Predicting youth help usage based on pre-usage risk factors

A predictive risk modeling approach

Master Thesis in Business Information Management

Rotterdam School of Management
Erasmus University Rotterdam

Cas de Weerd

400575

Supervisors Erasmus University:
Coach: dr. Jan van Dalen
Co-reader: dr. Otto Koppius

Supervisors Company: Frouwkje de Waart
Coach: Cathelijne Mieloo

Date: 02-08-2018

Preface

Acknowledgements

This study resembles a milestone for me as it is the final product of my studies. It has been a challenging and time-consuming process, but I have sincerely enjoyed the full effort. The thesis trajectory has largely contributed to my personal and professional development. For this, I want to thank a few people.

I would like to thank Cathelijne Mieloo for her outstanding effort and support as my coach during the thesis trajectory. Cathelijne always provided me with critical feedback, which largely contributed to the quality of this study. Besides, I would like to thank Frouwkje de Waart for granting me the opportunity to write my thesis at the municipality of Rotterdam. I truly enjoyed my time at the OBI department.

Furthermore, I am very grateful to dr. Jan van Dalen for his kindness and willingness to help. Besides, Jan has been a great motivator and substantially contributed to final results of this thesis. Also, I would like to thank Otto Koppius for his early interest and voluntary involvement in this study. The articles that Otto provided have helped me greatly during the entire process.

Last but not least, I would like to thank my family and my roommate Frank for supporting me during the process. In particular, I would like to thank my mother for the final spelling check and Frank for distressing me during the weekend with some outstanding parties and unforgettable nights out. It has been a blast.

Cas de Weerd

Abstract

Municipalities in the Netherlands are responsible for the provision of youth help services. Youth help services are support services to youngsters and families with child or parenting problems. The budget for youth help services has been declining, while the number of youth help trajectories is growing. This stresses the importance for municipalities to accurately predict future youth help demand. An accurate prediction of children and families who are expected to be entering a youth help trajectory in the future could improve prevention policies and the allocation of resources. In the past, research to youth help usage was aimed at identifying risk factors. However, these models could also be used for prediction. Moreover, these studies were mainly based on survey data, while linked-administrative data seems more suitable for the prediction of youth help usage. In this study a predictive risk modeling approach is applied to linked-administrative data, in order to predict first-time youth help usage in the following year. Six classification techniques are applied. The best model of these six is compared with a prior model of the city of Rotterdam. The dataset includes fifteen risk factors, derived from thirteen databases from the Central Bureau of Statistics, containing information of youngsters until the age of 23, who lived in Rotterdam between 2015 and 2017 (n=166.654). Results showed that the best performing model was a random forest model, trained on balanced training data. When compared to a previous model, the increase in predictive performance is: 51.2% in sensitivity, 36.7% in precision and 0.427 in F1-score at a risk threshold of 0,5. Besides, in terms of AUC, model performance increases from 0.62 to 0.80. In this model a combination of life events, child, parental and household factors are the most important predictors. Moreover, when previous youth help usage was added to the model the performance increases to a sensitivity of 76.2%, a precision of 58.7% and a F1-score of 0.66, at a risk threshold of 0,5. Moreover, the AUC increases to 0.91. However, this model does not predict the first-time usage, but total demand for youth help services in the following year. This study showed that predictive risk modeling applied to administrative data improves the prediction of first-time youth help usage in the following year. In theory this model could be used for individual prevention measures However, as the model performance is improved, the precision is still low. In future research this model could be improved by extending the number of risk factors and improving data quality.

Contents

List of Tables

List of Figures

1 Introduction

Since 2015, municipalities in the Netherlands are responsible for the funding and management of juvenile services in their region. In 2016, over 400 thousand youngsters between the age of 0 and 22 received support from the juvenile services in their municipality (CBS, 2017b;a). The juvenile services in the Netherlands provide various support services to youngsters and families for a variety of child and parenting problems. These services are divided into three categories: youth help, youth protection and youth probation. Youth help services are the main category, as they account for 91.3% of all the trajectories. Every day, youngsters and families are referred to the youth help institutions to receive appropriate support. Since the transition in 2015, the budget for youth help services has been declining yearly and this trend is expected to continue in the coming years (Rijksoverheid, 2017). Nevertheless, the number of youth help trajectories have been increasing since 2015 (CBS, 2016; 2017b). This stresses the importance of an accurate prediction of the future demand for youth help services. An accurate prediction of future youth help demand could contribute to effective prevention policies and an optimal allocation of the limited resources.

Research aimed at the prediction of youth help usage has, however, not been done before. But, the risk factors for youth help usage, which could possibly be used for prediction, have been extensively studied in the past (Bot et al., 2013; Mieloo et al., 2013; Sadiraj et al., 2016; Hermanns et al., 2005). In these studies several risk factors for child and parenting problems and the use of juvenile services have been identified. The main four categories of risk factors are: child factors, parental and household factors, environmental factors and stressful life events (Bot et al., 2013; Stevens et al., 2009; Hermanns et al., 2005). The presence of these risk factors could result in the need of help from youth help services.

However, the current literature regarding youth help usage, is mainly based on survey data (Bot et al., 2013; Stevens et al., 2009; Hermanns et al., 2005), while recent years have witnessed a strong increase in the number of Dutch government agencies sharing pseudonised person level administrative data. The usage of linked-administrative data seems promising, especially for predictive analytical approaches (Connelly et al., 2016). In the U.S., several studies have proven that linked-administrative data, combined with thorough knowledge about risk factors, creates opportunities to accurately predict certain outcomes closely related to youth help usage, for example: child maltreatment (Schwartz et al., 2017; Vaithianathan et al., 2012; Marshall and English, 2000). These studies commonly apply a predictive risk modeling approach.

Predictive risk modeling is a type of predictive analytics and is aimed at risk stratifying individuals in a population, based on the likelihood of these individuals to experience a certain outcome or event (Cuccaro-Alamin et al.,

2017). Predictive analytics is the application of data mining and modeling to existing data to discover patterns and make predictions (Guazzelli, 2012).

Predictive risk modeling has proven to be a successful tool for many purposes in the social services area and the healthcare industry. Some studies predict mental disorders like depression or PTSD, using social media data (Reece et al., 2017; Preoţiuc-Pietro et al., 2015; Benton et al., 2017), while others successfully predict the risk of heart failure, lung cancer, breast cancer, post-surgery mortality, emergency department use and homelessness (Brennan et al., 2009; Euhus, 2004; Fogel et al., 1998; Kanazawa et al., 1996; Lewis et al., 2013). The U.S. Department of Health and Human Services implemented a predictive risk modeling tool to identify and risk-stratify patient populations, particularly patients with multiple chronic conditions, for more focused care intervention (Weir and Jones, 2009).

The present study aims to improve the prediction of youth help usage by applying a predictive risk modeling approach to linked-administrative data.

2 Juvenile services

Juvenile services are support services for youngsters until the age of 18, but the trajectories can be extended until the age of 22. The juvenile services have three main responsibilities: The first responsibility is the help and care for youngsters and their parents when struggling with child or parenting problems. Child problems can be mental or behavioural problems. The second responsibility is promoting the participation in society and the independent functioning of youngsters with problems. The third responsibility is the support or take-over of activities related to personal care for youngters with mental or physical limitations (CBS, 2016). These three responsibilities result in the following services: youth help, youth protection and youth probation.

2.1 Youth help

The youth help is the largest part within the juvenile services area. In 2016, 388.655 youngsters received support from one of the youth help institutions, which accounts for 91.3% of the total juvenile services area(CBS, 2017b). Youth help is divided into two categories: Youth help without accommodation and youth help with accommodation. Figure 1 provides an overview of the number of youngsters in 2016 with youth help support.

2.1.1 Youth help without accommodation

Youth help without accommodation consists of four categories: performed by the district or neighborhood team, ambulant help at the location of the care provider, day treatment at the location of the care provider and help provided in the youngster's network. These categories have in common that the youngster stays and sleeps at home.

Performed by the district or neighborhood team
Almost every municipality in the Netherlands works with neighborhood teams. These teams are mainly responsible for the coordination of help in their neighborhood. These teams usually focus on help as well as prevention. This is the only service which can be considered as non-specialised support. This service is performed by non-specialised workers from the municipality. All the other juvenile services are performed by external specialists.

Ambulant help at the location of the care provider
Ambulant help or group conversations at the location of the care provider.

Day treatment at the location of the care provider
Group or individual help by a multidisciplinary team at the location of the care provider. It concerns, for example, a treatment combining physiotherapy, behavioural therapy and psychotherapy.

Help provided in the youngster's network
The help is not provided at the care providers location, but for example at home, at school or somewhere else in the youngster's network.

2.1.2 Youth help with accommodation

Youth help with accommodation also consists of four categories: foster care, family focused, forced admission and other forms of accommodation at the care provider. These categories have in common that the youngster is not staying, and especially sleeping, at his own home anymore. This could be permanent or just for certain periods, for example, weekends or just for a day.

Foster care
The youngster is (temporarily) admitted to a foster family. The parents, the youngster and the foster parents are being supervised by and receive support from the foster care provider.

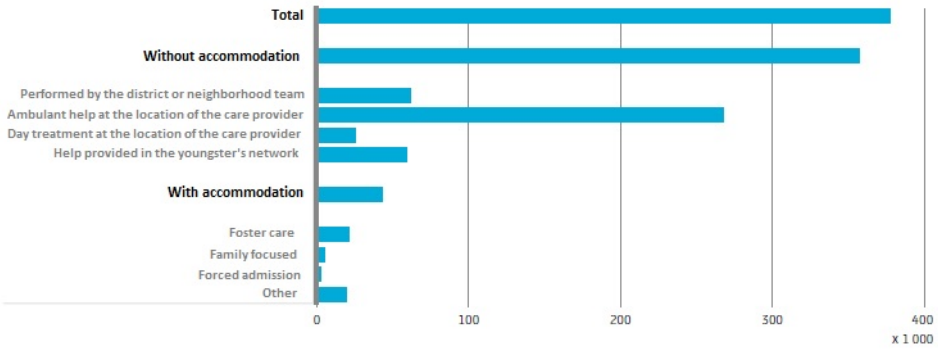
Family focused
All forms of accommodation resembling a family situation, apart from foster

care. This could be for example a care farm or family house.

Forced admission
The youngster will unvoluntarily be placed in a closed section of a care provider, because of, for example, severe behavioural problems. The focus of this treatment is not punishment, but help in a protected environment.

Other forms of accomodation at the care provider
This category consists of all other forms youth help with accommodation that have not been included in the three categories above. This category includes, for example, admission in an assisted living facility.

Figure 1: Overview of youngsters with youth help support 2016



2.2 Youth protection

In 2016, 30.250 youngster received support from the youth protection department, which accounts for 7.1% of the total juvenile services area (CBS, 2017a). Youth protection is a measure forced by a judge. A youngster is then 'placed under supervision' or 'placed under guardianship'. This happens when the healthy and safe development of a youngster is threatened and voluntary help is not effective or sufficient.

2.3 Youth probation

Youth probation is the smallest category of the juvenile services department. In 2016, 6.900 youngsters received support from the youth probation department, which accounts for 1.6% of the total juvenile services area (CBS, 2017a). Youth probation is a combination of guidance and control for youngsters who committed criminal offenses, including school absenteeism. The youngster

receives tailor-made guidance from a youth probation worker to prevent him or her from mistaking again. Youth probation can be imposed by the juvenile court judge or the public prosecutor. Youth probation can also be voluntarily initiated or on the initiative of the Dutch Child Protection council.

2.4 Entering a youth help trajectory

The path from having problems to receiving support from youth help institutions is mostly voluntary and always goes through a third party. If a child or a family has problems, they can get support from a youth help institution through the municipality, their general practitioner, the neighborhood team or the CJG (Center for Youth and Family). These professionals will assess the severeness of the situation and decide whether to forward the case to the appropiate youth help institution. The youth help institution will then take the appropiate measures. In some cases, a youth help trajectory is forced by a judge.

3 Objective and Research problem

In the past, much descriptive research has been done to the risk factors that could contribute to the prediction of youth help usage. These studies were mostly based on survey data (Bot et al., 2013; Mieloo et al., 2013; Sadiraj et al., 2016; Hermanns et al., 2005). In these studies several risk factors have been identified. The main four categories of risk factors are: child factors, like intelligence and birth weight; parental and household factors, like parenting style and parental drug use; environmental factors, like metropolitanism and unfavorable neighborhood; and life events, like parental divorce and being victim of a crime or accident (Bot et al., 2013; Stevens et al., 2009; Hermanns et al., 2005). The presence of these risk factors could result in the need of support from youth help institutions. The help seeking behaviour of youngsters and their families, when having child or parenting problems, is dependent on many factors. Many studies show that mostly the accumulation of risk factors is important for explaining the need of help (Stevens et al., 2009; Aalbers-van Leeuwen et al., 2002; Hermanns et al., 2005). These theories are based on the balance model of Bakker (1999), which explains the difference between the amount of support needed and supplied, based on a person's mental resilience and mental strain (Stevens et al., 2009; Bakker, 1999). Earlier research, however, finds little evidence of this accumulation effect (Mieloo et al., 2013; Bot et al., 2013; Sadiraj et al., 2016). For all the identified risk factors, the presence of a single risk factor has a significant, but small effect on youth help usage and the combinations of risk factors that have been studied, only slightly improve the results. Thus, despite the wealth of the current literature regarding the risk

factors for youth help usage, a clear understanding and an accurate prediction of children and families in need of youth help support has proven to be difficult.

