Linköping University | Department of Management and Engineering
Master's thesis | Computational Social Science
Spring term 2020 | ISRN LIU-IEI-FIL-A--20/03344--SE

Classification Tree Based Algorithms in Studying Predictors for Long-Term Unemployment in Early Adulthood

An Exploratory Analysis Combining Supervised Machine Learning and Administrative Register Data

Sanni Kuikka

Advisors: Maria Brandén & Martin Arvidsson
Examiner: Eduardo Tapia

LINKÖPINGS UNIVERSITET

Table of Contents

List of Tables, Figures and Appendices

Tables

Figures

Appendices

Abstract

Unemployment at young age is a negative life event that has been found to have scarring effects for future life outcomes, especially when continuing long-term. Understanding precursors for long-term unemployment in early adulthood is important to be able to target policy interventions in critical junctures in the life course. Paths to unemployment are complex and a comprehensive outlook on the most important factors and mechanisms is difficult to obtain. This study proposes a data-driven, exploratory approach for studying individual and family level factors during ages 0-24, that predict long-term unemployment at the age of 25-30.

A supervised machine learning approach was applied to understand associations deriving from longitudinal, individual-level administrative data from a full birth cohort in Finland. The data comprise information about physical and social wellbeing, life course events, as well as demographics, including the parents of the cohort members. Potential predictors were chosen from the data based on theories and previous research, and used to train a model aiming to correctly classify unemployed individuals. A CART algorithm was used to build a classification tree that reveals important variables, ranges of them as well as combinations of factors that together are predictive of long-term unemployment. A random forest algorithm was used to build several trees producing smoothed predictions that reduce overfitting of one tree. CARTs and random forest models were compared to each other to understand how they perform in a research task predicting life outcomes.

Both individual and family level factors were found to be predictive of the outcome. Combinations of variables such as GPA lower than ~7.5, ego's low education level, late work history start, depressive disorders and low parental education and income levels were found to be particularly predictive of unemployment. CART models correctly classified up to 87% of the unemployed, while misclassifying 70% of the employed and having 45% overall accuracy. Testing for CART model stability, finding consistency across several tree models improved robustness. Random forest correctly predicted up to 59% of the unemployed, while also correctly classifying 65% of the employed and producing robust results. The two algorithms together provided valuable insight for better understanding factors contributing to unemployment. The study shows promise for classification tree based methods in studying life course and life outcomes.

Keywords: long-term unemployment, life course research, register data, cohort study, supervised machine learning, CART, random forest

Acknowledgements

I would first like to express my gratitude to my main advisor and mentor Maria Brandén. Her positive encouragement has been a source of confidence and inspiration throughout the thesis process as well as the whole master's program. She believed in me and this project and helped me see the bigger picture. I would also like to thank my second advisor, Martin Arvidsson, who helped me navigate the mostly uncharted waters at the intersection of machine learning and life course research.

I would like to thank the Finnish Institute for Health and Welfare for providing me the opportunity to work with this unique data. In addition to thesis collaboration, spending the summer 2019 as a research intern at their Oulu office offered me valuable professional experience combining my interdisciplinary knowledge and skillset.

I want to thank my friend Turo for his invaluable help and being my extra set of eyes during the thesis project. He proofread my text and helped debug code in moments of frustration. I also want to thank my friend Silja who throughout this thesis project has been a great emotional support, and reminding me of how much I've worked and how far I've come. I thank them both for being my mirrors when I dove deep into my work, and for cheering me on and believing in me. Through you I see myself.

Thank you goes to my friends and family as well for their support and for encouraging me in my pursuit of this master's degree and my international adventures. Having people who care for me and who I can always return to means the world to me.

Finally, I thank my classmates for pioneering this master's program with me. The journey has not been easy, and having peer support has provided plenty of comfort. Moreover, learning from individuals with diverse cultural and disciplinary background has been a source of inspiration and personal growth.

Norrköping, June 2020

Sanni Maria Kuikka

1. Introduction

Unemployment is a negative life course event and when occurring and sustaining in early adulthood it can have scarring effects for future life outcomes (e.g. Abebe & Hyggen, 2019; Helbling et al., 2019; Mousteri et al., 2018). Childhood and adolescence individual factors as well as family conditions and processes of intergenerational transmission of disadvantage can together suggest to later unemployment (Caspi et al., 1998; Doku et al., 2018; Lallukka et al., 2019). The reasons behind and processes leading to unemployment are, however, complex by nature, and pinpointing the most important mechanisms and finding the most important factors can be challenging. Traditional quantitative social research approaches apply deductive reasoning aiming to generalize from particular by formulating hypotheses based on existing theories and designing a study to test the hypotheses. When studying the social world and life course, alternative approaches can be helpful to better understand the logic of such complexity and complement the research done so far.

A promising approach can be found in the machine learning literature. Supervised machine learning algorithms use a data-driven approach to analyze and derive meaning from big data (Burkov, 2019). They can provide new insight to understanding complex processes by approaching a research problem in a more inductive way: data are used to train an algorithm to understand patterns that derive from them in order to make conclusions. No functional form is specified in advance: parameters for the algorithm can be adjusted to find a more optimal fit, but data have the leading role. Such an approach can be used to complement deductive research methods traditionally applied in quantitative social research to better understand the complexity of the social word. Machine learning has been little utilized in the social sciences so far, but some very promising initiative has been made (see e.g. Boelaert & Ollion, 2018; Daoud & Johansson, 2019; Salganik et al., 2020). Due to the novelty of the field, more research is called for to better understand the insights such methods can offer.

In this thesis I study predictors for long-term unemployment in early adulthood, applying a data-driven approach using classification tree based algorithms, more specifically CART and random forest. Classification tree based algorithms are a type of supervised machine learning which apply a logic of recursive partitioning, where the algorithms identify important variables, ranges of those variables and higher-order interactions of them that together are predictive of the outcome (here: long-term unemployment). The patterns deriving from algorithms trained with a subset of data suitable for answering the research question are validated with another subset, by making predictions of the outcome and then comparing to the observed values. For the analysis, I'm using governmental administrative data from Finland: a full birth cohort of 1987 have been followed up to 30 years of age and individual-level data about life events and most aspects of physical, mental and social wellbeing as well as demographics, gathered longitudinally (Paananen & Gissler, 2012).

Individual and family level variables are chosen based on theories and previous research to capture important factor combinations and predict long-term unemployment between ages 25-30.

Two research questions are proposed, one regarding the topic and one the methodological approach:

1. Which variables of individual and family level characteristics are together predictive of long-term unemployment in early adulthood?

2. How do two different classification tree based methods, CART and random forest, compare to each other in predicting long-term unemployment in early adulthood?

2. Literature Review

2.1. Employment in Transition to Adulthood

Finding employment is a critical life course transition in a young person's life and one of the important markers for transitioning to adulthood (Hogan & Astone, 1986; Shanahan, 2000). Employment serves several different functions, varying from economic and material to psychosocial (Hess et al., 1994; Jahoda, 1981). For instance, having income provides economic independence from caregivers and/or the state, and allows one freedom to function in a society structured around market economy. In terms of psychosocial processes, having employment is connected to finding a meaningful role in the society and developing one's identity in relation to an occupation.

As employment plays such an important role in the transition process, it is intuitive that if the transition is unsuccessful, it can have negative consequences for the individual and through that, for the society as a whole. These lasting negative effects from being unemployed on life outcomes, often referred to as scarring effects, have been documented time and again. Abebe & Hyggen (2019) find scarring effects on wage and future employment from early unemployment. They also find heterogeneity in treatment effects moderated by several individual and family characteristics, such as gender, own and parental education levels, mental distress and problematic substance use. Clark & Lepinteur (2019) find early unemployment to have negative effects on life satisfaction at the age of 30. Their results are controlled for and independent of individual childhood and adulthood variables about intellectual performance, behavior, emotional health and current unemployment as well as family background variables such as parental education, family income level and family break-up. In a comparative cross-national study Mousteri, Daly, & Delaney (2018) find unemployment scarring effects on life satisfaction across all 14 European countries included in the study, both contemporarily and long-term. The results could not be explained by individual demographic, socio-economic or health related controls nor by country level controls, suggesting that the processes of scarring effects persistently operate beyond cultural definitions. Ending up unemployed can thus lead to disadvantage in the future and as such, it is a process that warrants more attention.

Through historical time, the transition to adulthood has changed from a series of subsequent fixed events, one following after another, to a process where events can even be reversible, such as finding employment and later losing it, and where they can no longer be conceptualized as discrete (Shanahan, 2000). These developments have placed new challenges for researching transitions to adulthood and has called for a reconceptualizing of the expected timing and order of life course events. In terms of employment - one of the non-discrete markers for adulthood - factors preceding the transition can be many and heterogeneity in the processes can be expected.

2.2. Paths to Unemployment

Several factors together play a role in how young individuals find their way into employment and integrate into the labor market. Economic conditions in the society inevitably influence the risk for an individual to become unemployed, as well as many structural and institutional factors, such as educational composition of the country, labor market structure, job availability, educational system and employment policies (Täht & Reiska, 2016). There are also differences in individuals, their characteristics, skills, opportunities and premises, that affect whether a person ends up unemployed (Caspi et al., 1998). Combinations of such factors can be assumed to interact and increase the risk for becoming, or remaining, unemployed.

