few steps have been done after removing the variables due to missingness, zero variance and near-zero variance:

- Textual variables with a large variety of answers (e.g., what shows the child watches) will not be used due to the time needed to process such a column. It is not a significant problem, as it only happens a few times throughout the datasets.
- Check whether the variables are in the correct format (e.g., numeric or factor).
- If ordinal variables have the option “not applicable” or any similar option, and that option is not very common in the questionnaire data, then the values corresponding to that option will be turned into NA’s. This way the ordinality of the variable can be retained and a transformation into factor is not necessary.