However, the current literature generally has a descriptive nature and applies traditional methods, like logistic regression, to survey data, while predictive analytics, applied to linked-administrative data, seems to have the potential to adress some of the problems in the prediction of youth help usage. The advantages of linked-administrative data for the prediction of youth help usage mostly relate to size (full population), objectivity, service delivery and the availability on individual level (Connelly et al., 2016). Besides, in the past, single relationships between the risk factors and youth help usage are examined, or pre-specified combinations of risk factors, whereas predictive risk modeling applies a more holistic approach. As many data points as possible are examined, even if no relationship has previously been specified (Cuccaro-Alamin et al., 2017).

The prediction of youth help usage has, at the moment, not sophisticately been done before. The models that have been developed in prior studies were not developed with the intent of prediction, but some of them could be applied as such, in order to assess predictive gain. For this reason, the present study aims to improve the prediction of youth help usage by applying a predictive risk modeling approach to linked-administrative data. Multiple classification models will be developed, which will predict the individual risk of the population of Rotterdam to enter a youth help trajectory in the following year. A risk threshold is chosen to classify observations as future youth help users. The performance of the models will be evaluated and compared, in order to choose the most suitable model. The models risk stratify the population and could be used to identify the youngsters and families who are most likely to be in need of support from youth help institutions in the next year. These high risk individuals and their families could possibly be approached for personal prevention measures. Besides, the distribution of resources among the neighborhood teams could be improved, by locating the neighborhoods in which these high risk individuals will most likely receive support.

Thus, the research question of this study is as follows:

How much can a predictive risk modeling approach, applied to linked-administrative data, improve the prediction of youth help usage?

3.1 Practical relevance

The practical relevance of this study mainly relates to prevention policies. At the moment, the development of children is monitored by institutions, like: CJG or schools. These institutions provide information to all their clients, relating to the healthy development of children. The outcomes of this study create the possibility to identify specific children and families with a high risk of youth help usage in the coming year. These families could now, if desired, individually be approached with certain prevention measures. Also, the neighborhoods in which these families live could be identified, in order to estimate by which neighborhood team these individuals are most likely to be receiving support. This could contribute to a more efficient distribution of resources among the neighborhood teams.

3.2 Academic relevance

Much research has been done aimed at identifying the factors that can be used in order to predict youth help usage (Bot et al., 2013; Mieloo et al., 2013; Sadiraj et al., 2016; Hermanns et al., 2005). However, all the identified risk factors seem to have a small effect on youth help usage and the combinations of risk factors that have been studied, only slightly improve the results. A clear understanding and an accurate prediction of youth help usage has thus proven to be difficult. In the aforementioned studies, traditional methods, like logistic regression, are applied to, mostly, survey data. This study will contribute to the current field of knowledge, by applying a predictive risk modeling approach to linked-administrative data, in order to adress some of the gaps in the understanding and prediction of youth help usage.

4 Theoretical background

The theoretical background describes the different concepts that are necessary to develop the intended predictive model and to be able to answer the posed research question. The first part willl provide an elaborate review of the current knowlegde about the most important risk factors for the use of youth help services, the second part will describe the concept of predictive risk modeling and the third part will substantiate the choice for the use of administrative data.

4.1 Risk factors for youth help usage

As the goal of this research is to build a predictive model, which is able to predict first youth help usage as accurately as possible, this section will review the main risk factors for youth help usage. Considering the broadness of the topic, many different risk factors have been identified. The most important risk factors can be divided into four categories: child factors, parental and household factors, environmental factors and life events (Bot et al., 2013; Stevens et al., 2009; Aalbers-van Leeuwen et al., 2002; Bakker et al., 2000; Bakker, 1999; Hermanns et al., 2005; Kijlstra et al., 2001; Woldringh and Peters, 1995). The articles of Bot et al. (2013); Bayer and Sanson (2003); Hermanns et al. (2005); Stevens et al. (2009) give an elaborate review about the main risk factors influencing child and parenting problems and the use of youth help services.

4.1.1 Parental and household factors

Parental and household factors are environmental characteristics within the family context, which are influencing child and parenting problems.

Parenting style
An important factor influencing the behaviour of a child is the the parenting style of the parents (de Roos et al., 2011). Parents play a major role in the development of a child and certain parenting styles lead to more child problems than others. An authoritative parenting style, which is characterised by a balance between authority as well as affection, leads to less behavioural problems (Baumrind, 1971). A parenting style which leads to more behavioural problems is the authoritarian parenting style (Lamborn et al., 1991). The authoritarian parenting style is characterised by a strong degree of authority and less affection. The main difference between the two parenting styles is thus the amount of affection given to the child. The importance of affection in parenting style is also found in the studies of Anderson et al. (1986), Bell and Chapman (1986) and Bayer and Sanson (2003). These studies found that non-conformative behaviour of a child often leads to a more authoritarian parenting style, showing less affection to the child, leading to even more behavioural problems. A good balance between a certain degree of authority and affection in parenting style is thus important for less problematic behaviour.

Parental (long-term) disease or condition
Living together in a family where one or both of the parents are suffering from a disease or condition leads to a lower well-being and life satisfaction (Sieh et al., 2013). The daily confrontation with the limitations of the parent and the insecurity about the further progress of the disease could lead to feelings of anger, fear and sadness (Sieh et al., 2013). These negative feelings have an

impact on the psychological well-being of a child. This is why children, living together with one or both parents suffering from a long-term disease or condition, are more likely to have severe mental and behavioural problems (Bot et al., 2013). Also, children in families with one or more parents suffering from mental health problems are more likely to have severe mental and behavioural problems (Bot et al., 2013; Bayer and Sanson, 2003; Hermanns et al., 2005). Parental diseases or conditions are thus important factors relating to mental and behavioural problems of their children.

Parent-child relationship
"Risky families are characterized by conflict and aggression and by relationships that are cold, unsupportive, and neglectful. These family characteristics create vulnerabilities and/or interact with genetically based vulnerabilities in offspring that produce disruptions in psychosocial functioning (specifically emotion processing and social competence), disruptions in stress-responsive biological regulatory systems and poor health behaviors, especially substance abuse." (Repetti et al., 2002, p.1) A bad parent-child relationship will thus more likely result in behavioural and mental problems of the child (Bot et al., 2013; Bayer and Sanson, 2003).

Marital discord
Marital discord is in many cases prior to a parental divorce. Both marital discord and parental divorce negatively influence the emotional and mental well-being of a child (Hermanns et al., 2005; Bayer and Sanson, 2003). Parental divorce will be addressed seperately in the risk factor category of life events.

Addiction within family
Addiction within the family can harm the well-being and development of a child in several ways. Drug use during pregnancy enhances the risk of a child being negatively affected at birth in a physical or mental way. Besides, an infant delivered to a drug-addicted woman is at risk for problems of growth and development (Chasnoff, 1988). Furthermore, parental drug use could a lead to a genetic susceptibility for addiction of the child. A child with genetic susceptibility for addiction has an increased chance for addiction as well, especially when growing up in a environment where drug use is present (Stanley and Vanitha, 2008; van der Zwaluw, 2011). Addiction problems within families highly increase the chance of child and parenting problems (Bot et al., 2013; Hermanns et al., 2005).

Parental education
A low level of parental education can lead to a higher risk of parenting problems (Bucx and de Roos, 2011). Furthermore, a high level of parental education leads to high expectations of the child's academic achievements (Davis-Kean, 2005). These parental expectations could lead to a higher amount of stress and anxiety, due to the experienced pressure to meet these expectations. Moreover, a low level of parental education leads to more physical

aggression (Nagin and Tremblay, 2001; Hermanns et al., 2005). However, the relationship between a low parental education and a higher amount of child and parenting problems has not been found significantly (Bot et al., 2013).

Ethnic background
Children in families with parents with a non-western background could have mental or behavioural problems, caused by cultural differences and the stress that comes along (Stevens et al., 2003; van den Broek et al., 2010). A negative relationship between non-western background and the use of youth help services has, however, been found (Jansen et al., 2015; Bot et al., 2013; Stevens et al., 2009). Youngsters with a non-western background are less likely to make use of youth help services. A possible explanation is that non-western people experience more barriers to search for help when this is needed (Jansen et al., 2015). This leads to a gap in the amount of help needed and the amount of help provided.

Teenage motherhood
Teenage motherhood is significantly related to all kind of mental health problems and behavioural problems (Harden et al., 2007; Bot et al., 2013; Levine et al., 2007; Hermanns et al., 2005). Teenage mothers generally have limited life experience and a lower income, due to lower wages. These factors could lead to extra need of help from, for example, grand parents or youth help services.

Single parent families
A single parent family is an important factor related to the use of youth help services. Children from single parent families are significantly having more mental and behavioural problems (de Roos et al., 2011; Bot et al., 2013; Hermanns et al., 2005). Besides, a single parent stands alone to fulfil all tasks and responsibilities related to parenthood and is for this reason more likely to be in need of help to fulfil these tasks and responsibilities. This is why single parent families make use of youth help services much more often (Stevens et al., 2009; Bot et al., 2013).

Poverty and income
A person is poor if he does not have enough income over a longer period of time to have access to what is deemed to be absolutely necessary in society (Soede et al., 2011). In the Netherlands, the income boundary to be considered as poor is a total gross household income of € 33.696 per year and € 26.179 for a single parent family. Bot et al. (2013) haven't found a significant relationship between poverty and child or parenting problems. However, Mieloo et al. (2013) did find a significant relationship between income and the use of youth help services. People with a below-average income have a higher chance of youth help usage. Nevertheless, income was measured based on house value in this study, due to data availability. This could lead to biased results based on the unit of measurement.

4.1.2 Environmental factors

Environmental factors are the broader environmental factors that are influencing the use of youth help services. These are the factors beyond the family context.

Unfavorable neighborhood
Unfavorable neighborhoods are neighorhoods characterised by a low level of income, a low educational level and with a high degree of unemployment (Roes, 2005). Besides, unfavorable neighborhoods typically have poor housing conditions (Bot et al., 2013). A neighborhood with a low level of income slightly increases the chance of child and parenting problems (Bot et al., 2013; Hermanns et al., 2005).

Metropolitanism
Weijters et al. (2007) suggest that living in a big city (>250.000 inhabitants) could have a negative impact on the well-being of the youth. The quality of life in big cities in the Netherlands is on average lower than in smaller cities (Boksic and Verweij, 2011; Bot et al., 2013). This could influence the number of youth related problems. Earlier research has, however, not found a significant relationship for the relationship between metropolitanism and child or parenting problems or juvenile services usage (Bot et al., 2013; Mieloo et al., 2013). In 2015, most of the youth help services to people below eighteen were provided in cities between 50 thousand and 150 thousand citizens. Most of the youth help services to people between 18 and 22 were, on the other hand, provided in cities with more than 250 thousand citizens (CBS, 2016). More recent statistics show that the percentage of people using youth help services is at it's highest in cities with more than 250 thousand citizens, for all age categories (CBS, 2017b). In particular, these statistics show that youth help usage of people between the age of 18 and 22, living in a city with more than 250.000 inhabitants, is two times the Dutch average youth help usage in both years (CBS, 2016; 2017b). This suggests that age could be an important factor influencing the relationship between metropolitanism and the use of youth help services.

4.1.3 Child factors

Child factors have been found to be the most important category of factors influencing the well-being of youth and youth help usage (Bot et al., 2013). Child factors are contributing much more to the explanation of youth related problems in than the parental and environmental factors.

Temperament
Children with inhibited temperamental styles easily become anxious, angry or

upset. This could lead to an angry, over-controlling or punitive child rearing (Bayer and Sanson, 2003). These rearing characteristics are closely related to the authoritarian parenting style and negatively influence the parent-child relationship (Lamborn et al., 1991; Anderson et al., 1986). For this reason, temperament is an important risk factor for child and parenting problems (Bayer and Sanson, 2003; Bot et al., 2013; Hermanns et al., 2005).

Cry babies
A cry baby can place a lot of stress and emotional drain on his parents and this could result in angry and agressive parental behaviour. For this reason, a cry baby is an important risk factor for child-abuse (Hermanns et al., 2005) and an important risk factor for parenting problems (Bot et al., 2013). Thus, a cry baby is likely be an important risk factor for youth help usage.