Young people in Europe are increasingly experiencing episodes of unemployment in the work life (Müller & Gangl, 2003). When overall unemployment rates are high, it also translates to young workers: youth often have significantly higher unemployment rates compared to the overall (Eurostat, 2020). The jobs available are insecure and often short-term, and young people are in a vulnerable position because of having accumulated less labor market experience and networks than senior workers. Individuals often linger in a stage of prolonged transition, sometimes referred to as *moratorium* (Erikson, 1968), *holding tank* (Coleman, 1994) or *emerging adulthood* (Arnett, 2000), which can be characterized in terms of multiple transitions in and out of the workforce before making strong attachments to the labor market. Often, these kinds of insecure jobs with multiple transitions are termed *precarious employment*. Precarious employment arrangements, whether occurring simultaneously with studying or not (Wolbers, 2003), can serve as a stepping stone into the labor market but do not necessarily help finding better employment arrangements later on (de Graaf-Zijl et al., 2011). Areas of residence and work commute vary in their ratio between job availability and number of working age people as well as unemployment rate, thus contributing to the risk for becoming unemployed (Lallukka et al., 2019). The setup of the educational system primes pathways to the labor market as well as assigns a ruleset to accessing different types of education (Levels et al., 2014). When such a setup is unequal, some can be negatively affected in terms of future employment. Policies enforced on the unemployed in terms of activity measures or unemployment benefits play their part as well. In the case of universalistic welfare state policies, while policies offer income security in the face of job loss or re-training opportunities with prolonged unemployment, they may also contribute to increased amounts of precarious jobs in terms of e.g. fixed-term contracts to increase flexibility in the labor market (Täht & Reiska, 2016). Because such complex macro systems are contributing to paths to unemployment, it can be difficult to rank which processes are more important than others.

Individual differences contributing to the risk of becoming unemployed at a young age can be viewed from multiple theoretical perspectives. Caspi et al. (1998) approach the problem in an interdisciplinary way, combining explanations from acquisition of different types of capital: human (economics), social (sociology) and personal (psychology). The skills, qualifications, knowledge and resources needed for becoming employed can be categorized as human capital (Becker, 1975). These can be conceptualized as e.g. important basic skills such as literacy and numeracy or educational attainment and qualifications. The more human capital an individual has, the more employability as well. Thus, individuals worse equipped with human capital have a higher risk to become unemployed. Social structures, often networks of familial, friendly or collegiate nature, that provide or control access to resources needed for acquiring and maintaining a job are referred to as social capital, and can be crucial in processes of being left without a job (Coleman, 1988). Not having access to networks that can help in finding a first job or moving from a precarious job arrangement to a more stable one is an increased risk for unemployment, to which young people are naturally more vulnerable to. Personal capital, as referred to by Caspi et al. (1998), consists of behavioral and psychological characteristics and resources linked to motivation and capacity to work. Motivation to work has several explanations in the scientific field (Jahoda, 1981) but is nevertheless one key factor in who finds employment and maintains it and who does not. Factors of personal capital can be conceptualized, for example, in terms of symptoms of mental illness, antisocial behavior and substance abuse. All the above-mentioned individual perspectives can work simultaneously and have combined effects in paths to unemployment, while also interacting with the macro systems described earlier.

Not only are individual level characteristics likely to be important but also characteristics of one's parents. Intergenerational transmission of (dis)advantage refers to family surroundings one grows up in as well as parental resources having an impact on the individual, and therefore their life course and life outcomes. Intergenerational transmission of socioeconomic (dis)advantage is one of the key topics in sociology and has been studied for over a century with different approaches (Ganzeboom et al., 1991). The processes of accumulating different types of capital can all have roots in the family: a vast body of literature has found positive associations between the human, social and personal capital of parents and children (see e.g. Black & Devereux, 2010; Black et al., 2005; Caspi et al., 1998; Currie, 2009; Doku et al., 2018; Mood, 2017). Parental economic resources can translate to investments into the child and their future employability, in other words, to the accumulation of the child's human capital. Social capital of the children that originates from the family can be expressed both in terms of parents having access to useful networks that can benefit the child, as well as parents being physically and emotionally available to the child. This also translates to the interplay between social and human capital: if the family has strong social capital – close relationships – the human capital of the parents can transmit to the children (Coleman, 1988). Lack of parental human and/or social capital can thus translate to less chances of finding or maintaining employment as well as poor employability of the child. Social capital in the family can also be understood in terms of structurally organized home and family processes of informal social control, where children are negatively affected if the home environment does not provide enough safety and stability in terms of parental resources to take care of the child. A disorganized and unstable home environment has been associated with antisocial and criminal behavior (Sampson & Laub, 1994), which again has been connected to unemployment and precarious work arrangements in later life (Sanford et al., 1994). Such family conditions can be a fertile ground also for mental health and substance abuse problems, that can operate either through diminished opportunities to accumulate human and social capital, or psychological distress that affects abilities to apply for and maintain a job (Wadsworth et al., 2005). Moreover, motivation to work and values towards it can be transmitted intergenerationally. Finally, it has also been found that unemployment spells itself are reproduced intergenerationally and that scarring effects on life satisfaction are less severe for young people with parental unemployment: a social-norm effect of the intergenerationally disadvantaged individuals in terms of family unemployment can be observed (Clark & Lepinteur, 2019).

2.3. Determinants for Unemployment

Plenty of literature has studied youth unemployment, typically addressing individuals aged up to 25. This study addresses a stage in the life referred to as early adulthood (25-30y): it is defined in terms of a period after the phase of emerging adulthood, when more stable work force attachment is assumed to form. Due to shortage of literature in the specific age range, a review is provided of predictors for unemployment at young age in general. The duration of unemployment has been found important for scarring effects which is the reason to focus on prolonged status of unemployment: the longer the unemployment the more severe the scars (Clark et al., 2001). The typical pattern of precarious work arrangements during youth can also lead to more short unemployment spells in between temporary work arrangements, pronouncing the need to make a distinction between a likely transient situation and a more continuous state.

The Finnish Institute for Health and Welfare recently conducted a study about social and health related precursors for long-term unemployment in early adulthood (25-28y) using the same data that I used in this study (Lallukka et al., 2019). Their research design is different to this study in terms of theoretical framework, outcome variable operationalization and methods, while at the same time serving as an interesting benchmark for this study. They defined long-term unemployment as lasting over 12 consecutive months between 25-28 years (up to the age where data were available at the time of analysis). They studied several social and health determinants for both ego and parents. Grade point average was found to be strongly associated to early adulthood unemployment, as well as ego’s long-term unemployment before the age of 25. Mental health of ego and parents were also among the most important predictors, as well as parental education level, young age of the mother and involvement with child protective services, specifically after the age of 12. They also found that residing in a region with unemployment rates higher than the national average increases the risk. One of their key findings was that accumulation of measures of disadvantage increased the probability for unemployment.

Other studies have also found that individual and intergenerational characteristics begin shaping unemployment trajectories well before the transition to labor market happens. Doku et al. (2018) studied predictors for youth unemployment trajectories between ages 16-28, also in the context of Finland. The study revealed both individual and intergenerational predictors, the latter measured in two generations back. Ego’s poor school achievement and low education level were strong predictors for high unemployment risk. Low education level and socioeconomic status of not only the parents but also the grandparents were found to predict ego’s high risk for unemployment. Parents and grandparents living in rental dwellings as opposed to owner-occupied predicted high risk for unemployment for ego. Caspi et al., (1998) studied long-term unemployment between ages 15-21 in New Zealand using predictors measured at different stages during childhood. Most important predictors measured at the age 15 were not having obtained a school certificate, having low reading skills and showing symptoms of delinquency. They found that accumulation of disadvantage factors, such as parent’s low occupational status, single-parent family and ego’s behavioral problems, starting from early childhood and continuing to adolescence, were strong predictors for unemployment. They also tested whether the predictors were the same for lack of education, potentially explaining the predictors for unemployment: they found that factors such as having not obtained a certificate, low reading skills and delinquency continued to significantly contribute to unemployment having adjusted for lack of education.

2.4. Machine Learning for Understanding Life Course and Life Outcomes

Understanding life paths leading to unemployment is of importance for gaining more insight about the social world but also for policy-making purposes: possessing knowledge about which factors together predict the outcome makes it possible to provide support and implement policy interventions to correct the processes before they lead to unwanted outcomes. The previous chapters shed some light over the complexity of the mechanisms leading to early adulthood unemployment. As many simultaneous processes are at play, it is difficult to analytically pinpoint the most important ones. Most of the previous quantitative research leans on deductive approach, where hypotheses are built based on theory and then statistically tested to assess the generalizability of the theory. In this study I use instead a more inductive approach applying supervised machine learning. Using this approach can help reveal new insight to the phenomenon under inspection, but it can also be used to validate previous research in terms of the social problem of unemployment.

Machine learning (ML), sometimes referred to as artificial intelligence, is an algorithmic way of statistical modelling (Breiman, 2001). An algorithm is *"a set of mathematical instructions or rules that, especially if given to a computer, will help to calculate an answer to a problem"* (Cambridge Dictionary, 2020). ML typically has an exploratory approach drawing close to inductive reasoning: starting with observations, deriving a statistical model from the observations using an algorithm, and ending up with associations from the data. It is a highly data-driven approach to research, and quality of data plays an increasingly important role: a trained algorithm is merely a reflection of the data used for training. Different types of learning processes, supervised, unsupervised and semi-supervised exist for different purposes (Burkov, 2019). For life outcome research, supervised learning (SML) is intuitive: it flexibly searches for functions $f(X)$ to predict an output (Y) given an input (X), and forecast (Y) for future inputs (X) (Molina & Garip, 2019). It aims to predict outcomes in a dataset based on an algorithm that is trained with another dataset: a process called cross-validation. In the social sciences, specifically in sociology, machine learning is a relatively new approach and there are not yet established standards for research projects using ML.

In the field of life course and life outcomes research, there are so far few examples using machine learning as a research method. Hobcraft & Sigle-Rushton (2009) studied resilience factors among individuals who experienced time in the care system during their childhood. They used SML to predict having no academic qualifications at age 30 (on a sample of individuals with a premise of heightened risk for obtaining no academic degree), using several survey-based childhood precursors related to family socioeconomic background, behavioral & psychological characteristics and academic test scores. They used a decision tree algorithm for the classification task and were able to distinguish ranges of certain variables that either suggested a risk for no academic degree or, conversely, identified patterns of resilience. Some of their findings include a risk combination of low test scores and aggression, strongly suggesting to not having an academic degree, and a resilience combination of test scores not being in the lowest quartile, not having lived in social housing and not having high scores in aggression, suggesting to having a degree. They also constructed a logistic regression model and while getting similar predictions from both models, they concluded that SML approach distinguished combinations of risk and resilience factors that would have been difficult to distinguish using regression only. Another example is the Fragile Families challenge, where a scientific mass collaboration was conducted to predict six life outcomes using large-scale birth cohort survey data (Salganik et al., 2020). Six waves of survey data specifically designed for the purpose to understand children born into families with unmarried parents were collected from 0 to 15 years of age by interviewing parents, teachers and the child (once old enough). Several research teams undertook the same task to predict child grade point average, child grit, household eviction, household material hardship, primary caregiver layoff and primary caregiver participation in job training, as accurately as possible, using the variables in the survey data as predictors, and using any method available. Most teams used SML methods similar to and deriving from the ones used by Hobcraft & Sigle-Rushton (2009). Each of the 160 submissions were assessed using the same criteria. What this common task method study revealed was that the predictions produced were not very accurate, with R^2 varying between 0.05-0.2 depending on the outcome. The best performing complex SML models were only slightly better at predicting than fairly simple linear or logistic regression models. Their conclusions include the notion that even though poor predictions were obtained with these data, predictability of life outcomes can vary between settings and types of data: they suggest using a similar approach with administrative governmental data.

There are clear differences between these two example cases that should be pointed out for the purpose of future research. Hobcraft & Sigle-Rushton (2009) are experienced social scientists specialized in quantitative analysis to understand social exclusion in the society. The predictors were chosen by the researchers based on their expertise in structures of disadvantage and their study had a small sample size (N = 440) of a particular nature. Their success was facilitated by substance expertise combined with a novel methodological approach, small data from a particular case, and not aiming for accuracy of predictions (R^2 0.11 for logistic regression and 0.12 for SML) but gaining more insight from the higher-order interactions of predictors revealed by the tree models. The best performing models in Fragile Families challenge were constructed by data scientists who have no expertise in sociology (Rigobon et al., 2018). They had 2121 households in their data with more than 20000 predictors that were all used to predict several outcomes. What we can learn from these two examples can be summarized in three main points: 1) more units in the data can provide more generalizability, but simultaneously introduce more complexity due to varying life courses, 2) the amount and quality of predictors are important, a few handfuls of carefully chosen predictors based on topic-specific theory and research can work better than thousands of precursors used to predict several outcomes and likely producing noise for many models, and 3) it is worthwhile not to focus solely on prediction accuracy in understanding what data-driven methods can teach us about diverging life courses. Building on this, I choose to study one outcome - long-term unemployment in early adulthood - with similar methods used in both of the examples, utilizing theory and previous research as well as my training in the field of sociology and social welfare to choose a set of predictor variables, and use governmental administrative data, though large in number of units, as suggested by Salganik et al. (2020) as means to achieve better predictions in these kinds of endeavors.

Due to the exploratory and data-driven nature of ML, traditional ways for establishing causality do not apply in such research projects. However, there have been recent developments towards the direction of studying causality with ML methods (Athey & Wager, 2019). While establishing causality is outside the scope of the current study, associations between factors, that have been previously found predictive of the outcome using traditional approaches testing for hypothesis, can suggest towards causal patterns. Questions addressing selection versus causation in life course research have been asked repeatedly, and previous research has paid attention to this divide (e.g. Caspi et al., 1998). However, when drawing conclusions based on results drawn from an exploratory study using ML methods that do not directly address causality, one needs to be careful when making such interpretations. While understanding the restrictions of supervised machine learning approaches in general in terms of interpretability of "black box" algorithms, their empirical efficacy in terms of flexible models, predictive power, cross-validation and balancing between over-fitting and biasedness are useful features when looking into a complex social world problem using such detailed big data (Boelaert & Ollion, 2018).

3. Data and Methods

I conducted the analysis using governmental administrative data from a Finnish Birth Cohort, where all 60254 children entered in the Medical Birth Record in 1987 are followed throughout their lives from fetal stage onwards (Paananen & Gissler, 2012). Currently the entries reach up to 2017, with some variation across different registers. The data include individual-level and longitudinal administrative information from 10 different population register holders, covering most aspects of wellbeing, life events as well as demographics, and linking them together with unique personal identification numbers. In addition to the children born in 1987, their parents are linked to the data and administrative register information collected from them as well, partly already from the time before the birth of the child. Such rich data enable distinguishing factors in the life course, both individual and intergenerational, for predicting life outcomes. For the purpose of studying long-term unemployment as an outcome, individuals who died during the review period were excluded from the population, as well as individuals with an intellectual disability, following the definition by Westerinen (2018), resulting in an analytical sample the size of 58179. The data were provided by the Finnish Institute for Health and Welfare, who collect them from administrative register holders as well as coordinate research done using the data.

Administrative data do not often capture personality characteristics, such as attitudes towards work, which is a disadvantage of this type of data (Dohmen & Van Landeghem, 2019). The data available in this study do not explicitly address macro conditions either, but using a birth cohort from a single year, each individual can be assumed to have been exposed to same societal conditions and macro events, such as the recessions in the beginning of the 1990's and around 2008 in this case. Nevertheless, such economic conditions have likely contributed to more polarization: as a result, individuals who are in disadvantaged positions to begin with can become even more so (Rinne & Järvinen, 2010). This is difficult to assess with data from one birth cohort only.

Using such data containing individual-level data of sensitive nature, such as medical diagnoses and criminal records, calls for some ethical remarks. Register data have not originally been collected for research, but administrative and demographic purposes. It is nevertheless increasingly used as one type of big data for research outside its original intended purpose. Informed consent of the subjects, ownership of data and datafication of information related to big data have been under intense debate in the recent years due to increased volumes of data collection and repurposing for predictive analytics (Mai, 2016; Safran et al., 2007; Vayena & Blasimme, 2018). Historical data from registers regarding populations, patients or customers are drawn and algorithmically repurposed in ways that are not always transparent, potentially jeopardizing individual rights or even democratic processes (Hao, 2019; O'Neil, 2017). Individuals who are under inspection in studies such as the current one and whose personal information is turned into data, processed and analyzed, have not given their informed consent: being registered in population databases is seen as civic duty. The ethical responsibilities of database holders and researchers repurposing such data algorithmically are pronounced in this time in history. In the case of the Finnish Birth Cohort Study, data from population registers have been later collected and merged for research purposes by the license of register holders. The individual researcher using the data undergoes a process of personal security check and signing non-disclosure forms for every register holder separately. To ensure data privacy, the data are accessed with institute credentials through a two-phase identification protocol and a secure VPN connection, using a laptop owned and administered by the Finnish Institute for Health and Welfare.

3.1. Long-term Unemployment in Early Adulthood

I defined long-term unemployment in early adulthood as being registered as a job seeker for more than 365 consecutive days at any time during 1.1.2012 – 31.5.2017 (latest entry in the register at the time): 27.2% of the individuals in the analytical sample fulfill this condition. The cohort members were ~25-30 years of age at that time: an important age where the prolonged transition period - holding tank phase - can be assumed to have passed for most, when secondary and potentially also university education has finished, and when work life attachment and establishing a career are ongoing. The data were derived from the Ministry of Economic Affairs and Employment in Finland. Being registered as a job seeker in the Employment and Economic Development Office is a requirement for receiving unemployment benefits, the primary welfare subsidy for unemployed adults included in the work force, yet not a guarantee for receiving them. Individuals registered as job seekers for the purpose of having access to unemployment benefits can be found in varying situations, such as unemployed looking for work, part-time workers getting partial unemployment benefits to compensate the loss of income due to lack of fulltime employment, and previously unemployed currently fulltime degree students for whom studying further has been found to increase their employability and unemployment benefits (which are higher than student benefits) granted as an incentive to re-educate (TE-services, 2018). They can also be temporarily laid off, in training or in services promoting employment, and the type and amount of potential unemployment benefits can vary depending on one's label in the register. This definition of unemployment can be contrasted to the term *unemployed jobseeker*, which is used by governmental agencies to distinguish individuals who are fully unemployed and looking for fulltime work from all registered as unemployed. The Employment Bulletin for November 2015, the last archived bulletin so far during the review period, reveals that when looking at the whole Finnish adult population, approximately 54%

of the registered jobseekers (i.e. those who are defined as unemployed in this study) were, in fact, fully unemployed jobseekers (Official Statistics of Finland, 2015). Out of those, the distribution by age group is balanced among people aged 20-64, indicating that the percentage of unemployed jobseekers holds also in the outcome variable. Thus, the operationalization including all job seekers provides a nationally drawn definition about different types of unemployment and seeking a job, made for the purpose of the social welfare system, rather than capturing a pure measure of unemployment. For this reason, the operationalization does not yield an internationally comparable definition of unemployment either. Another drawback of the operationalization is not being able to separate the concept unemployed from NEET - not in Employment, Education, or Training (Holte et al., 2019). Nevertheless, it can be seen as to capture a real-life phenomenon related to not having a smooth transition into employment and representing the important life course transition failing in some way, thus serving as an interesting outcome variable.