Gender
The number of boys making use of youth help services is much larger than the number of girls. In 2016, 159 thousand girls and 219 thousand boys made use of youth help services (CBS, 2017b). This substantial difference is explained by the fact that the male gender is more likely to express externalising behaviour than the female gender (Bot et al., 2013). Problems with boys are thus more visible than with girls and faster result in the need of support . The male gender is therefore is an important risk factor for youth help usage (Bot et al., 2013; Hermanns et al., 2005; Mieloo et al., 2013).

Age
Age has a significant influence on the amount of behavioural and mental problems and seems to have a significant influence on youth help usage (Bot et al., 2013). An important aspect is puberty. Children, especially girls, have an increased risk of despression during puberty (Bayer and Sanson, 2003; Bot et al., 2013). Nevertheless, the percentage of youngsters receiving youth help support between the age of 4-11 was 13% and the youth help usage of youngsters between the age of 12-17 was 12%. Youth help usage of children with age of 0-3 and 18-22 were substantially lower, respectively 3% and 1% (CBS, 2017b). The low number of youth help users between the age of 18-22 is mainly due to the fact that the trajectories are only extended above the age of eighteen in special situations.

Birth weight
Children with a low birth weight (below 2500g) have an increased chance of autism, intellectual disability and ADHD (Bot et al., 2013) A low birth weight is thus an important risk factor for mental and behavioural problems and is expected to be positively related to youth help usage (Bot et al., 2013; Hermanns et al., 2005).

Long-term disease, condition or disability
The presence of a long-term disease, condition or disabilty is a very important

risk factor for child and parenting problems (Bayer and Sanson, 2003; Bot et al., 2013). For this reason, and due to the fact that youth help institutions are responsible for the support to children with an intellectual disability, is the presence of a long-term disease, condition or disability expected to be an important predictor of youth help usage.

Intelligence
A low or an extremely high intelligence seem to be important risk factors for mental and behavioural problems and youth help usage (Bot et al., 2013; Hermanns et al., 2005). Specifically, children who are either highly intelligent (IQ>115) or weakly intelligent (IQ<85) seem to have an increased chance to have mental or behavioural problems. Children with a very high intelligence often feel different or isolated, due to their deviating interests and capabilities. Children with low intelligence often have mental and behavioural problems or an intellectual disability (Bot et al., 2013).

Educational level
The educational level of a child seems to be an important predictor of child and parenting problems. The relationship can be explained in various ways. First, children with a low educational level generally have a lower intelligence. Secondly, mental and behavioural problems could complicate a child's progress in higher educational levels, causing a child to end up in a lower level of education. Thirdly, children, especially during puberty, have the tendency to copy each others behavioural problems. The higher number of behavioural problems on lower educational levels and the copying behaviour of children reinforces itself (Bot et al., 2013).

4.1.4 Life events

The experience of a stressful life events during childhood or adolescence has been proven to be an important risk factor for mental or behavioural problems (Bouma et al., 2008; Bot et al., 2013; Hermanns et al., 2005; Bayer and Sanson, 2003). However, the current literature mainly focuses on the impact of a stressful life event as a single predictor. Research to specific life events has not been done elaboratively. Nevertheless, Bot et al. (2013) studied the impact of three specific life events on child and parenting problems: Parental divorce, the death of a loved one and being victim of a crime or accident.

Parental divorce
In the Netherlands, every year approximately 0.9% of the parents divorce. This number relates to around 33 thousand children with separating parents. Among the three studied life events, parental divorce seems to be the most important risk factor for child and parenting problems. 27.6% of the children with seperated parents had any form of child or parenting problems.

The death of a loved one
The death of a loved one also has a significant relationship with child and parenting problems, but it was the weakest predictor among the three studied life events. 18.1% of the children who experienced the death of a loved one had any form of child or parenting problems. However, the data was collected in a survey, so the results could be biased by the multi-interpretability of the definition of "a loved one".

Victim of a crime or accident
Having been a victim of a crime or incident has a strong significant relationship with the amount of child and parenting problems. 22.4% of the children who had been victim of a crime or accidents had child or parenting problems.

Table 1 provides an overview of important risk factors for youth help usage.

Table 1: Overview of risk factors for youth help usage

Parental & Household factors	*Child factors*
Parenting style	Temperament
Parental (long-term) disease or condition	Cry babies
Parent-child relationship	Gender
Marital discord	Age
Addiction within family	Low birth weight
Parental education	Long-term disease, condition or disability
Ethnic background	Intelligence
Teenage motherhood	Educational level
Single parent families	
Poverty and income	
Life events	*Environmental factors*
Parental divorce	Unfavorable neighborhood
Death of a loved one	Metropolitanism
Victim of a crime or accident	

Many risk factors, like: single parent families, marital discord and parental divorce, or: intelligence, educational level and income, are strongly related to each other. Besides, many risk factors can be derived from admnistrative data. Section 5 will extensively describe the data.

4.2 Predictive risk modeling

Predictive analytics is the application of data mining, modeling, and analytic techniques to existing data to discover patterns and make predictions (Guazzelli, 2012). Predictive risk modeling (PRM) is a type of predictive analytics and is a statistical method of identifying characteristics that

risk-stratify individuals in a population, based on the likelihood that each individual will experience a specific outcome or event. The outcome of the model's mathematical algorithm is a risk score (Cuccaro-Alamin et al., 2017). Traditionally, risk assessments are made based on previously researched relationships, while in predictive risk modeling as many data points as possible are examined, even if no specified relationship has previously been identified.

Predictive risk modeling has various advantages in comparison to traditional risk assessment methods. First, due to the inclusion of many data points, PRM models can identify earlier unobserved relationships between variables (Marshall and English, 2000; Cuccaro-Alamin et al., 2017). Second, PRM-models are flexible learning models and have the ability to adjust to risk changes over time (Cuccaro-Alamin et al., 2017). Third, predictive risk modeling has proven to be inherently more consistent than other risk assessment methods (Cuccaro-Alamin et al., 2017). Fourth, PRM-models operate indenpently, whereas traditional risk assessment methods are, for example, dependent on worker training and compliance(Cuccaro-Alamin et al., 2017).

Predictive risk modeling is currently used in many industries, for many different purposes. In the insurance industry, the creditcard industry or by the tax authorities, it is used for fraud detection (Nyce, 2007; Ngai et al., 2011). In the e-commerce industry, it is used to predict buying patterns and helps to identify specific customer segments for personalized advertisement purposes (Linoff and Berry, 2011). In the healthcare industry, it has successfully been used to predict the risk of heart failure, lung cancer, breast cancer, post-surgery mortality, and emergency department use, among other things (Cuccaro-Alamin et al., 2017).

An important limitation of predictive risk modeling can be illustrated by an example in the United States. In the United States a predictive risk modeling tool is used as a screening tool to identify high risk families by the Child Protection Services. The by the algorithm identified high risk families are being subjected to further investigation. The purpose of the tool is to spend the limited amount of resources to the right children and families and has been proven to be very useful and accurate. Nevertheless, predictive risk modeling has several important limitations.

Child Protective Services in the United States suggests that the underlying data in their risk assessment tool reflects ingrained biases against African-Americans and others, due to the fact that African-Americans are more often reported with the suspicion of child abuse. Black children are thus over-surveilled in the system in comparison to white children, because the underlying data for the tool is not based on who abuses, but on who gets reported (Hurley, 2018) The outcomes of the predictive risk modeling algorithm are thus biased due to biases in the underlying data. In general, biases in the underlying data of the algorithm could thus result in biased predictive outcomes.

Moreover, a predictive model will rarely be a 100% accurate. Different risk thresholds yield a varying number of type 1 and type 2 errors, which are inversely related to each other. Depending on the situation, an appropriate risk threshold should be used. A third important limitation is the lack of transparency. Some models, like a neural network, do not demonstrate what its decisions are based on, which complicates interpretability (Schwartz et al., 2017).

Predictive risk modeling thus has several important advantages over traditional risk assessment methods. These advantages are promising and currently being exploited in many different industries. Important limitations of PRM are mainly related to biased data, prediction errors and possible low interpretability.

4.3 The potential of linked-administrative data

Exciting new opportunities for social science research arise from the use of administrative data sources, but these are currently under appreciated by the research community (Connelly et al., 2016). The main difference between administrative data and the data traditionally used in social sciences is characterized by the distinction between "found" and "made" data. Generally, social science research is based on data sources that are "made" by researchers, like experimental or observational data, which are specifically gathered for research purposes. Administrative data is derived from the operation of administrative systems and usually collected for the purpose of registration, transaction and record keeping. Besides, administrative data is usually associated with the delivery of a service (Connelly et al., 2016). Little attention has been given to the use of administrative data for social science research, because access to administrative records is usually restricted to the research community. The prior research to youth help usage is also mainly based on observational data (Bot et al., 2013; Appleyard et al., 2005). However, several governments, especially Nordic countries, the U.K. and the U.S., are recently making administrative data available to the research community (Connelly et al., 2016).

Linked-administrative data has certain characteristics that are especially suitable for the prediction of youth help usage. First, as mentioned before, administrative data is highly associated with service delivery. Secondly, the sample size of administrative data is usually much larger than for example of large social survey data and can sometimes even cover the entire population of interest (Card et al., 2010). This leads to a high amount of training and testing data for a predictive algorithm to develop, which eventually benefits model performance. Thirdly, administrative records provide access to information of groups who might be less likely to participate in primary social science research (Connelly et al., 2016). Fourthly, it is especially useful for studying "sensitive"

issues that people might like to disclose to a researcher in an experimental or observational study, like mental health problems or substance abuse (Goerge and Lee, 2002). At last, administrative data contains individual information, which is necessary for the individual identification of youngsters and their families. These issues are particularly important for the prediction of youth help usage.

However, besides several advantages, the use of administrative data for research has several challenges and limitations as well. First, there are some legal and privacy issues. Administrative data is not collected for the purpose of research and the public might have concerns with the linkage of their data from different sources and the use of their data for research purposes (Stevens and Laurie, 2014). For this reason, datasets should be constructed based on pseudonised records to ensure that individuals or families can not be identified by the researchers. Section 5 provides detailed information regarding this issue. Secondly, powerful computing capacity is needed in order to perform the analyses on massive amounts of data. Thirdly, administrative data requires an extensive process of preparing and shaping the data in order to perform the analyses, because the data is not collected for research purposes (Goerge and Lee, 2002). This process includes the construction of new variables, for which background knowlegde is essential. Fourthly, the reliability and accuracy of administrative records are influenced by the administrative accuracy of the institutions providing data. For example, policy changes in administrative policy could influence the correctness of administrative data. This requires scientists to check the process of data collection of the data providing institutions.

Several studies exist, which are using linked-administrative data in combination with predictive modeling, to predict certain outcomes closely related to youth help usage. Marshall and English (2000) used the neural network methodology applied to the Child Protective Services risk assessment data in order to predict the risk of child maltreatment. 37 variables, based on known risk factors, were substracted from administrative records and included in the model. The model produced a prediction accuracy of 81%. Vaithianathan et al. (2012) used a prospective population based predictive risk model to predict child abuse in New-Zealand. Linked-administrative records from work and income, child and family health and welfare systems were used. The model included 132 variables and yielded an overall model performance of AUC=76%. Schwartz et al. (2017) studied whether predictive analytics and machine learning could improve the accuracy and utility of the child welfare risk assessment instrument used in Broward County (Ft. Lauderdale, Florida). The study linked 58 datasets from the Broward Sherrifs Office, ChildNet and the Children's Services Council. The results of the study show that the accuracy of the risk assessment tool was substantially improved from 60% to 90% accuracy with an AUC of 0.81.

Linked-administrative data thus seems a valuable data source for the prediction of youth help usage, but currently underexploited in the research community. Linked-administrative data, in combination with a thorough data preparation process and contextual knowlegde, can provide a huge amount of objective individual data, sometimes including the whole population of interest. These characteristics are promising for predictive analytical approaches.

5 Data and Methods

This section will describe the data and the methods, used to develop the predictive models.

5.1 Data

Much information about important risk factors for youth help usage can be derived from administrative databases of the Central Bureau of Statistics. Thirteen databases with pseudonised individual administrative information about the registered Dutch population have been used in order to create a single linked-admnistrative dataset. Some databases of the Central Bureau of Statistics are longitudinal and some have seperate tabs for each year. The institutions which are providing data are mostly governmental, like: juvenile services institutions, municipal population registers or tax authorities. These institutions provide pseudonised administrative information about their clients.

As administrative juvenile services data is only available for the years 2015 to 2017, the constructed dataset contains information about the citizens, until the age of 23, who have lived in Rotterdam during these years. In order to select these people, the GBAADRESOBJECTBUS, the VSLGWBTAB and the GBAPERSOONTAB have been merged. These databases contain adress characteristics and age information of the Dutch population. The resulting output has repetitively been merged with other databases to construct the final dataset, containing the information from all the databases on an individual level. The number of observations in this dataset is 506.754.