3.2. Predictors

I chose a set of predictor variables based on theories and previous research as described in the literature review (in particular, see Lallukka et al. (2019)). Predictors were defined as occurring before the outcome variable review period starts and grouped to represent individual and family level characteristics. Distributions and descriptive statistics are displayed in Table 1a-c. Individual level predictors aim to capture ego’s characteristics, resources, skills and measures of wellbeing important for transitioning to working life and were chosen as follows:

- Sex
- GPA
- School completed on time
- Education level
- Timing of work history start
- Young motherhood
- Marriage status
- Behavioral diagnosis
- Mood diagnosis
- Anxiety diagnoses
- Substance abuse diagnosis
- Criminal conviction
- Number of criminal offences

Sex was coded by a dichotomous classification (female/male) according to legislative processes related to registering a child in the Medical Birth Record. Information on grade point average (GPA) as well as whether school was completed on time (i.e. together with the majority of the birth cohort) were drawn from the Finnish National Board of Education. GPA was available when applying for secondary education and finishing or having finished comprehensive school. For cases where GPA had missing values (1386 observations, 2.5%), a data imputation technique commonly used in machine learning of replacing NA with the average value (7.8) of the variable across the population was used (Burkov, 2019). This can potentially affect the results since the individuals with missing GPA can be assumed to either not have finished school on time or having low GPA and not applying for further studies. Nevertheless, I decided to use this strategy instead of excluding all cases with missing value from the analysis: Lallukka et al. (2019), using the same data, find GPA to be the most predictive variable and comparing to their results can serve as one type of sensitivity check for the imputation technique. Comprehensive school finished on time was coded as a dichotomous 0(no) / 1(yes) variable based on whether the person appeared in the register for applying for secondary education in 2003 along with the majority of the cohort. The highest education level obtained before 2012 was derived from Statistics Finland Register of Educational Achievements and coded with values 0 (no secondary) / 1 (secondary, either vocational or high school) / 2 (bachelor) / 3 (master or higher).

Timing of the first paid job was derived from the Finnish Center for Pensions and coded as the year of observation, with missing information coded as “2012”, representing later work history start or no start at all. Young motherhood was defined as giving birth before the age of 25, coded as 0 (male or female no birth) / 1 (female yes) and data for it derived from the Finnish Medical Birth Register. No data were available for young fatherhood. Marital status was retrieved from the Population Register Centre and coded as 0 (never married) / 1 (married) / 2 (divorced). No information about cohabitation could be obtained from the data to capture a common arrangement of living together with a partner without being married. Hospital Discharge Register provides data for all medical diagnoses with ICD 10 & 9 classifications (World Health Organization, 2020). The four chosen categories are coded as 0(no) / 1(yes) for any occurrence. Behavioral diagnoses include ICD-10 F90-92[1] categories for disturbance of activity and attention, conduct disorders and mixed disorders of conduct and emotions, including e.g. ADHD and ADD. Mood diagnoses include ICD-10 F3[1] category of mood affective disorders such as depression and bipolar disorders. Anxiety diagnoses include ICD-10 F4[1] category of neurotic, stress-related and somatoform disorders, such as phobias, anxiety and dissociative disorders. Substance diagnoses include ICD-10 F1[1] category of mental and behavioral disorders due to psychoactive substance use. Finnish Legal Register Centre provides data for criminal offences and convictions.

[1] From before 1997 matched with the equivalent diagnosis codes in ICD-9

Convictions variable was defined as dichotomous according to if a conviction has ever been placed (1) or not (0). Number of offences is recorded for each person as a numeric value.

Family level predictors aim to capture parental resources that can be intergenerationally transmitted and situation at home that affects the child growing up. They were chosen as follows:

- Education level (mother & father)
- Mean income (mother & father)
- Parental unemployment (mother & father)
- Parents' marital status
- Single parent household during early childhood
- Born to a teen parent
- Child protective services involved

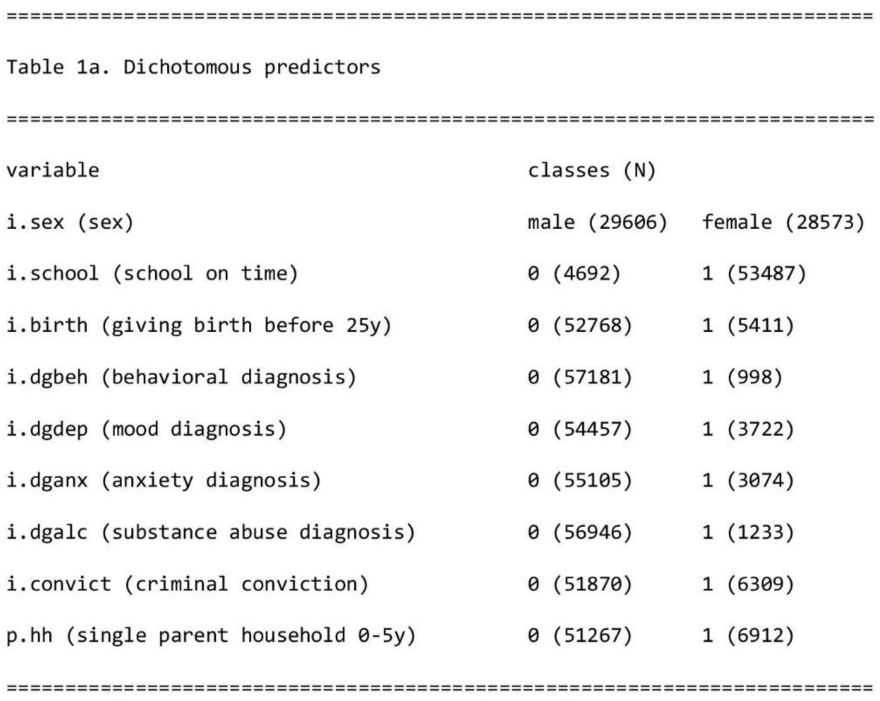
Education level data for both parents separately were derived from Statistics Finland and defined as the highest education level achieved before 2009 (the end of recording period). Coding of the variable differs from the one made for education level of the cohort member: a category 'post-secondary not university'[2] was added between secondary and bachelor degrees to account for degrees obtained before the educational reform in the 1990's, and the categories "master" and "postgrad" were separated, resulting in an integer range 0-5. Income data were derived from the Earnings Register of the Finnish Centre for Pensions and defined as the mean yearly income (without indexing) converted to Euro[3] across all years since an individual's work history started, separately for both parents. Missing values were again imputed with the average: for mothers, 180 missing values were replaced with 9099€ and for fathers, 872 missing values were replaced with 12435€. Parental unemployment was not operationalized similarly to the outcome variable due to lack of similar data: instead I defined it as the sum of days of received unemployment benefits between the years 1994-2011 (when such data were available), separately for each parent. Data for parents' marital status were derived from the Population Register Centre. The variable captures different timings and is categorized as follows: not married / married / divorced before school / divorced during school / divorced later / single parent at birth / parent dead. For cases with several logs of different statuses, e.g. married, divorced, married (244 observations), mean imputation was used (*married* most common value). Single parent household variable derives from the same data source but represents a restricted timing between ego's age 0-5 and accounts for different scenarios resulting in either two parent household (0) or one parent household (1). Teen parent is defined as ego's parents having been younger than 19

[2] Depending on the degree, they were categorized post-reform partly into secondary and partly into university education

[3] The Finnish Mark was used as a currency until 2002 when Euro replaced it

at the time of the child’s birth, coded as 0 (no) / 1 (one teen parent) / 2 (two teen parents), and data being derived from the Medical Birth Register. Register on Child Welfare provided data for children being placed in out-of-home care. The variable is coded as the year of entry into a placement, where no placement is coded as 2012.

Table 1 a-c. Descriptive tables for predictors (i. = ego, p. = parent)

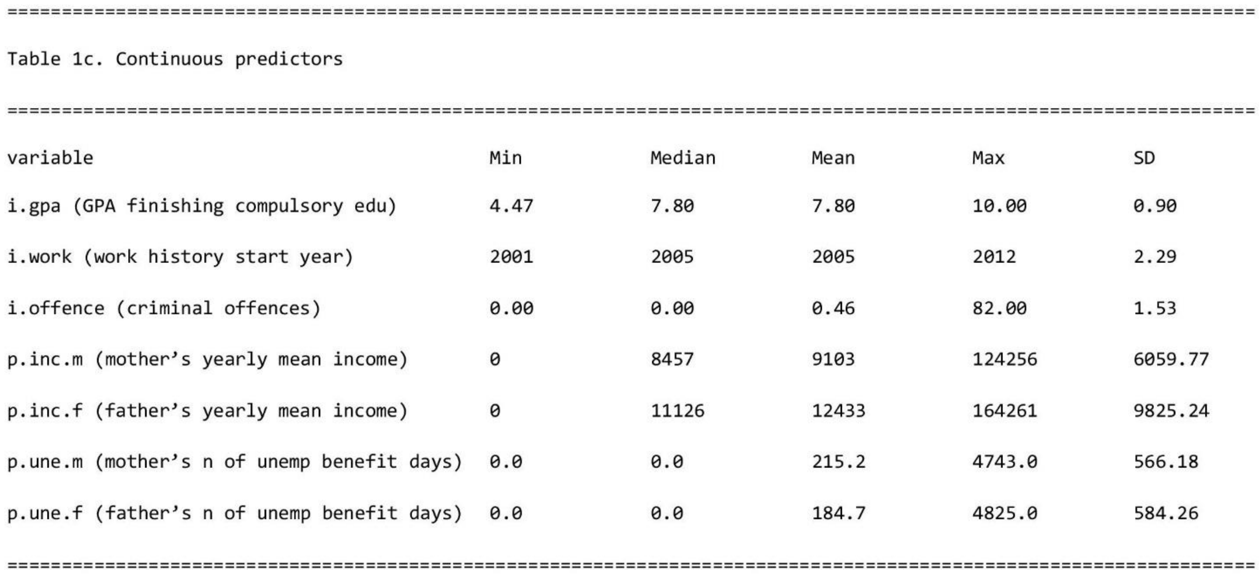


```
===========================================================================
Table 1a. Dichotomous predictors
===========================================================================
variable                                     classes (N)
i.sex (sex)                                  male (29606)   female (28573)
i.school (school on time)                    0 (4692)       1 (53487)
i.birth (giving birth before 25y)            0 (52768)      1 (5411)
i.dgbeh (behavioral diagnosis)               0 (57181)      1 (998)
i.dgdep (mood diagnosis)                     0 (54457)      1 (3722)
i.dganx (anxiety diagnosis)                  0 (55105)      1 (3074)
i.dgalc (substance abuse diagnosis)          0 (56946)      1 (1233)
i.convict (criminal conviction)              0 (51870)      1 (6309)
p.hh (single parent household 0-5y)          0 (51267)      1 (6912)
===========================================================================
```

Table 1b. Multiple class predictors

variable	classes (N)
i.married (marriage status)	never (52377); married (5432); divorced (370)
p.teenparent (parent’s younger than 19y)	no (56764); one parent (1222); two parents (193)
i.chserv (placement outside home)*	< 2000 (885); >= 2000 < 2012 (873); never (56421)
i.edu (education level)	no 2nd (7822); 2nd (40811); BA (8893); MA (653)
p.edu.m (mother’s education level)	no 2nd (9042); 2nd (26243); post 2nd not 3rd (12596); BA (5082); MA (4765); postgrad (451)
p.edu.f (father’s education level)	no 2nd (14230); 2nd (25245); post 2nd not 3rd (8070); BA (4685); MA (5067); postgrad (882)
p.married (parental status)	married (33802); not (6190); divorced (5614); div school (5070); div early (4749); dead (1972); single p. (782)

*coded for each year separately but due to small n of cases tabled as ranges



```
======================================================================================================

Table 1c. Continuous predictors

======================================================================================================

variable                                 Min         Median      Mean        Max         SD

i.gpa (GPA finishing compulsory edu)     4.47        7.80        7.80        10.00       0.90

i.work (work history start year)         2001        2005        2005        2012        2.29

i.offence (criminal offences)            0.00        0.00        0.46        82.00       1.53

p.inc.m (mother’s yearly mean income)    0           8457        9103        124256      6059.77

p.inc.f (father’s yearly mean income)    0           11126       12433       164261      9825.24

p.une.m (mother’s n of unemp benefit days) 0.0       0.0         215.2       4743.0      566.18

p.une.f (father’s n of unemp benefit days) 0.0       0.0         184.7       4825.0      584.26

======================================================================================================
```

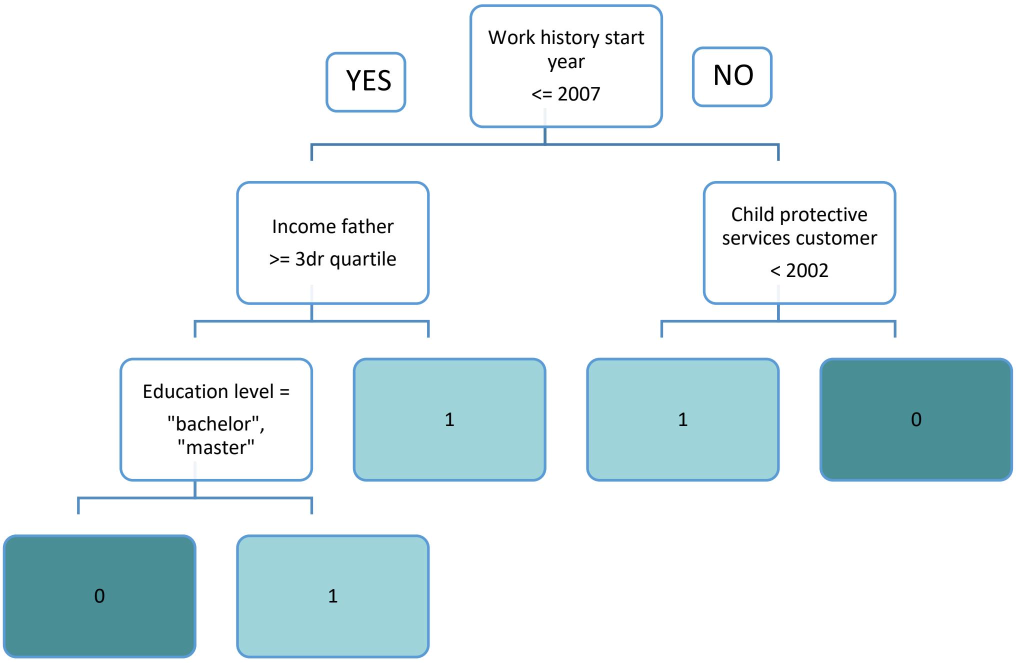
3.3. Analytical Approach

In this study I used supervised machine learning algorithms, more specifically classification trees (CART) and random forests (RF), to a) detect non-linearities and predictive interactions and b) predict the outcome. Rather than focusing on accuracy of prediction in the whole model, which has proved to be a difficult task in the past (Salganik et al., 2020), I put emphasis on correctly classifying the outcome variable class “unemployed” to be able to interpret the results in terms of critical factors and combinations of characteristics that are risk factors for long-term unemployment. CART algorithms were chosen to make use of their data-driven way of creating complex but parsimonious higher-order interactions found to be predictive of the outcome, while RF was chosen to smoothen the predictions and create robust models. Decision trees were built using the *rpart* package (Therneau, Atkinson & Ripley, 2019) to allow for non-linear effects and predictive interactions of several variables. To correct for potential overfitting from a single tree, a forest of trees was grown using the *randomForest* package (Breiman, Cutler, Liaw & Wiener, 2018). The two methods were then compared against each other to see if the simpler model setup using a single tree performs well on its own or if a forest gives additional advantage in terms of overfitting and predictive performance. The analysis was carried out using R version 3.6.2. Preliminary analysis was done using a random 10% subset of the data while the final models were evaluated using the full analytical sample. A cross-validation approach typical for machine learning was used in all analysis: data were split in two and the model trained with a large subset (70%) and evaluated with a smaller (30%).

3.3.1. CART & Random Forest

Both rpart and randomForest belong to the family of algorithms that build on decision trees: they perform either classification or regression tasks, depending on the outcome variable type (Hastie et al., 2009). While rpart only builds one tree, randomForest builds several and averages the results across all trees. A classification tree algorithm applies the logic of recursive partitioning and using a training dataset, optimizes if-else splits from predictor variables based on an *impurity measure*, to create tree branches representing combinations of variable ranges that together predict the outcome as accurately as possible. The approach does not specify a functional form in advance, but allows for a flexible model deriving from the training data. This feature also holds the key to such type of analysis: the training data used need to be rich, detailed and diverse enough with as little missing values as possible, in order to obtain good performance. Parameter settings in each algorithm, which are discussed in more detail below, can help improve the performance of the model. In the case of this study the outcome is dichotomous (employed / unemployed), making the task a classification one. See Figure 1 for an imaginary example of a classification tree. The variables on a branch and their ranges can be interpreted as interacting with each other to predict an outcome – individual and family level childhood characteristics that together suggest unemployment in early adulthood.

Figure 1. Example classification tree (0 = employed, 1 = unemployed)


YES
Work history start year <= 2007
NO
Income father >= 3dr quartile
Child protective services customer < 2002
Education level = "bachelor", "master"
1
1
0
0
1

To make the splits, the default impurity measure for both rpart and randomForest classification is Gini Impurity: it measures the probability of a random observation in the training data being incorrectly classified, based on the distribution of classes of the outcome variable in the data (Hastie et al., 2009). It aims to determine which (range of a) variable gives most information about the class the observation belongs to: the more frequent class in the data, the lower probability of misclassifying. After each new split, the calculation is adjusted to the subset that is left in the new node, aiming again to find out which variable provides most information about the correct class. Using this logic, the algorithm learns the training data and produces a tree model. After the trees are constructed using the training data, each individual case in the test data is analyzed through the tree, and based on their characteristics, they will follow different paths through the branches of the tree, finally ending up in one of the leaf nodes indicating either 0 (employed) or 1 (unemployed). In the case of this study, where the outcome class (or *label*) is known in the test set as well, I am able to compare the observed and predicted labels for the individuals in test data. The aggregated results of such analysis are displayed in a confusion matrix output (see Figure 2). Evaluation measures for decision tree classification are derived from the confusion matrix. In this classification task I defined unemployment as the positive outcome. The cells in the matrix are defined as True Positive (the unemployed individuals who are predicted as to be unemployed), False Negative (unemployed predicted as employed), True Negative (employed predicted as employed) and False Positive (employed predicted as unemployed).

Choosing the best evaluation measures depends on the research task at hand. Some typical ones include *accuracy* ($\frac{TP+TN}{N}$) which is the overall share of correct predictions out of all cases, *recall* ($\frac{TP}{TP+FN}$) which is the share of correctly predicted positive labels out of all positive observations, and *precision* ($\frac{TP}{TP+FP}$) which is the share of correctly predicted positive labels out of all positive predictions (Burkov, 2019). In my analyses, I chose to prioritize selecting the best model based on low *false negative rate* ($\frac{FN}{FN+TP}$) in order to minimize the risk that the model is falsely classifying unemployed individuals to be employed. I added *false positive rate* ($\frac{FP}{FP+TN}$) to check for to what extent the cases observed as employed are predicted as unemployed. While accuracy in not necessarily the best evaluation measure for life outcome prediction tasks (Salganik et al., 2020), I nevertheless wanted to see how the model performs in terms of accuracy as well. Finally, I included *F1-score*, a harmonic mean of precision and recall ($\frac{2\,(precision * recall)}{precision + recall}$), to aid interpreting the overall performance in terms of positive observations and predictions. Evaluation of the model performance prioritizes minimizing FNrate after which balancing it with FPrate, then maximizing accuracy and maximizing F1-score. The range for all evaluation measures is 0-1.

Figure 2. Confusion matrix labels

		PREDICTED	
		0	1
OBSERVED	0	True negative (TN)	False positive (FP)
	1	False negative (FN)	True positive (TP)

A great advantage of building one classification tree is its interpretability and reproducibility, while a weakness is the tendency to overfit and produce large variance. This can be adjusted by limiting the depth of the tree by allowing less predictors and by pruning branches to combine leaf nodes together and using an average value of the combined leaves. This is done to some extent by using the complexity parameter (see further section 3.3.2), but this approach can simultaneously introduce more bias into the model. To correct for the overfitting of a single tree, I grow a forest of trees to use an average prediction across all trees. Random forest algorithms build several decision trees following the logic described above, but each of them from a different subset of the training data. The trees are uncorrelated and each individual in the test set is evaluated for all the trees to produce as many outcome predictions as there are trees. An average is calculated of those predictions for each individual, resulting in smoothed predictions that mitigate overfitting. The predictions in the test set are again compared to the observed values, providing another confusion matrix to calculate evaluation measures from. Both models produce *importance scores* for each predictor variable, referring to a single predictor's contribution to the model. The importance score for each variable is obtained by calculating the loss of information if that variable was not used for splits, and then summing the loss: the more the variable contributes to reducing error in the model, the higher the importance. I compare the two models in terms of these outputs, as well as their performance based on the evaluation measures.