The dataset contains a row for each person in every year. But, for the analysis, one row per person is selected in order to make the predictions. As we are aiming to predict the first use of youth help services, for youth help users, the most recent situation before first usage is used in the prediction analysis. So, for a person who entered a youth help trajectory in 2016, his personal situation before the start of the first trajectory is used. For people who have not participated in a youth help trajectory between 2015 and 2017, the most recent

situation in 2017 is used in the analysis. The final dataset for prediction contains 166.654 observations with an event rate of 12.7%.

A summary of the used databases can be found in table 2. Table 3 provides a description of the dataset, before the selection of a single row per person.

Table 2: Summary of the used databases

Database	*Description*
GBAADRESOBJECTBUS	Address characteristics of people who have been registered in the municipal population registers (GBA). Time unit: longitudinal since 1995. Methodology: The data is derived from the municipal population registers.
GBABURGERLIJKESTAATBUS	Marital status characteristics of people who have been registered in the municipal population registers (GBA). Time unit: Longitudinal since 1995. Methodology: The data is derived from the municipal population registers.
GBAHUISHOUDENSBUS	Household characteristics of people who have been registered in the municipal population registers (GBA). Time unit: Longitudinal since 1 October 1994. Methodology: The data is derived from the municipal population registers.
GBAOVERLIJDENTAB	The date of death of people who have been registered in the municipal population registers (GBA). Time unit: Longitudinal since 1995. Methodology: The data is derived from the municipal population registers.
GBAPERSOONTAB	Personal characteristics of people who have been registered in the municipal population registers (GBA) (e.g. gender, year of birth, country of birth). Time unit: Longitudinal since 1995. Methodology: The data is derived from the municipal population registers.
HOOGSTEOPLTAB	Highest graduated educational level and highest "followed" educational level and educational direction of the Dutch population. The target population is the GBA population. The database only contains records of people whose highest level of education is known. Time unit: since 1999. Methodology: The database is composed from several educational registers and supplemented with the yearly educational survey EBB (Since 1996) Educational registers: CHRIHO: Central register of higher education registrations (Since 1983), ERR: Graduation results register secondary education (Since 1999), Educational number files from: Secondary education, lower education and adult education, WSF: Files from student grants (Since 1995), UWV: Files from educational history as noted by job seekers.
INHATAB	Information about household incomes of the Dutch population. Time unit: yearly databases from 2011. Methodology: The data is obtained from administrative systems of the Dutch tax authorities and DUO.

Table 2: Summary of the used databases

Database	*Description*
KINDOUDERTAB	The persons registered in the municipal population register (GBA) and the identification numbers of their legal parents. Time unit: 2010-2016. Methodology: Data is derived from the municipal population register.
MEDICIJNTAB	Information about medication delivery, covered by the basic insurance of the Dutch population registered in the municipal population register (GBA). Time unit: 2006-2016. Methodology: The data is derived from the administrative system of the college for health insurance (CVZ).
VSLGWBTAB	Municipality, district and neighborhood codes of BAG objects (e.g. buildings or homes). Time unit: 1999- September 2017
JGDHULPBUS	Information about all the Dutch youth help trajectories. Time unit: 2015-2017. Methodology: The data is derived from the administrative systems of the Dutch youth help institutions. The data is updated and revised twice a year.
JGDBESCHERMBUS	Information about all the Dutch youth protection trajectories. Time unit: 2015-2017. Methodology: The data is derived from the administrative systems of the Dutch youth protection institutions. The data is updated and revised twice a year.
JGDRECLASBUS	Information about all the Dutch youth probation trajectories. Time unit: 2015-2017. Methodology: The data is derived from the administrative systems of the Dutch youth probation institutions. The data is updated and revised twice a year.

Table 3: Data description

Variable	*Class*	*Values*	*Source*	*Description*	*Risk factor*
URINPERSOON	Character	e.g R000000001	GBAPEROONTAB	Unique identifier of a person (Linking key)	
URINPERSOONMa	Character	e.g. R000000002 NA = 3.6%	KINDOUDERTAB	Unique identifier of a person's mother (Linking key)	
URINPERSOONpa	Character	e.g. R000000003 NA = 16.2%	KINDOUDERTAB	Unique identifier of a person's father (Linking key)	
year	Numeric	2015-2017		Year identifier (Linking key)	
GBAGESLACHT	Factor	Male = 50.9% Female = 49.1%	GBAPEROONTAB	Gender	Male gender
GBAHERKOMST GROEPERING	Factor	Autochtonous = 41.9% Non-Western-Immigrant = 46.4% Western-Immigrant = 11.7%	GBAPEROONTAB	Autochtonous: Both parents born in The Netherlands. Non-Western-Immigrant: Person originates from a non-western country based on birth country of one of the parent(s). Western-Immigrant: Person originates from a western country based on birth country of one of the parent(s).	Ethnic background
OPLNIVSOI2016 AGGHMETNIRWO	Factor	High = 6.5% Middle = 18.1% Low = 66.5% NA = 8.9%	HOOGSTEOPLTAB	Highest followed level of education in year x. High: HBO, WO, Doctor. Middle: Secondary education (except VMBO). Low: No education, primary education, VMBO	Educational level
ChildAddiction	Factor	No = 99.98% Yes = 0.02%	MEDICIJNTAB	Has a person had anti addictive drug prescriptions in his/her life. ATC4 = N07B. *Based on WHO ATC Classification.	Addiction within family

Table 3: Data description

Variable	*Class*	*Values*	*Source*	*Description*	*Risk factor*
ChildDisease	Factor	No = 96.5% Yes = 3.5%	MEDICIJNTAB	Has a person had drug prescriptions in 2015 that suggest a long term disease or condition. ATC4 = A09A, A14A, A14B, B01A, B05D, B05Z, D03B, H01A, H01B, H01C, L01X, L02B, L03A, L04A, N03A, N04B, N05A, N06A, N06B, N06C, N06D. *List is contructed in consultation with professionals.	Long term disease or condition
BirthDate	Date	e.g. 2012-01-01	GBAPEROONTAB	Birth date of a person	
age	Numeric	0-22 μ = 11.8 σ = 6.6	BirthDate year	The age of a person in year x.	
scaled_age	Numeric	e.g. 1.3	age	Scaled age	Age
Age_class	Factor	Age 0-3 = 18.4% Age 4-11 = 35.0% Age 12-18 (Puberty) = 29.0% Age 19-22 = 17.7%	age	Age categories separating ages with varying juvenile services usage and identifying puberty in year x.	Age
BirthDateMom	Date	e.g. 1970-01-01 NA = 2.6%	GBAPEROONTAB	Birth date of the mother	
child_teenmom	Factor	No = 96.4% Yes = 1% NA = 2.6%	BirthDate BirthDateMom	Is a person the child of a teen mom. (age difference with mother <= 18 years).	Teen mom
mom_teenmom	Factor	No = 96.3% Yes = 0.1% NA = 3.6%	BirthDate BirthDateMom	Is a person a teen mom. (age difference with child <= 18 years).	Teen mom

Table 3: Data description

Variable	*Class*	*Values*	*Source*	*Description*	*Risk factor*
SPF	Factor	No = 74.4% Yes = 25.6%	GBAHUIS-HOUDENSBUS	Has a person been part of a single parent family for a time period over two years in his/her life.	Single parent family
HouseInc	Numeric	μ =€ 71.865 σ = 83.630 NA = 7.2%	INHATAB	Gross household income of a person's household in year x.	
scaled_HouseInc	Numeric	e.g. 1.3	INHATAB	Scaled HouseInc	Household income
Poverty	Factor	No = 71.1% Yes = 21.7% NA = 7.2%	HouseInc	Is a person's household income below the poverty threshold in year x. If SPF: HouseInc <26179 Not SPF: HouseInc <33696	Poverty
ParDivorce	Factor	No = 71.6% Yes = 11.6% NA = 16.8%	GBABURGER-LIJKESTAATBUS	Have the parents of a person been divorced.	Parental Divorce
ParentalDisease	Factor	No = 77.7% Yes = 19.3% NA = 3.0%	MEDICIJNTAB	Have one or both of a person's parents had drug prescriptions in 2015 that suggest a long term disease or condition. ATC4 = A09A, A14A, A14B, B01A, B05D, B05Z, D03B, H01A, H01B, H01C, L01X, L02B, L03A, L04A, N03A, N04B, N05A, N06A, N06B, N06C, N06D. *List is contructed in consultation with professionals.	Parental long term disease or condition
ParentalDeath	Factor	No = 96.8% Yes = 0.2% NA = 3%	GBAOVERLIJDEN-TAB	Have one or both of the parents died within two years before year x.	Death of a loved one

Table 3: Data description

Variable	*Class*	*Values*	*Source*	*Description*	*Risk factor*
ParentalAddiction	Factor	No = 98.5% Yes = 1.2% NA = 3%	MEDICIJNTAB	Have one or both of the parents of a person had anti addictive drug prescriptionsin his/her life. ATC4 = N07B. *Based on WHO ATC Classification	Addiction within family
ParentalEducation	Factor	High = 29.9% Middle = 33.4% Low = 19.8% NA = 16.9%	HOOGSTEOPLTAB	Highest graduated level of education of the parent with the highest graduated level of education in year x. High: HBO, WO, Doctor. Middle: Secondary education (except VMBO). Low: No education, primary education, VMBO	Parental educational level
YouthHelp	Factor	No = 92.3% Yes = 7.7%	JGDHULPBUS	Has a person made use of any of the youth help services in year x.	Youth help usage
YouthProtection	Factor	No = 98.4% Yes = 1.6%	JGDBESCHERM-BUS	Has a person made use of any of the youth protection services in year x.	Youth protection usage
YouthProbation	Factor	No = 99.4% Yes = 0.6%	JEUGDRECLAS-BUS	Has a person made use of any of the youth probation services in year x.	Youth probation usage

5.1.1 Handling of missing values

As indicated in table 3, several variables contain a large number of missing values. These values can be missing for several reasons and will be treated as appropiately as possible. An important reason for missing parental information is immigration. Administrative information about parents (born) in other countries is often incomplete. Observations with missing values arising from this issue can not be totally excluded from the analysis, because that would lead to the structural exclusion of immigrants in our analysis. But, in most cases, information about at least one of the two parents is available. Only for a select group of first generation immigrants (N=5023), all parental and household information is missing. For this group, 50% of all the predictors had to be estimated. Therefore, this group is excluded from the analysis. A second important reason for missing data is that many of the databases from the Central Bureau of Statistics are not updated until the year 2017. For example, the database containing information about household incomes is updated until 2016. This leads to missing values of all household incomes in 2017. Nevertheless, many missing values can be reasonably estimated.

URINPERSOONMa and URINPERSOONpa
For missing values of URINPERSOONMa and URINPERSOONpa the "-" symbol is imputed, because these variables are linking keys and not predictors in the analyses. If both URINPERSOONMa and URINPERSOONpa are missing, the record is excluded from the dataset.

Child of teen mom and teen mom
If URINPERSOONMa is unknown, the birth date of the mother is unknown and so the child_teenmom and mom_teenmom variables are unknown. These missing values are imputed with "No", because if the biological mother of a child is unknown and not registered in the dutch population registers, then it is unlikely that this woman is still actively involved in the child's raising. When the woman is not involved the child's raising anymore, the risk factor should not be accounted for.

Educational level
The child's highest followed level is unknown for most of the children below the age of five, because compulsory education in The Netherlands starts at the age of five. So, for children below the age of five with missing values, the new factor level "unknown (pre-school age)" is created. Missing values of children between the age of five and twelve are assigned to factor level "Low", because primary education is compulsory for Dutch citizens of this age. Missing values of children older than 12 are assigned to factor level "Middle", because secondary education is still mandatory until the age of eighteen. After the age of 18, children with missing values will at least have followed secondary education, so this group is also assigned to factor level "Middle".

Household income and poverty
For year 2017, all the observations have missing values, because the household income information is only available until 2016. For 2017, the average household income per person of 2015 and 2016 is imputed. If there is no household income information available for the years 2015 or 2016, the missing value is imputed with the mean household income per ethnic background group. Table 4 provides an overview of the household incomes per ethnic background group.

Table 4: Average income per ethnic background

Ethnic background	*Average household income (€)*
Autochthonous	92.876
Western immigrant	71.904
Non-Western immigrant	52.903

Parental divorce
Parental divorce contains missing values when one of the two biological parents are unknown. In these cases, it is unlikely that the child has experienced a parental divorce. For this reason, missing values arising from unknown URINPERSOONMa or URINPERSOONpa are imputed with "No".

Parental Education
The percentage of missing values of the parental education variable is exceptionally high, namely: 16.9%. Furthermore, the absence of this information occurs much more often for immigrants than autochtonous persons. Imputing the mean educational level for such large quantities leads to biased results. Besides, the reliability of the data of this variable is low, because due to a large number of missing values, the administrative information is supplemented with several kinds of observational information, like survey data. Due to the large number of missing values and the low data quality, the variable is excluded from the analysis.

5.1.2 Data description of Mieloo's model

In the second part of the analysis, the best performing predictive model from the first part of the analysis, will be compared with the multivariate logistic regression model from Mieloo et al. (2013), based on the same test set. The predictors in this study are very similar to the predictors in our main analysis, but structured slightly different. Table 5 describes the data structure of Mieloo's model applied to our test set.

Table 5: Data structure of Mieloo's model

Variable	*Class*	*Values*	*Source*	*Description*	*Risk factor*
URINPERSOON	Character	e.g R000000001	GBAPEROONTAB	Unique identifier of a person (Linking key)	
GBAGESLACHT	Factor	0 = Female = 49.1% 1 = Male = 50.9%	GBAPEROONTAB	Gender	Male gender
SPF	Factor	0 = No = 73.8% 1 = Yes = 26.2%	GBAHUIS-HOUDENSBUS	Has a person been part of a single parent family for a time period over two years in his/her life.	Single parent family
HouseInc	Factor	0 = High = 39.4% 1 = Low = 60.6%	INHATAB	Gross household income above or below average.	Low household income
OPLNIVSOI2016 AGGHMETNIRWO	Factor	0 = High = 88.8% 1 = Low = 11.2%	HOOGSTEOPLTAB	Highest followed level of education in year x. Low: Practical education, VMBO-b/k, MBO-1/2. High: Other education	Low educational level
YouthHelp	Factor	0 = No = 87.1% 1 = Yes = 13.9%	JGDHULPBUS	Has a person made use of any of the youth help services in year x.	

Table 6 describes the parameter values that are estimated to the test set for comparison.

Table 6: Parameter values of Mieloo's model, source: Mieloo et al. (2013)

Variable	*B*
GBAGESLACHT	0.248
SPF	0.850
HouseInc	0.364
OPLNIVSOI2016 AGGHMETNIRWO	0.976
Constant	-4.697

5.2 Research methods

This study has an exploratory nature and therefore an inductive research approach (Buchdahl, 1956). A predictive risk modeling approach is applied and several supervised classification techniques are compared in order to develop the best model for the prediction of youth help usage.

The analysis will consist of three parts. In the first part (paragraph 6.1), six classification techniques are applied to the data in order to develop the best performing model for the prediction of youth help usage. Because youth help usage can be seen as a "rare event", and only occurs in 12.7% of our population, an unbalanced sample exists. For this reason, the models are trained on unbalanced and balanced data, in order to examine which method yields a

higher performance on the real-world test data. For balancing the training data, the Synthetic Minority Over-sampling Technique is applied.

In the second part (paragraph 6.2), a multivariate logististic regression model that could be used for prediction, formed in earlier research on the demand of youth help services (Mieloo et al., 2013), will be estimated to and evaluated on our test set. The data structure of this model is described in table 5. Thereafter, the best performing classification technique from the first part of the analysis is applied to the same test data, using the same predictors. The increase in predictive performance between the traditional model, the best performing classification technique, using the same variables, and the best performing model from the first part of the analysis, provides information to assess the contribution of a predictive risk modeling approach, applied to administrative data, to the prediction of youth help usage.

In the third part (paragraph 6.3), the predictive performance of the models is evaluated when historical youth help usage, historical youth protection usage and historical youth probation usage are included as predictors in the analysis.

This section describes the Synthetic Minority Over-sampling Technique, the validation process, the applied classification techniques, the performance measures and the variable importance measures.

5.2.1 SMOTE: Synthetic Minority Over-sampling Technique

When the classification categories of a dataset are not approximately equally distributed, a dataset is imbalanced (Chawla et al., 2002). In machine learming, real-world imbalanced datasets often occur with the prediction of "rare" events, like fraud or rare diseases (Chawla et al., 2002). The training of a predictive model, based on an imbalanced dataset, severely reduces model performance. In our dataset, only 12.7% of the population has made use of youth help services. For this reason, the Synthetic Minority Over-sampling Technique is applied in order to create a more balanced dataset and improve model performance. The Synthetic Minority Over-sampling Technique is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Neighbors from the k nearest neighbors are randomly chosen, depending upon the amount of over-sampling required (Chawla et al., 2002).

5.2.2 The validation process

First, the dataset is divided into a training set, containing 80% of the observations, and a test set, containing 20% of the observations. Then, the Synthetic Minority Over-sampling Technique is applied to the training set, in order to balance the data. Thereafter, ten-fold cross-validation is applied to the balanced training data and the unbalanced training data, in order to train the six classification models. The performance of the models, based on the cross-validated dataset, is evaluated. However, within the process of cross-validation, the performance of the models on the test folds of the balanced training data, is based on balanced test data, whereas in the real-world situation the distribution of youth help users is not balanced. Thus, the performance of the models, based on balanced training data, describes the ability of the models to predict youth help usage under the false assumption of a balanced distribution of youth help usage. For this reason, the performance of the models trained on balanced data and the performance of the models trained on unbalanced data will be evaluated on the initial unbalanced test set. The performance of the models on the unbalanced test set will be more important for this study, as the real-world distribution of youth help usage is unbalanced. Because the explicit future use of the model has not yet been determined, the intuitive risk threshold of 0.5 is applied, in order to classify the observations as positive.

5.2.3 Classification techniques

Six classification techniques are applied to the data, in order to predict youth help usage. The performance of these models is evaluated and compared, based on several performance measures, which will be discussed in paragraph 5.2.4. The six applied classification techniques are: classification tree, logistic regression, neural network, naïve bayes, random forest and gradient boosted machine. All these methods have another approach in the segmentation of data.

The classification tree consist of nodes and branches. Each node contains the test of an attribute, with each branch from the node representing a distinct value of the attribute. A unique path is created by following the branches down the tree. Each path eventually ends up at a leaf. Based on its attributes, every observation in the dataset will belong to only one leaf from the tree. All observations in the same leaf have are classified with the same value in relation to the target variable (Provost and Fawcett, 2013). The logistic regression model is not a tree-based model and predicts the likelihood of a certain event based on a linear combination of independent variables (Provost and Fawcett, 2013). The neural network implements complex non-linear numeric functions and works with several layers. In each layer a new model is formed and the output is

passed on to the next layer. The next layer forms a new model, improved by the output of the previous layer. The combined output of several layers leads to the final estimation of the model parameters (Provost and Fawcett, 2013). The naïve bayes classifier is a simple probabilistic classifier, which is based on the application of Bayes' theorem and the assumption of conditional independence. The model classifies each new observation by estimating the class probability for each class and selects the class with the highest probability (Provost and Fawcett, 2013). The random forest model is a tree-based model and applies bootstrap aggregation in order to improve the predictions. The model combines the result of multiple decision trees, which are constructed by the random selection of a partial set of the total number of predictors. Every model, thus, has a different set of predictors. This technique reduces the model's variance at the expense of a slightly higher bias (van Dalen, 2018). The gradient boosted machine is also a tree-based model, but applies boosting techniques in order to improve the predictive performance of the tree. Boosting involves the creation of multiple models. Each model is improved specifically based on the prediction errors of the previous model. Thus, the model improves the areas where it didn't perform well before. Boosting reduces variance and bias, but can be sensitive to over-fitting (van Dalen, 2018).

Subsequently, in order to reduce the overall risk of overfitting, ten-fold cross-validation is applied to all the models. Cross-validation is a method to estimate the accuracy of an inducer by dividing the data into k mutually exclusive subsets of approximately equal size. The model is trained and tested k times on different folds of the data. The performance estimate is the mean of the performance of the k number of folds(Provost and Fawcett, 2013).

5.2.4 Performance measures

In order to evaluate the performance of the applied models, several classification performance measures will be used. This paragraph will describe the performance measures in detail and discuss the importance for the selection of the best model.

Confusion matrix
The confusion matrix of a binary classifier is a contingency table with columns labeled as: predicted class and actual class. It provides an overview of the ratio of correctly and incorrectly predicted observations. From the confusion matrix, other performance measures can be derived. Figure 2 provides an example of the confusion matrix.

Figure 2: Confusion matrix

		Actual value		
		0	**1**	**total**
Predicted outcome	**0**	True negative (TN)	False negative (FN)	
	1	False positive (FP)	True positive (TP)	
	total			

Accuracy
The accuracy of a classification model is defined as the proportion of correct predictions (Provost and Fawcett, 2013). Accuracy is calculated by:

$$Accuracy = \frac{\text{Correctly predicted observations}}{\text{All observations}} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy (1) is a simple measure of performance, but can be misleading with unbalanced samples (Provost and Fawcett, 2013).

Sensitivity
The sensitivity of a model is also called the "True positive rate" and is defined as the proportion of actual positives that are also predicted as such (Provost and Fawcett, 2013). Sensitivity is calculated by:

$$Sensitivity = \frac{\text{True positives}}{\text{All actual positives}} = \frac{TP}{TP + FN} \quad (2)$$

Specificity
The specificity of a model is also called the "True negative rate" and is defined as the proportion of actual negatives that are also predicted as such (Provost and Fawcett, 2013). Specificity is calculated by:

$$Specificity = \frac{\text{True negatives}}{\text{All actual negatives}} = \frac{TN}{TN + FP} \quad (3)$$

Precision
The precision of a model is also called "Positive predictive value" and is defined as the ability of a model to correctly predict positive values (Provost and Fawcett, 2013). Precision is calculated by:

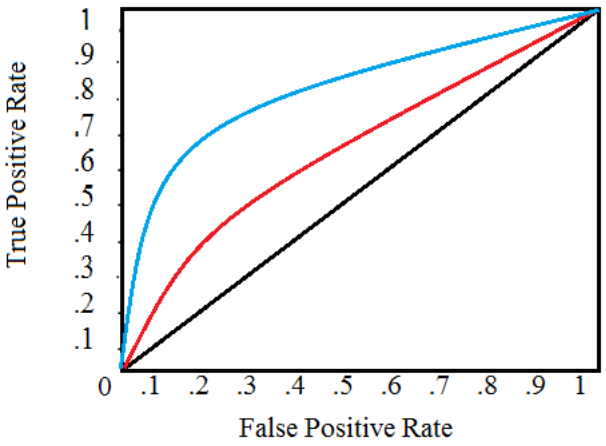
$$Precision = \frac{\text{True positives}}{\text{All predicted positives}} = \frac{TP}{TP + FP} \tag{4}$$

F1 score
The F1 score measures the performance of a model as the harmonic average between precision (4) and sensitivity(2). The outcome is a value between zero and one, whereas one would implicate perfect precision and sensitivity. The F1 score is calculated by:

$$F1 = 2 * \frac{\text{precision * sensitivity}}{\text{precision + sensitivity}} \tag{5}$$

ROC curve
The ROC curve is graphical illustration of model performance. The ROC curve is created by plotting the true positive rate against the false positive rate for multiple risk thresholds (Provost and Fawcett, 2013). The further away the curve is from the diagonal, the higher the performance of a model. Figure 3 provides an example of an ROC curve.

Figure 3: Example ROC curve



Area Under the ROC Curve (AUC)
The "Area Under the ROC curve" is a general summary statistic of the performance of a classifier. It simply represents the area under the classifier's curve, expressed as a fraction of the unit square (Provost and Fawcett, 2013).The AUC measures the probability that a classifier will rank a randomly chosen positive observation higher than a negative one. The outcome of the AUC is always between zero and one (Schwartz et al., 2017).

Relating to the practical implementations of this study, the best performing model will be chosen based on it's ability to identify future youth help users (Positive values). For this reason, the sensitivity, precision and F1 score will be leading in the decision.

5.2.5 Variable importance

Variable importance in section 6 is measured in two ways: the "mean decrease in accuracy" and the "mean decrease in gini-coefficient". Moreover, the outcomes of these measures are graphically illustrated in a variable importance plot. The mean decrease in accuracy is computed from permuting out-of-bag data. For each tree, the error rate on the out-of-bag portion is computed, before and after permuting each predictor variable. The difference between the two is then averaged over all trees and normalized by the standard deviation of the differences. The predictors with the highest increase normalized increase in error rate, are considered as the most important variables in the prediction (Archer and Kimes, 2008). The mean decrease of the gini-coefficient determines variable importance based on the total decrease in node impurities, from splitting on this variable, averaged over all trees. The decrease in node impurities is measured by the gini-coefficient (Archer and Kimes, 2008).

6 Prediction analysis

This section will first develop and compare six classification models, based on balanced and unbalanced training data. Thereafter, the best performing model will be chosen. Then, the performance of the best model will be compared with the performance of a model from earlier research, applied to our testing data. At last, traditional juvenile services usage will be added as predictors to evaluate it's consequence on model performance.