The two algorithms are expected to complement each other in answering the research questions by indicating the combinations of variables that together predict the outcome (CART) and balancing the overfitting when predicting the outcome (RF). Comparing the two algorithms reveals if they yield similar results to each other or if one has advantages over the other in a life outcome prediction task. Multicollinearity of predictors is handled by both algorithms in terms of choosing only one of possible correlating predictors when making splits. This can be a problem when building a single tree where only one of the correlating predictors can exist, but should be adjusted well with a forest of trees where all the correlating predictors can exist in different trees. Predictors only producing noise are assumed to have been excluded by the initial screening process done by the researcher based on theories and previous research.

3.3.2. Parameter Optimization

The process of choosing and adjusting the algorithm parameters aims to find an optimal fit for the model and help improve the model performance. This chapter introduces the parameters that I chose for both models, the values that were tested for each and the testing strategy for parameters. In addition to parameters inside the algorithms, class weighing options for data are introduced: based on the preliminary analysis the underrepresentation of the unemployed in the data was an issue, producing high false negative rates. In the case of my research, the optimal fit obtained by parameter optimization and class weighting aim for as many correct predictions as possible specifically for the outcome variable class "unemployed".

The parameters chosen for rpart include *cp*; a complexity parameter, *minsplit*; the minimum amount of observations in a node to attempt a split, and *minbucket*; the minimum amount of observations needed in a leaf node. The complexity parameter indicates the decrease in lack of fit needed to attempt a split: if a variable split does not provide enough gain for the model performance, it will not be attempted. Minsplit and minbucket help reduce the overfitting, and their values are typically specified as minbucket = minsplit */ 3* (they were tested pairwise according to this formula). Based on preliminary analysis, and accounting for the multiple number of cases in the final analysis compared to the preliminary, the following vectors of parameter values were tested:

- cp: [0, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01]
- minbucket: [10, 20, 30, 40, 50, 60]
- minsplit: [30, 60, 90, 120, 150, 180]

For randomForest, the parameters chosen were *ntree*; number of trees grown, and *mtry*; number of variables randomly sampled for candidates at each split. The number of trees that predictions are averaged across varies depending on the task and no rule of thumb exists. Preliminary analysis done with 10% of the analytical sample was performed with a large range (5, 50, 500 or 5000 trees) but this was reduced in the final analysis both because the preliminary analysis showed little difference between the number of trees, as well as limitations in computational power when growing hundreds or thousands of trees with a large dataset. The mtry value suggested by Breiman et al. (2018) is the (rounded up) square root of the number of predictors, in this case, 5. Preliminary analysis revealed some variance when adjusting this parameter, and therefore it was tested for with the full analytical sample as well. The equivalent of minbucket in randomForest is *nodesize,* but preliminary analysis revealed no difference when adjusting the value, so it was left out (this decision was later confirmed by the final rpart analysis results as well). The final parameter values tested for were the following:

- Ntree: [100, 500]
- Mtry: [3, 4, 5]

Preliminary analysis revealed high false negative rates for both models, which suggested a problem of imbalanced classes in the outcome variable. This stems from when the classes are not roughly equal in size, the algorithms interpret the dominant class (i.e. employed) to be more meaningful and prioritize correct predictions for the overrepresented (Menardi & Torelli, 2014). Different strategies of weighting the classes were tested to ensure the model addresses the research problem aiming to find important variable combinations that predict unemployment, rather than optimizing predicting the opposite, employment. Using classification tree based models to study marginalized groups of people, such as the unemployed, using data from a full cohort or population can thus be inherently problematic because such groups are typically in minority. Implementing higher weights or multiplying cases for the underrepresented class in the training data is one solution to this problem.

Four weighting strategies were compared to one with no weights: 1) applying equal weights to classes, 2) creating a new training set by doubling the amount of observed as unemployed, 3) creating a new training set by quadrupling the amount of observed as unemployed, and 4) ROSE oversampling on unemployed. The first approach functions through the built-in weight parameters in the algorithm, considering both classes of the outcome variable equally. The second and third approach multiply the cases with observed unemployment in the training data, resulting in more equal class sizes with doubling (43% unemployed) or overrepresentation of the unemployed with quadrupling (60% unemployed). The fourth approach, ROSE oversampling, synthetically creates a balanced dataset using a smoothed bootstrap approach: artificial cases of unemployed are created based on the observed cases (Lunardon et al., 2014). ROSE technique results in equal class sizes in the training data. All four approaches are tested for both tree and forest models, and compared against the unweighted model to evaluate which balancing strategy helps correctly classify the unemployed.

Testing for parameters and class balancing approaches was followed by stability tests for splitting the data to assess how random split to train and test sets affects the results. Stability tests were conducted as follows: 10 different 70%train / 30%test samples were created for both weighted and quadrupled[4] samples, by assigning a randomly sampled seed value for each. 10 more samples were created similarly, this time with 80%train / 20%test split. Stability could thus be assessed both in terms of how much the results are affected by the specific cases that end up in training data, as well as the ratio of training/test data. These two steps of analysis would have ideally been done simultaneously instead of subsequently, to test for all possible combinations, but due to limitations in computational power[5], they

[4] In the case of quadrupled sample, the split ratios do not remain as expressed in the text due to the multiplication of unemployed cases in the training data: 70/30 becomes 81/19, 80/20 becomes 88/12.

[5] Using an external computing solution, e.g. cloud computing services, was not an option due to the highly restricted data access to protect individual data of sensitive nature

were split into subsequent steps. Some manual individual cross-checks were done after completing both steps to bypass memory restrictions, and to ensure that parameter selection process was not affected by instability of the model produced by a random train/test split.

I chose the final models based on the tests: stability tests produced 10 models for each split ratio, and the best performing CART and RF in terms of FNrate were chosen. One tree model and one forest model were produced and compared to each other in terms of performance, predictors and importance scores. The tree model was visualized and interpreted in terms of the branches it produced: which variables interact with each other to produce predictions for the outcome. Robustness of the tree model results was increased by comparing the chosen model to the other 9 produced by stability tests as well as their importance scores, looking for consistency across the trees.

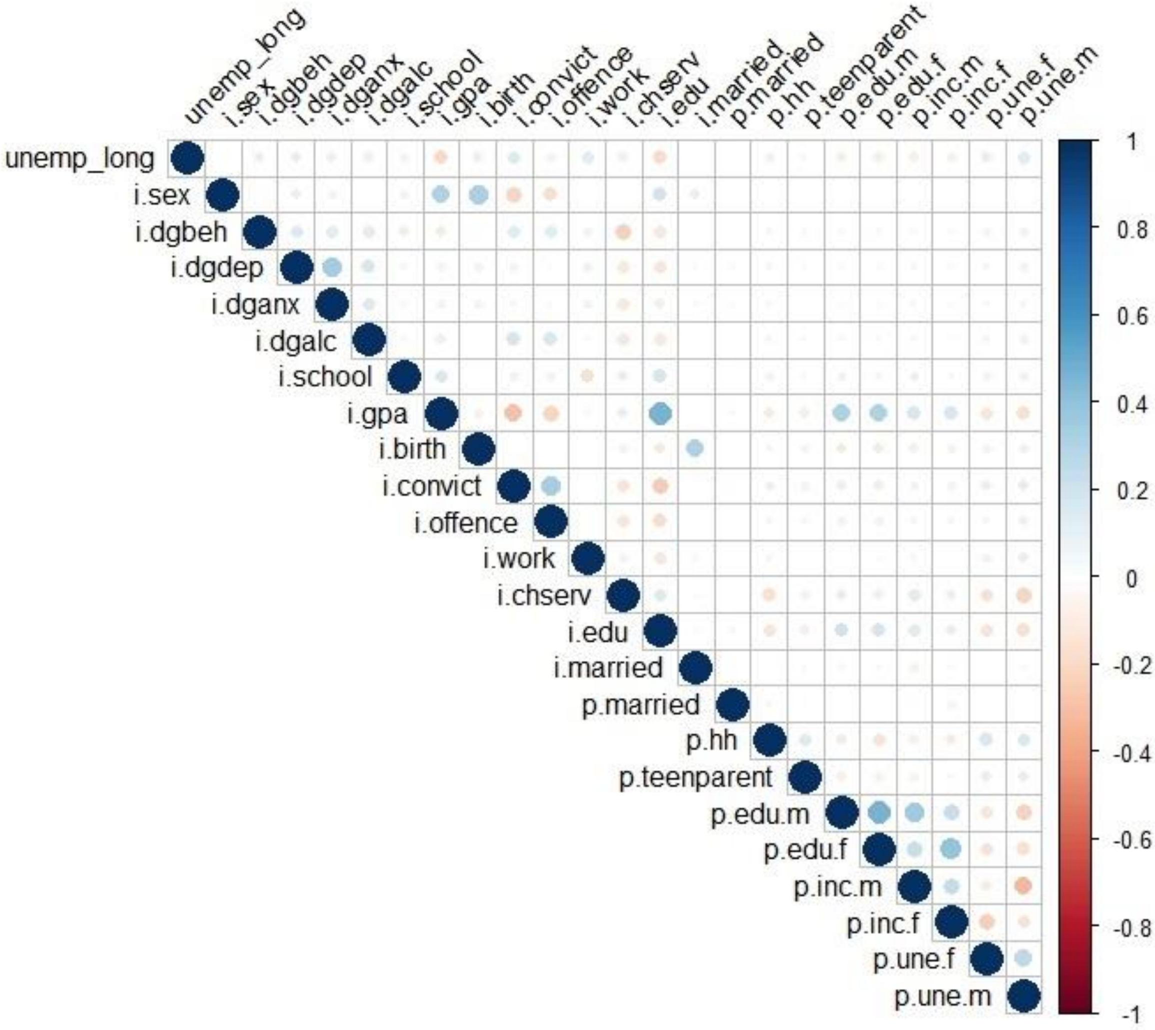
4. Results

This section begins with a descriptive correlation plot revealing bivariate associations in the data. I move on to describing the parameter and balancing test results, after which the stability tests results. Across all tests I compared evaluation measures prioritizing in the following order: 1) low false negative rate, 2) low false positive rate, 3) high accuracy and 4) high F1-score, to optimize a model performing well in terms of predicting unemployment. After the test results I draw the final models for both CART and random forest. I interpret the best performing CART by looking for consistency with the other 9 models produced by stability tests, both in terms of visual tree branches and importance scores: this approach reduces sensitivity of a single tree and produces more robust results. CART model results are then compared to the random forest model, in terms of most predictive variables based on importance scores as well as evaluation measures. CART and random forest models are compared to each other and conclusions drawn from how each model behaves throughout the analysis.