6.1 The best performing model

Table 7 provides an overview of the performance of the classification models on the training set, based on unbalanced training data. The performance measures show that the models are quite accurate, mostly strong in predicting negative cases, considering the specificities around 99%, but weak in predicting positive cases, considering the sensitivities between 12% and 20%. Also, the F1 score is low.

Table 7: Performance of unbalanced models on training data

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1*	*AUC*
Classification tree	0.881	0.160	0.990	0.703	0.260	0.686
Logistic regression	0.882	0.131	0.991	0.688	0.220	0.762
Neural network	0.888	0.176	0.991	0.749	0.284	0.793
Naive bayes	0.870	0.201	0.967	0.472	0.282	0.745
Random forest	0.888	0.174	0.992	0.759	0.284	0.705
Gradient boosted machine	0.883	0.127	0.993	0.720	0.216	0.767

Table 8 provides an overview of the performance of the classification models on the training set, based on balanced training data. The performance measures show that, in general, the performance of the models increases and specifically the ability of the models to predict positive cases. The accuracy and specificity of the models decrease, but the sensitivity, precision, F1 and AUC increase.

Table 8: Performance of balanced models on training data

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1*	*AUC*
Classification tree	0.771	0.739	0.800	0.766	0.752	0.832
Logistic regression	0.746	0.727	0.762	0.731	0.729	0.818
Neural network	0.807	0.779	0.833	0.806	0.792	0.895
Naive bayes	0.717	0.864	0.585	0.650	0.742	0.808
Random forest	0.839	0.807	0.867	0.844	0.854	0.919
Gradient boosted machine	0.764	0.802	0.730	0.725	0.762	0.850

Also, the ROC-plots presented in figure 4 and figure 5 show that the models based on balanced training data yield a higher performance on the training set.

Figure 4: ROC-plot unbalanced models on training data

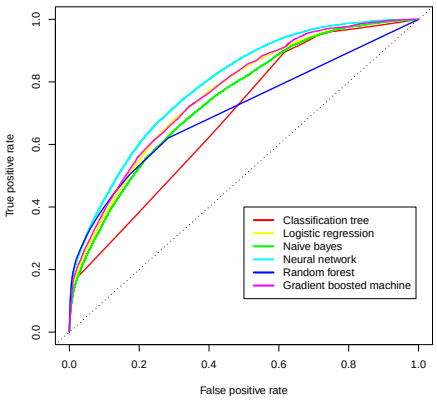


Figure 5: ROC-plot balanced models on training data

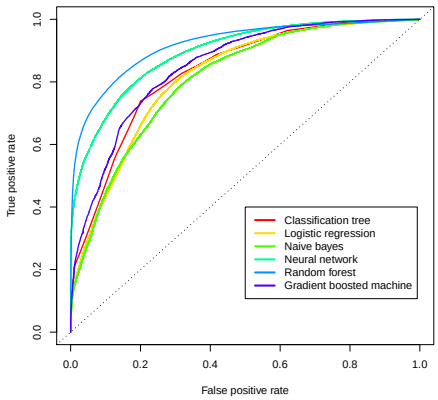


Table 9 provides an overview of the performance on the test set of the models, trained on unbalanced training data and table 10 provides an overview of the performance on the test set of the models, trained on balanced training data. The sensitivity, precision and F1-score show that, also on unbalanced test data, the performance of the models trained on balanced data is much higher than for the models trained on unbalanced data. In terms of AUC, the performance is quite similar. In terms of accuracy and specificity, the performance is lower.

Table 9: Performance of unbalanced models on test data

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1*	*AUC*
Classification tree	0.884	0.170	0.990	0.717	0.275	0.694
Logistic regression	0.881	0.139	0.991	0.704	0.232	0.766
Neural network	0.887	0.189	0.991	0.756	0.303	0.799
Naive bayes	0.870	0.205	0.971	0.509	0.292	0.748
Random forest	0.890	0.196	0.994	0.822	0.316	0.820
Gradient boosted machine	0.882	0.132	0.993	0.740	0.224	0.770

Table 10: Performance of balanced models on test data

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1*	*AUC*
Classification tree	0.760	0.513	0.797	0.273	0.356	0.716
Logistic regression	0.740	0.591	0.763	0.270	0.371	0.755
Neural network	0.789	0.538	0.826	0.315	0.397	0.781
Naive bayes	0.610	0.741	0.591	0.212	0.330	0.737
Random forest	0.822	0.512	0.868	0.367	0.427	0.804
Gradient boosted machine	0.713	0.579	0.732	0.244	0.343	0.744

Similar to and logically consistent with the AUC, the ROC-plots presented in figure 6 and figure 7 show that the overall performance on the unbalanced test set of the models, trained on balanced data and trained on unbalanced data, is quite similar, when graphically illustrated in the ROC-plot.

Figure 6: ROC-plot unbalanced models on test data

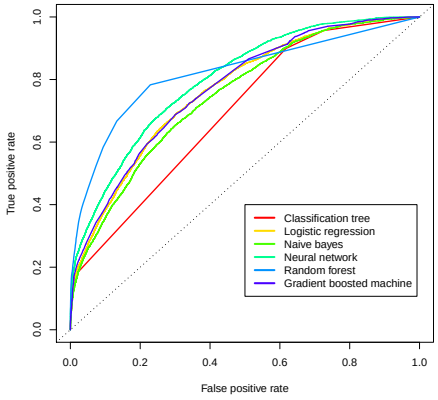
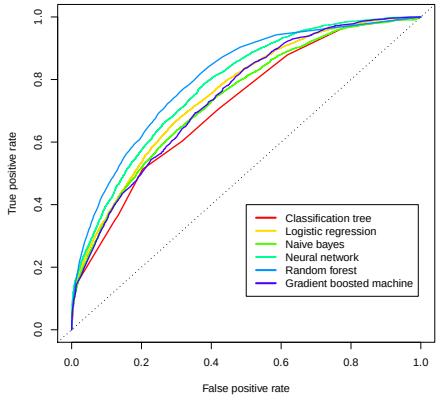


Figure 7: ROC-plot balanced models on test data

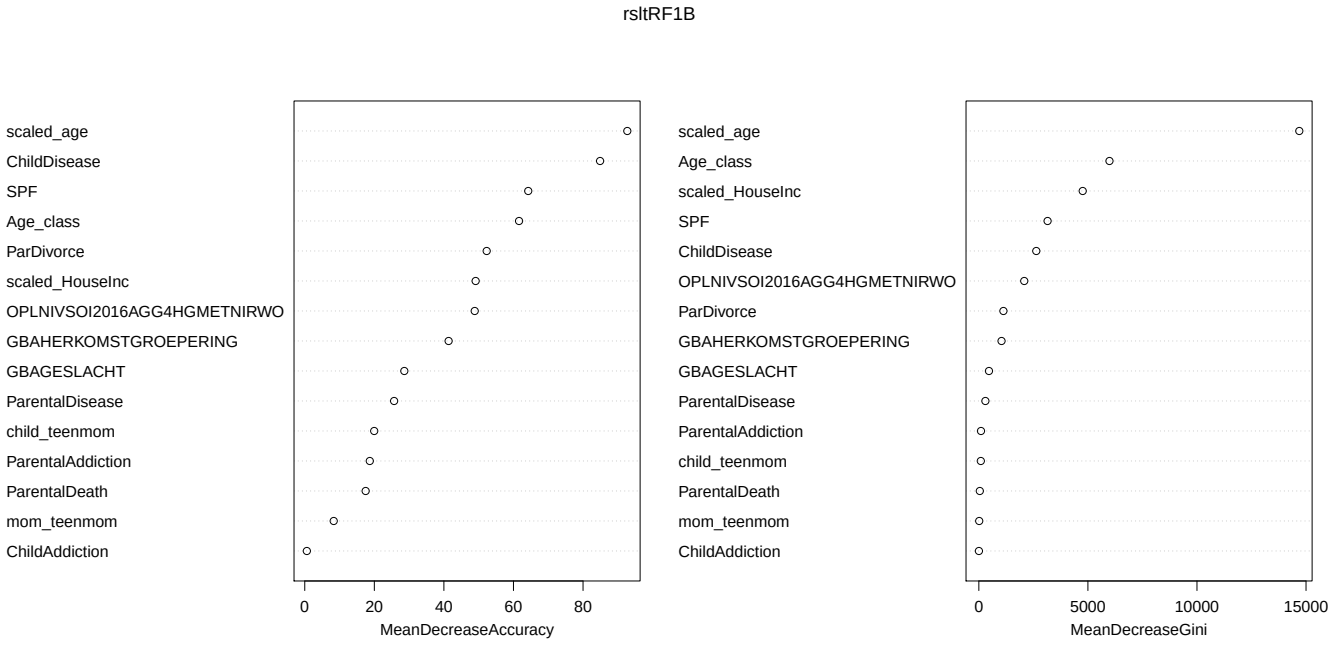


Thus, considering the sensitivity, precision and F1 of the models on the unbalanced test data, the random forest model, based on balanced training data, is outperforming the other models in terms of predictive performance. Hence, this model is considered as the best performing model in the first part of the analysis and will be used in paragraph 6.2 for further comparison.

6.1.1 Variable importance

In figure 8 the variable importance plot of the random forest model is shown. This figure shows that age (continuous), age-class, long-term disease or condition, having been part of a single parent family, household income and parental divorce are the six most important predictors for youth help usage in this model. These factors are a combination of child, parental and household factors.

Figure 8: Variable importance plot



6.2 The comparison with Mieloo's model

Table 11 provides an overview of the performances of the three models based on the same test set. The outcomes of the logistic regression seem typical with a sensitivity, precision and F1 of 0 and a specificity of 1. These outcomes are explained by the fact that the chosen risk threshold that classifies an observation as positive, is above the risk of 0.5. The summary of the predicted risk scores of the models, provided in table 12, shows that the logistic regression model with a limited set of predictors, only has the ability to risk stratify observations between the risk of 0.009 and 0.105. Thus, the classification

threshold of 0,5 is never reached and no observations are classified as positive. Consequently, the predictive performance of this model is low. Also, the mean risk of youth help usage, as predicted by the model, is much lower than the event rate in our population (12.7%).

When a random forest model, trained on balanced training data, is applied to the same test data, using the same data structure as the logistic regression model from Mieloo et al. (2013), a gain in predictive performance is seen, in terms of sensitivity, precision and F1-score. Besides, the model is able to classify observations above the risk threshold of 0.5 and the mean risk of youth help usage, as predicted by the model, is closer to the event rate in our population (12.7%).

At last, when all the predictors are included in the random forest model, and when some of the variables from Mieloo's model are structured slightly different, the performance increases to the performance of the best model from paragraph 6.1. Also, the mean risk of youth help usage, as predicted by the model, is the closest to the event rate in our population (12.7%).

The increase in predictive performance between the three models, illustrated by ROC-curves, is shown in figure 9

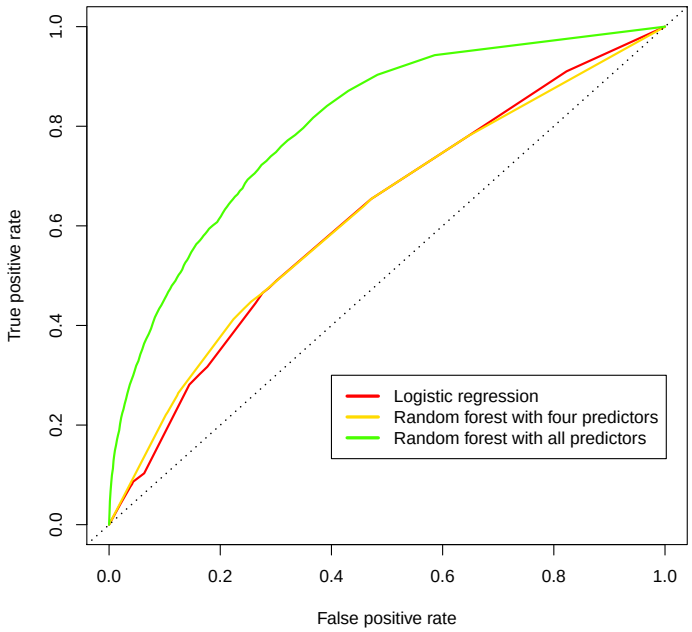
Table 11: Model performance on test data

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1*	*AUC*
Logistic regression	0.871	0	1	0	0	0.621
Random forest with four predictors	0.729	0.412	0.776	0.215	0.283	0.623
Random forest with all predictors	0.822	0.512	0.868	0.367	0.427	0.804

Table 12: Summary of predicted risk scores

Model	*Min*	*Mean*	*Max*
Logistic regression	0.009	0.022	0.105
Random forest with four predictors	0.000	0.279	1.000
Random forest with all predictors	0.000	0.193	1.000

Figure 9: ROC-plot of models on test data



6.3 The contribution of historical juvenile services usage

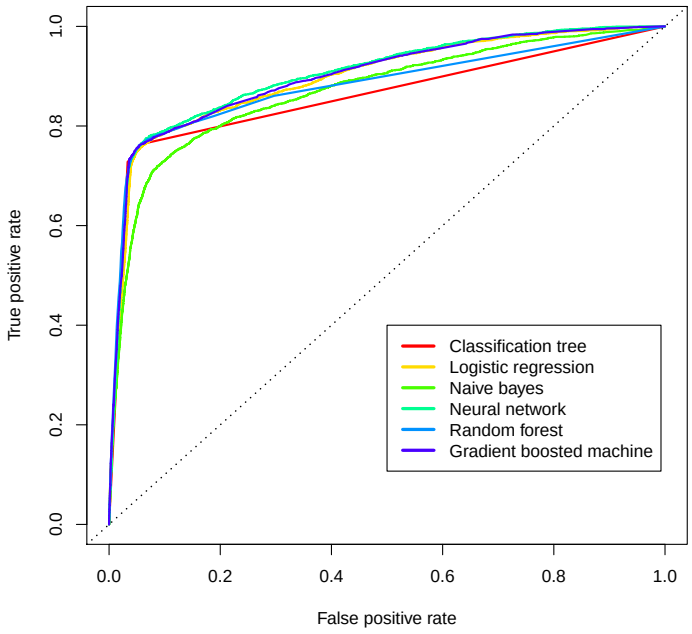
The main goal of this study is to predict first youth help usage, as this is relevant for prevention policies and forecasting the inflow of new entrants in youth help trajectories. However, during the analysis we added historical youth help usage, historical youth protection usage and historical youth probation usage as predictors in the models, in order to investigate it's effect on the performance of the models. The models now predict the total number of youth help users in a certain year, including the people who have been part of a youth help trajectory in the preceding year(s).

Table 13 shows that the predictive performance of the models increases when historical juvenile services usage is added to the predictors. Besides, with this new data structure, the neural network becomes the best performing model for the classification of youth help usage. It is the best performing model in terms of: accuracy, specificity, precision, F1-score and AUC. Only in terms of sensitivity, some models have a slightly higher performance. Figure 10 shows the predictive performance of the models, illustrated by ROC-curves.

Table 13: Model performance including historical juvenile services usage

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1*	*AUC*
Classification tree	0.926	0.784	0.938	0.501	0.611	0.871
Logistic regression	0.940	0.766	0.954	0.571	0.654	0.900
Neural network	0.943	0.762	0.957	0.587	0.663	0.906
Naive bayes	0.914	0.699	0.931	0.448	0.546	0.875
Random forest	0.942	0.768	0.956	0.584	0.664	0.891
Gradient boosted machine	0.943	0.764	0.957	0.588	0.664	0.901

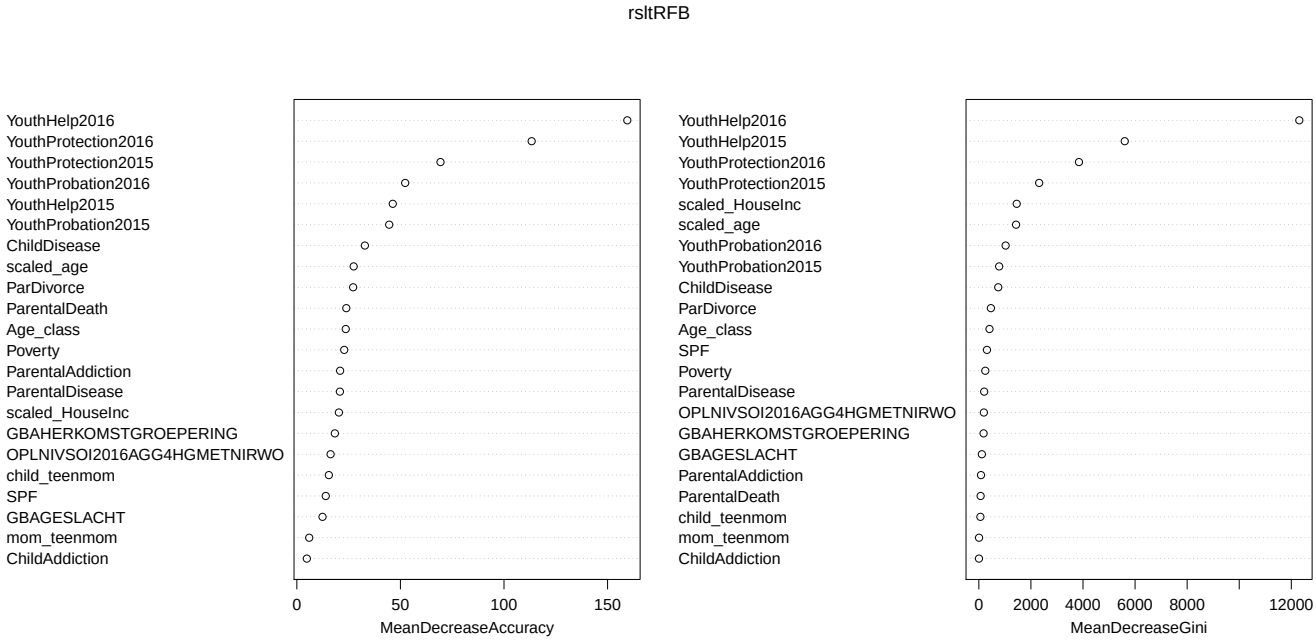
Figure 10: ROC-plot of models with historical juvenile services usage



6.3.1 Variable importance

Consistently with the increase in model performance after adding historical juvenile services usage to the predictors, the variable importance plot in figure 11 shows that the historical juvenile services usage predictors are generally the most important predictors for youth help usage.

Figure 11: Variable importance, including historical juvenile services usage



6.4 Practical consequences

At the moment, the development of children is monitored by institutions, like: CJG or schools. These institutions provide information to all their clients, relating to the healthy development of children. Now, in theory, the model can serve to identify specific children and families with a certain risk to enter a youth help trajectory in the following year. These families could be approached for appropiate prevention measures. However, this holds under the assumption that a personal prevention policy is desired by the municipality and that the pseudonised information can legally be made personal again.

An important aspect in the choice of a suitable risk threshold for classification is the nature of the prevention measure. The prediction errors will have a different consequence for every prevention measure. For every measure, it is important to choose an appropiate risk threshold, resulting in a varying number of type 1 and type 2 errors.

For example, if the municipality of Rotterdam would like to spread some flyers with extra information regarding youth help services or child- and parenting

problems, to all the youngsters with more than 50% chance of entering a youth help trajectory in the following year, the result would be as follows:

Consistent with the performance of the random forest model in table 10, approximately 51% of all the people who would actually enter a youth help trajectory in the following year, would receive the flyer. Besides, from all the youngsters receiving a flyer, approximately 63% would not enter a youth help trajectory in the following year. With regard to a flyer with information, one could argue that it is morally accepted to 'falsly' approach this 63%. In terms of cost-effectiveness, 37% of the investment would be spent effectively.

However, with a more confronting prevention policy, like a conversation with a youth help services employee, the cost of being falsly classified as a future youth help user increases, because it could emotionally hurt children and families. In this situation a higher risk threshold would be appropiate. For example, a high risk threshold of 80% would decrease the model's sensitivity and increase it's precision. Now, approximately 26% of all the people who would actually enter a youth help trajectory in the following year, would be approached. Besides, 54% of all the people being approached, would be correctly approached. Besides, in terms of cost-effectiveness, 54% of the budget would be spent effectively.

Summarizing, the best performing model could be used for personal prevention policies. However, with regard to the nature of the policy, it is important to consider the appropiate risk threshold for classification and to consider the moral impications and budget effects of falsly classified observations. Is it important to know that type 1 and type 2 errors are inversely related to each other.

7 Discussion and conclusion

7.1 Main findings

The results of the final predictive model clearly show that a predictive risk modeling approach, applied to linked-administrative data, significantly increases the prediction of youth help usage. The random forest model is the best performing model among the six classification models that have been compared. The six most important predictors in this model are a combination of life events, child, parental and household factors, namely: age, age-class, having been part of a single parent family, household income, having a long-term disease or condition and parental divorce. Besides, if historical juvenile services usage is added to the predictors in the analysis, the predictive

performance significantly increases and the historical juvenile services predictors become the most important predictors among all variables.

7.2 Discussion

When compared to Mieloo's model (Mieloo et al., 2013), formed in prior research to youth help usage, this study shows that the application of a predictive risk modeling approach, applied to linked-adminstrative data, significantly increases the prediction of youth help usage. Mieloo's model was able to risk stratify the population between the risk of youth help usage of 0.9% and 10.5%. So, for a risk threshold above 10.5% the model classifies all observations as negative, resulting in a model which is not able to correctly classify a single future youth help user above this risk threshold. For this reason, the sensitivity, precision and F1-score of this model, which are the most important performance measures in this study, are zero when the same risk threshold of 50% is applied. On the other hand, the predictive performance of a random forest model, applied to the same data structure, trained on balanced training data, already increases the predictive performance to a sensitivity of 41%, a precision of 22% and an F1-score of 0.28. Moreover, when, consistently with a predictive risk modeling approach, a large set of risk factors is added to the predictors in the dataset, the model performance increases to a sensitivity of 51,2%, a precision of 36.7% and a F1-score of 0.43. Besides, in terms of AUC, where model performance for different risk thresholds is included in one measure, the model performance increases from 0.62 to 0.80.

With regard to variable importance, this study finds that the most important predictors for youth help usage are a combination of life events, child, parental and household factors, namely: age, age-class, having been part of a single parent family, household income, having a long-term disease or condition and parental divorce. Contrary to these results, prior research generally found that the child factors were consistently more important than other factors related to youth help usage (Bot et al., 2013; Hermanns et al., 2005). Besides, this study shows that historical youth help usage, historical youth protection usage and historical youth probabation usage are most important among all predictors, when they are included in the analysis. However, not all risk factors that have been found in prior research have been included in our analysis and variable importance in a random forest model is calculated differently than in a logistic regression model, due to its tree-based nature.

7.3 Managerial implications

The results of this study have several practical consequences for the municipality of Rotterdam. As described in paragraph 6.4, the main consequence of this study is the ability to early identify youngsters and families with a certain risk to enter a first youth help trajectory in the following year. These individuals could now, theoretically, be approached for individual prevention measures. Besides, the allocation of resources among the neighborhood teams could be improved by locating the neighborhoods in which these individuals are most likely to receive support. The implementation of individual prevention measures has, however, some important managerial consequences that should be considered.

First, there are some concerns relating to privacy and the use of personal data. A characteristic of administrative data is that it is not collected for the purpose of research. The "participants" of research, based on administrative data, have usually not specifically approved to their data usage for research. For this reason, the research data is pseudonised. If a personal prevention policy would be preferred, the advantages and disadvantages of this policy should be thoroughly weighed and the legal and moral aspects relating to privacy and personal data usage should be highly considered.

Secondly, moral issues arise from the fact that the PRM-model classifies children and families as families with a high risk of child and parenting problems, while no child or parenting problems have yet been observed. The model produces the risk scores based on personal information that is just related to an increased chance of child and parenting problems, like household income or gender. The results of the model are never 100% accurate and thus, some families will be falsly classified as families that are likely to be in need of youth help services next year. On the other hand, children and families in need of support could be earlier identified and possibly be prevented from entering youth help trajectories. The correctly classified children and families will thus, in most cases, benefit from an individual prevention policy. Besides, the effect of falsly classified observations is dependent on the nature of the prevention policy. As mentioned in paragraph 6.4, with a less confronting prevention policy, the effect of falsy classified observations may be negligible. Thus, whether our model could serve as an operational tool is fully dependent on the nature of the implemented measure. For less confronting prevention measures we argue that the model is sufficiently accurate to be used, but for more confronting policies, the performance of the model should first be improved, because the precision of the model is still low. The moral dilemma's related to the implementation of predictive tools by government agencies, using personal data of its citizens, are part of an ongoing discussion in society and the scientific field (Shroff, 2017; Hurley, 2018)

Thirdly, at the moment, the data from the Central Bureau of Statistics is used and only available for this specific study. If a PRM-tool would operationally be desirable, another collaborative structure would be necessary. The data should continuously be available, with records containing personal information. Besides, the merging and data preparation of new incoming data should be automated.

7.4 Limitations and future research

In this study, fifteen risk factors have been included as predictors in the analysis. Information about these risk factors is substracted from the databases from the Central Bureau of Statistics, that were made available for this research. However, the bureau has many more databases containing (administrative) information about other important risk factors, that could be used to improve the performance of predictive models. In comparison with other studies applying a predictive risk modeling approach, the number of predictors included in the analysis is still relatively low. For example: Marshall and English (2000); Vaithianathan et al. (2012) included, respectively, 37 and 132 variables, as predictors in the model. Thus, successive research on this topic should focus on model extension by gathering more information relating to other risk factors for juvenile services usage, for example: metropolitanism or birth weight.