4.1. Bivariate correlations

A bivariate correlation plot was drawn using the full analytical sample (N 58179) to understand associations between variables in the data: multicollinearity allows for only one of potentially correlating predictors to be chosen for a single tree. Pearson correlation was used to produce a matrix for visualization, results are displayed in Figure 3. The plot does not reveal very high correlation between any two variables in the data: examples of strongest correlating variable pairs are ego's GPA and education level (.47), mother's & father's educational levels (.47), father's education and income (.39), mother's education and income (.36) as well as ego's depressive and anxiety diagnoses (.35). Whether these variable pairs exist in the same tree models can provide insight to how much correlation rpart and randomForest can handle between predictors.

Figure 3. Correlation plot



4.2. Parameters, balancing and stability

Beginning with CART models, I cross-tested all three parameters (*cp*, decrease in lack of fit; *minsplit*, minimum N of obs. in a node for a split; *minbucket*, minimum N of obs. left in a leaf node) with all the five class balancing approaches used to adjust the fact that there were fewer unemployed than employed in the data set (no weights, equal weights, doubling, quadrupling, ROSE). The tests were done with randomly splitting to train and test samples and repeated several times to get an overall idea of the test results. Results from one test with a random 70/30 split are displayed in Table 2 (only including cp 0.003). A first important pattern is that regardless the type of balancing used, the values of minsplit and minbucket had almost no effect on any of the evaluation measures. Therefore, minsplit and minbucket were left out of the final models, in order for the algorithm to optimize the performance. For the complexity parameter, 0 was consistently performing worse than > 0 (not presented in the table). Varying between different train/test splits, 0.002-0.005

performed best and often similarly to each other. The decision for the final complexity value was done based on looking at the visualized trees: cp 0.003 pruned trees to an interpretable size and was thus selected as the final complexity parameter value.

Different class balancing approaches yielded differing results in terms of how well the models performed. Overall, the models that performed best in order to predict unemployment were the quadrupled and weighted ones. Accuracy (the share of any correct predictions out of the whole sample) reached a maximum of approximately .74 in unweighted models. This means that for those models, a random observation in the data would be correctly classified with a 74% chance. Such high accuracy has a trade-off: false negative rate is high and an unemployed individual would be misclassified with an 81% chance. In addition, the F1-score (harmonized mean of precision and recall) was low across all models, reaching its maximum of .46-.47 in quadrupled and weighted samples: a result of no model being able to capture all cases equally. However, there were large differences in FNrate and FPrate across the models. The unweighted model had, as expected based on the preliminary results, high FNrates (.81). Simultaneously it produced very low FPrates (.05), suggesting to the typical feature of imbalanced class problem where the dominant class is prioritized (Menardi & Torelli, 2014). Using the inbuilt weight parameter to equalize classes, the FNrate dropped (.46) and simultaneously FPrate rose (.30), suggesting a slightly better balance between these two measures and more equal chances for a random observation from *either* class to be correctly labeled. Both the doubling strategy (.63) and ROSE oversampling (.76) underperformed in terms of FNrate, as well as other measures: based on this, artificially creating cases (ROSE) does not seem to help in improving model performance in this type of research task. Quadrupled sample clearly had the best performance in terms of FNrate: only 14% of the unemployed cases were mislabeled, producing a high 86% rate for correct classification. The trade-off from such high performance for the unemployed is high FPrate and lower accuracy: any random individual would be correctly classified with a 46% chance, while the employed would have a 69% chance of being misclassified.

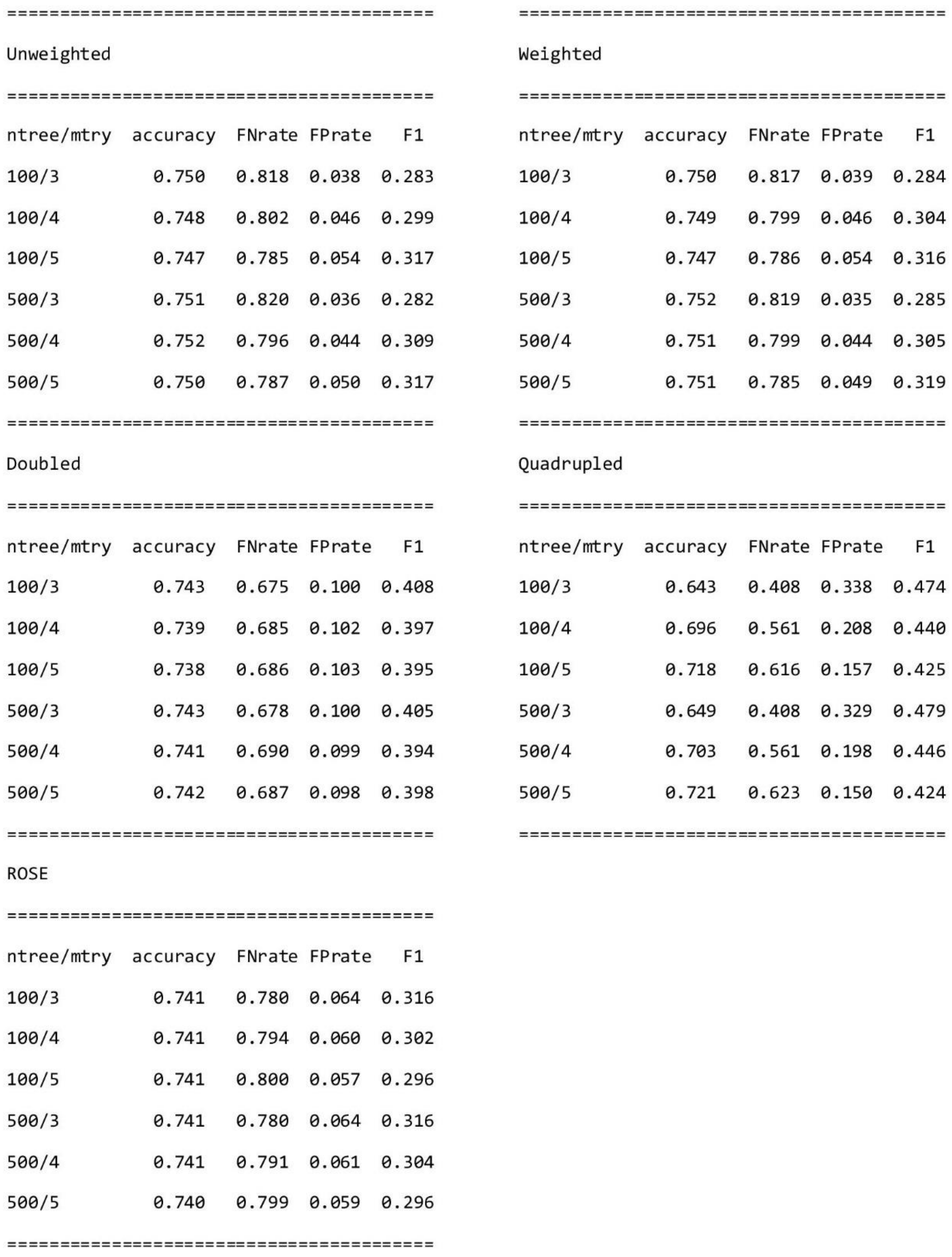
In conclusion, unbalancing the classes in favor of the unemployed was able to capture the unemployed much better than a balanced sample, with a trade-off of many employed being misclassified. What is interesting is that the models did not perform inversely when employed were overrepresented (unweighted) and when unemployed were (quadrupled). This can be partly due to the balancing technique used: another explanation could be that the variables chosen as predictors simply perform better in capturing employment than unemployment. According to the priority order in evaluation measures, a quadrupled sample clearly outperformed the rest. Nevertheless, a weighted sample produced lower FNrates compared to all models except the quadrupled, and *in addition* balanced FNrate with FPrate and produced a rather high overall accuracy in the model, thus performing nicely in all top-3 evaluation criteria. Moving forward in the results section, CART results are compared to the ones obtained from random forest models to confirm the final balancing approach.