Besides, the data quality of the predictors based on medical prescription data could be improved. The WHO ATC classification codes categorize medicine types into general categories. Not all the medicines in these general categories would directly imply the presence of a risk factor for which it is used. For example, a nicotine addiction would not directly increase the risk of child and parenting problems, but yet the prescription of nicotine patches is included in the category “anti-addictive medicin”. Medical data, derived from, for instance, electronical patient files, would provide more accurate and reliable information.

Moreover, at the moment, the final predictive model is able to identify youth help users for the following year. However, for prevention purposes, it would be even more interesting to be able to classify future youth help users a few years in advance. If a person is very likely to be entering a youth help trajectory in the following year, prevention measures might already be too late. The available data at the moment, unfortunately, does not allow for long-term predictions, because only juvenile services data for 2015 until 2017 is available. In the future, a model predicting youth help usage a few years in advance, will improve the managerial relevance of the model.

References

Aalbers-van Leeuwen, M., Van Hees, L. and Hermanns, J. (2002). Risico-en protectieve factoren in moderne gezinnen: reden tot optimisme of reden tot pessimisme?, *Pedagogiek* **22**(1): 41–54.

Anderson, K. E., Lytton, H. and Romney, D. M. (1986). Mothers' interactions with normal and conduct-disordered boys: Who affects whom?, *Developmental Psychology* **22**(5).

Appleyard, K., Egeland, B., Dulmen, M. H. and Alan Sroufe, L. (2005). When more is not better: The role of cumulative risk in child behavior outcomes, *Journal of Child Psychology and Psychiatry* **46**(3): 235–245.

Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures, *Computational Statistics & Data Analysis* **52**(4): 2249–2260.

Bakker, I., Bakker, K., van Dijke, A. and Terpstra, L. (2000). *O + O = O2: naar een samenhangend beleid en aanbod van opvoedingsondersteuning en ontwikkelingsstimulering voor kinderen en ouders in risicosituaties*, Nederlands Instituut voor Zorg en Welzijn / NIZW.

Bakker, K. (1999). Sociale kwetsbaarheid en sociale competentie: een kaderstelling, *K. Bakker, M. Pannebakker & J. Snijders (Red.). Kwetsbaar en Competent. Sociale Participatie van Kwetsbare Jeugd. Theorie, Beleid en Praktijk. Utrecht: NIZW* .

Baumrind, D. (1971). Current patterns of parental authority., *Developmental Psychology* **4**(1p2).

Bayer, J. K. and Sanson, A. V. (2003). Preventing the development of emotional mental health problems from early childhood: recent advances in the field, *International Journal of Mental Health Promotion* **5**(3): 4–16.

Bell, R. Q. and Chapman, M. (1986). Child effects in studies using experimental or brief longitudinal approaches to socialization., *Developmental Psychology* **22**(5).

Benton, A., Mitchell, M. and Hovy, D. (2017). Multi-task learning for mental health using social media text, *arXiv preprint arXiv:1712.03538* .

Boksic, S. and Verweij, A. (2011). *Leefbaarheid in balans*, Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

Bot, S., de Roos, S., Sadiraj, K., Keuzenkamp, S., van den Broek, A. and Kleijnen, E. (2013). *Terecht in de jeugdzorg: voorspellers van kind- en opvoedproblematiek en jeugdzorggebruik*, Sociaal Cultureel Planbureau.

Bouma, E. M., Ormel, J., Verhulst, F. C. and Oldehinkel, A. J. (2008). Stressful life events and depressive problems in early adolescent boys and girls: the influence of parental depression, temperament and family environment, *Journal of Affective Disorders* **105**(1): 185–193.

Brennan, T., Dieterich, W. and Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system, *Criminal Justice and Behavior* **36**(1): 21–40.

Buchdahl, G. (1956). Inductive process and inductive inference, *Australasian Journal of Philosophy* **34**(3): 164–181.

Bucx, F. and de Roos, S. (2011). *Opvoeden in Nederland*, Sociaal en Cultureel Planbureau.

Card, D., Chetty, R., Feldstein, M. and Saez, E. (2010). Expanding access to administrative data for research in the united states. american economic association, ten years and beyond: Economists answer nsfâĂŹs call for long-term research agendas.

CBS (2016). *Jeugdhulp 2015*, Centraal Bureau voor de Statistiek.

CBS (2017a). *Jeugdbescherming en jeugdreclassering 2016*, Centraal Bureau voor de Statistiek.

CBS (2017b). *Jeugdhulp 2016*, Centraal Bureau voor de Statistiek.

Chasnoff, I. J. (1988). Drug use in pregnancy: Parameters of risk, *Pediatric Clinics of North America* **35**(6): 1403–1412.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* **16**: 321–357.

Connelly, R., Playford, C. J., Gayle, V. and Dibben, C. (2016). The role of administrative data in the big data revolution in social science research, *Social Science Research* **59**: 1–12.

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R. and Putnam-Hornstein, E. (2017). Risk assessment and decision making in child protective services: Predictive risk modeling in context, *Children and Youth Services Review* **79**: 291–298.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment., *Journal of Family Psychology* **19**(2): 294.

de Roos, S., Bucx, F. and Geijer, M. (2011). *Gezondheid en probleemgedrag van kinderen: de rol van ouders en de bredere opvoedomgeving*, Sociaal en Cultureel planbureau.

Euhus, D. (2004). Risk modeling in breast cancer, *The Breast Journal* **10**(s1).

Fogel, D. B., Wasson III, E. C., Boughton, E. M. and Porto, V. W. (1998). Evolving artificial neural networks for screening features from mammograms, *Artificial Intelligence in Medicine* **14**(3): 317–326.

Goerge, R. M. and Lee, B. J. (2002). Matching and cleaning administrative data, *New Zealand Economic Papers* **36**(1): 63–64.

Guazzelli, A. (2012). Predicting the future, part 1: What is predictive analytics, *IBM: Developer Works* .

Harden, K. P., Lynch, S. K., Turkheimer, E., Emery, R. E., D'onofrio, B. M., Slutske, W. S., Waldron, M. D., Heath, A. C., Statham, D. J. and Martin, N. G. (2007). A behavior genetic investigation of adolescent motherhood and offspring mental health problems., *Journal of Abnormal Psychology* **116**(4): 667.

Hermanns, J., Öry, F. and Schrijvers, G. (2005). Helpen bij opgroeien en opvoeden: eerder, sneller en beter.

Hurley, D. (2018). Can an algorithm tell when kids are in danger? **URL:** *https://www.nytimes.com/2018/01/02/magazine/cananalgorithmtellwhen-kidsareindanger.html?_r=0*

Jansen, W., Mieloo, C., Anschutz, J. et al. (2015). Gap between the use of and need for youth care: research in rotterdam neighbourhoods, *Nederlands Tijdschrift Voor Geneeskunde* **159**: A7664–A7664.

Kanazawa, K., Kubo, M. and Niki, N. (1996). Computer aided diagnosis system for lung cancer based on helical ct images, *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, Vol. 3, IEEE, pp. 381–385.

Kijlstra, M., Prinsen, B. and Schulpen, T. (2001). *Kwetsbaar jong! Een quick scan van de kansen op achterstand van kinderen van 0 tot 4 jaar in risicosituaties*, NIZW.

Lamborn, S. D., Mounts, N. S., Steinberg, L. and Dornbusch, S. M. (1991). Patterns of competence and adjustment among adolescents from authoritative, authoritarian, indulgent, and neglectful families, *Child Development* **62**(5): 1049–1065.

Levine, J. A., Emery, C. R. and Pollack, H. (2007). The well-being of children born to teen mothers, *Journal of Marriage and Family* **69**(1): 105–122.

Lewis, G. H., Georghiou, T., Steventon, A., Vaithianathan, R., Chitnis, X., Billings, J., Blunt, I., Wright, L., Roberts, A. and Bardsley, M. (2013). Impact of âĂŸvirtual wardsâĂŹ on hospital use: a research study using propensity matched controls and a cost analysis, *Final report. NIHR Service Delivery and Organisation Programme* .

Linoff, G. S. and Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*, John Wiley & Sons.

Marshall, D. B. and English, D. J. (2000). Neural network modeling of risk assessment in child protective services., *Psychological Methods* **5**(1): 102.

Mieloo, C., Anschutz, J., Rietveld, L. and van den Einde-bus, A. (2013). *Vraagontwikkelingsonderzoek: Risicofactoren voor zorggebruik*, Gemeente Rotterdam.

Nagin, D. S. and Tremblay, R. E. (2001). Parental and early childhood predictors of persistent physical aggression in boys from kindergarten to high school, *Archives of General Psychiatry* **58**(4): 389–394.

Ngai, E., Hu, Y., Wong, Y., Chen, Y. and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *Decision Support Systems* **50**(3): 559–569.

Nyce, C. (2007). Predictive analytics white paper, *American Institute for CPCU. Insurance Institute of America* pp. 9–10.

Preoţiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A. and Ungar, L. (2015). The role of personality, age, and gender in tweeting about mental illness, *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 21–30.

Provost, F. and Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*, " O'Reilly Media, Inc.".

Reece, A. G., Reagan, A. J., Lix, K. L., Dodds, P. S., Danforth, C. M. and Langer, E. J. (2017). Forecasting the onset and course of mental illness with twitter data, *Scientific Reports* **7**(1): 13006.

Repetti, R. L., Taylor, S. E. and Seeman, T. E. (2002). Risky families: family social environments and the mental and physical health of offspring., *Psychological Bulletin* **128**(2): 330.

Rijksoverheid (2017). Rijksbegroting 2017.
URL:
http://www.rijksbegroting.nl/2017/voorbereiding/begroting,kst225923_13.html

Roes, T. (2005). *De sociale staat van Nederland 2005*, SCP.

Sadiraj, K., Ras, M. and Pommer, E. (2016). *Cumulaties in de jeugdhulp*.

Schwartz, I. M., York, P., Nowakowski-Sims, E. and Ramos-Hernandez, A. (2017). Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The broward county experience, *Children and Youth Services Review* **81**: 309–320.

Shroff, R. (2017). Predictive analytics for city agencies: Lessons from children's services, *Big Data* **5**(3): 189–196.

Sieh, D. S., Visser-Meily, J. M. A. and Meijer, A. M. (2013). Differential outcomes of adolescents with chronically ill and healthy parents, *Journal of Child and Family Studies* **22**(2): 209–218.

Soede, A., Hoff, S. and Vrooman, C. (2011). *Armoede volgens de budgetbenadering*, Centraal Bureau voor de Statistiek.

Stanley, S. and Vanitha, C. (2008). Psychosocial correlates in adolescent children of alcoholics-implications for intervention, *Int J Psychosoc Rehabil* **12**: 67–80.

Stevens, G. W., Pels, T., Bengi-Arslan, L., Verhulst, F. C., Vollebergh, W. A. and Crijnen, A. A. (2003). Parent, teacher and self-reported problem behavior in the netherlands, *Social Psychiatry and Psychiatric Epidemiology* **38**(10): 576–585.

Stevens, J., Pommer, E., van Kempen, H., Zeijl, E., Woittiez, I., Sadiraj, K., Gilsing, R. and Keuzenkamp, S. (2009). *De jeugd een zorg*, Sociaal en Cultureel Planbureau.

Stevens, L. A. and Laurie, G. (2014). The administrative data research centre scotland: A scoping report on the legal & ethical issues arising from access & linkage of administrative data.

Vaithianathan, R., Maloney, T., Jiang, N., De Haan, I., Dale, C., Putnam-Hornstein, E., Dare, T. and Thompson, D. (2012). Vulnerable children: Can administrative data be used to identify children at risk of adverse outcomes, *Report Prepared for the Ministry of Social Development. Auckland: Centre for Applied Research in Economics (CARE), Department of Economics, University of Auckland* .

van Dalen, J. (2018). Session 6 - more analytics, data science in organizations.

van den Broek, A., Kleijnen, E. and Keuzenkamp, S. (2010). *Naar Hollands gebruik? Verschillen in gebruik van hulp bij opvoeding, onderwijs en gezondheid tussen autochtonen en migranten. Verdiepingsstudie Diversiteit in het Jeugdbeleid.*, Sociaal en Cultureel Planbureau.

van der Zwaluw, C. (2011). *Genes in a bottle: The interplay between the social environment, individual characteristics, and genetics in alcohol use*, PhD thesis, [Sl: sn].

Weijters, G., Scheepers, P. and Gerris, J. (2007). Distinguishing the city, neighbourhood and individual level in the explanation of youth delinquency: A multilevel approach, *European Journal of Criminology* **4**(1): 87–108.

Weir, S. and Jones, W. (2009). Selection of medicaid beneficiaries for chronic care management programs: Overview and uses of predictive modeling, *Center for Health Policy and Research* .

Woldringh, C. and Peters, J. (1995). *De relatie tussen risico- en protectieve factoren en het functioneren van het kind*, Instituut voor Toegepaste Sociale Wetenschappen.