Table 2. Parameter and class balancing test results, CART cp 0.003

```
=========================================        =========================================
Unweighted                                       Weighted
=========================================        =========================================
bucket/split accuracy FNrate FPrate   F1         bucket/split accuracy FNrate FPrate   F1
10/30           0.743  0.810  0.050  0.286       10/30           0.661  0.457  0.296  0.465
20/60           0.743  0.810  0.050  0.286       20/60           0.661  0.457  0.296  0.465
30/90           0.743  0.810  0.050  0.286       30/90           0.661  0.457  0.296  0.465
40/120          0.743  0.810  0.050  0.286       40/120          0.661  0.457  0.296  0.465
50/150          0.743  0.810  0.050  0.286       50/150          0.661  0.457  0.296  0.465
60/180          0.743  0.810  0.050  0.286       60/180          0.661  0.457  0.296  0.465
=========================================        =========================================
Doubled                                          Quadrupled
=========================================        =========================================
bucket/split accuracy FNrate FPrate   F1         bucket/split accuracy FNrate FPrate   F1
10/30           0.709  0.625  0.166  0.412       10/30           0.459  0.144  0.689  0.462
20/60           0.709  0.625  0.166  0.412       20/60           0.459  0.144  0.689  0.462
30/90           0.709  0.625  0.166  0.412       30/90           0.459  0.144  0.689  0.462
40/120          0.709  0.625  0.166  0.412       40/120          0.459  0.144  0.689  0.462
50/150          0.709  0.625  0.166  0.412       50/150          0.459  0.144  0.689  0.462
60/180          0.709  0.625  0.166  0.412       60/180          0.459  0.144  0.689  0.462
=========================================        =========================================
ROSE
=========================================
bucket/split accuracy FNrate FPrate   F1
10/30           0.706  0.761  0.120  0.306
20/60           0.706  0.761  0.120  0.306
30/90           0.706  0.761  0.120  0.306
40/120          0.706  0.761  0.120  0.306
50/150          0.706  0.761  0.120  0.306
60/180          0.706  0.761  0.120  0.306
=========================================
```

Next, I repeated the exercise for random forest. Two randomForest parameters (*ntree*, N of trees grown; *mtry*, N of predictors randomly sampled as a candidate for a split) were equally tested with all five class balancing approaches. Results are displayed in Table 3. The first notable pattern is that the number of trees has minor impact on how well the model performs. Therefore, ntree was set to minimum test value (100) according to the logic of choosing the simplest model. Mtry was not as simple to interpret: the results were not consistent across balancing approaches. Eventually mtry = 3 was selected based on the results from the top candidate balancing approach: quadrupled. Comparing the class balancing strategies, the results somewhat differ from CART: a quadrupled sample managed to bring FNrate down (.41) but not as much as it did in CART (.14). In addition, it simultaneously balanced FNrate with FPrate, similarly to how the weighted sample behaved with CART models: this was the case using mtry = 3, while the other values tested produced more unequal predictions between the classes, favoring the employed. The weighted sample in random forest on the other hand behaves similarly to the unweighted: keeping FNrate high (.82). This suggests to CART models tending to overfit to match the current data, specifically in the quadrupled sample. This can be due to the fact that quadrupled sample is sensitive to the cases sampled in the training set: whoever is picked, that specific combination of factors is magnified 4-times, leaving less room for variation. Nevertheless, according to these tests, random forest using quadrupled sample manages to correctly predict 59% of the unemployed, while the overall accuracy of the model is 64%.

Table 3. Parameter and class balancing test results, random forest

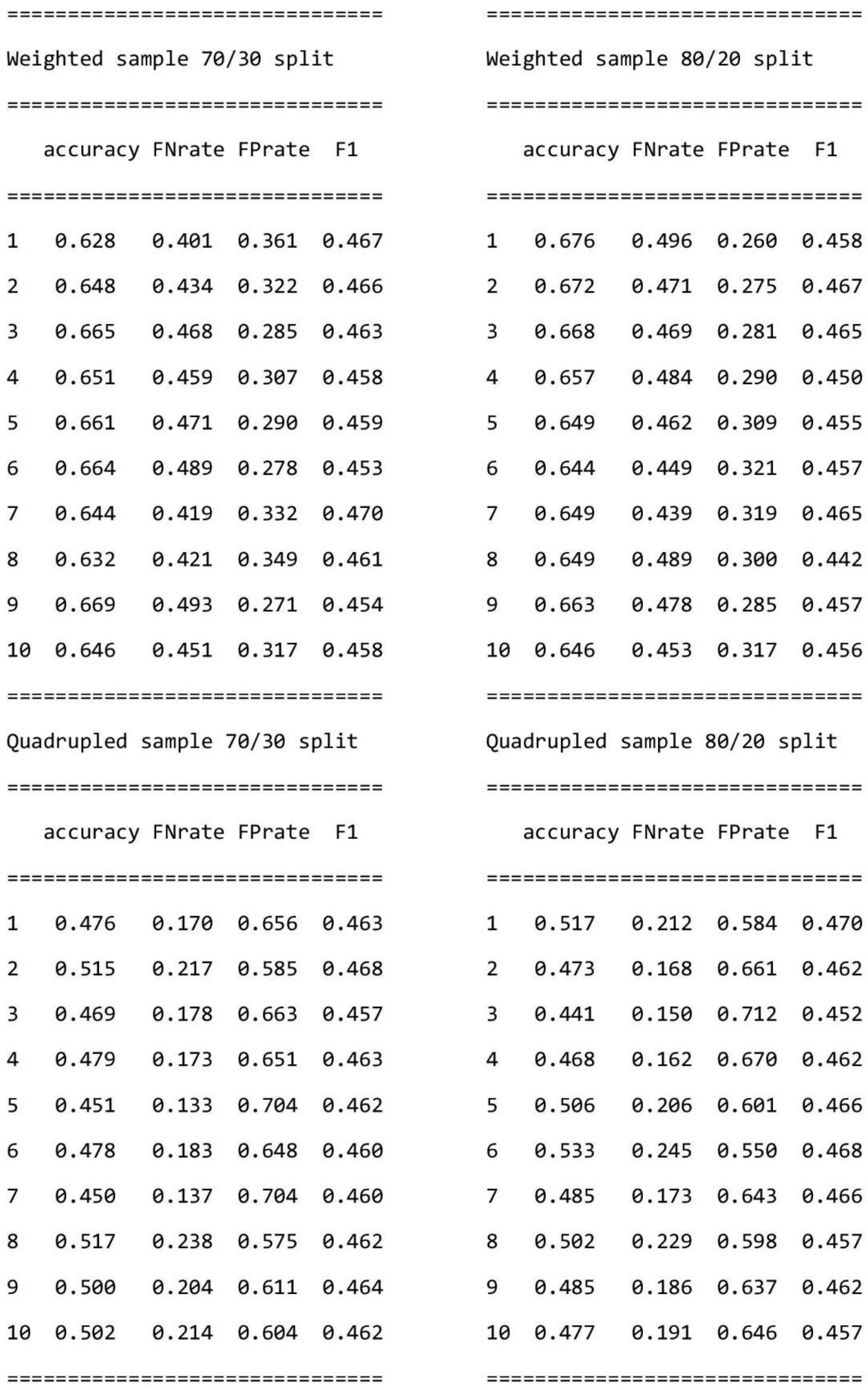


```
=========================================       =========================================
Unweighted                                      Weighted
=========================================       =========================================
ntree/mtry  accuracy  FNrate FPrate   F1        ntree/mtry  accuracy  FNrate FPrate   F1
100/3          0.750   0.818  0.038  0.283      100/3          0.750   0.817  0.039  0.284
100/4          0.748   0.802  0.046  0.299      100/4          0.749   0.799  0.046  0.304
100/5          0.747   0.785  0.054  0.317      100/5          0.747   0.786  0.054  0.316
500/3          0.751   0.820  0.036  0.282      500/3          0.752   0.819  0.035  0.285
500/4          0.752   0.796  0.044  0.309      500/4          0.751   0.799  0.044  0.305
500/5          0.750   0.787  0.050  0.317      500/5          0.751   0.785  0.049  0.319
=========================================       =========================================
Doubled                                         Quadrupled
=========================================       =========================================
ntree/mtry  accuracy  FNrate FPrate   F1        ntree/mtry  accuracy  FNrate FPrate   F1
100/3          0.743   0.675  0.100  0.408      100/3          0.643   0.408  0.338  0.474
100/4          0.739   0.685  0.102  0.397      100/4          0.696   0.561  0.208  0.440
100/5          0.738   0.686  0.103  0.395      100/5          0.718   0.616  0.157  0.425
500/3          0.743   0.678  0.100  0.405      500/3          0.649   0.408  0.329  0.479
500/4          0.741   0.690  0.099  0.394      500/4          0.703   0.561  0.198  0.446
500/5          0.742   0.687  0.098  0.398      500/5          0.721   0.623  0.150  0.424
=========================================       =========================================
ROSE
=========================================
ntree/mtry  accuracy  FNrate FPrate   F1
100/3          0.741   0.780  0.064  0.316
100/4          0.741   0.794  0.060  0.302
100/5          0.741   0.800  0.057  0.296
500/3          0.741   0.780  0.064  0.316
500/4          0.741   0.791  0.061  0.304
500/5          0.740   0.799  0.059  0.296
=========================================
```

Based on the results described above, I chose both the quadrupled and weighted samples to continue performing stability tests with: creating 10 train/test data sample for both 70/30 split and 80/20 split, and testing both tree and forest models on all data sets. Results for CART are displayed in   Table 4 and for random forest in   Table 5. Stability tests confirm the patterns observed in previous analyses: for trees, weighted samples balance the FNrate and FPrate better, having higher overall accuracy, while quadrupled manages to produce low FNrates, and for forests, the weighted sample maintains high FNrate while quadrupled balances FNrate and FPrate. Comparing evaluation measures across all 10 samples, there is some variation across the strength of the measures, but overall patterns remain unchanged. More variation can be observed in the quadrupled sample than weighted, confirming the sensitivity to individual cases being multiplied. Forest seems to balance this tendency slightly compared to trees, likely due to averaging across trees rather than relying on one tree only. Comparing 70/30 and 80/20 splits, there seems not to be much difference due to the train/test ratio: I chose to continue using 70/30 split as I had used in the analysis up until this point. I chose the quadrupled sample[6] for the final models because it manages to bring the FNrate down for trees, while managing to keep FNrate relatively low for forest. The overfitting tendency of trees is, in this case, in favor of what the model is intended to capture - factors contributing to unemployment. Final models are chosen between the 10 tests based on the lowest FNrate: number 5 for CART and number 6 for random forest.

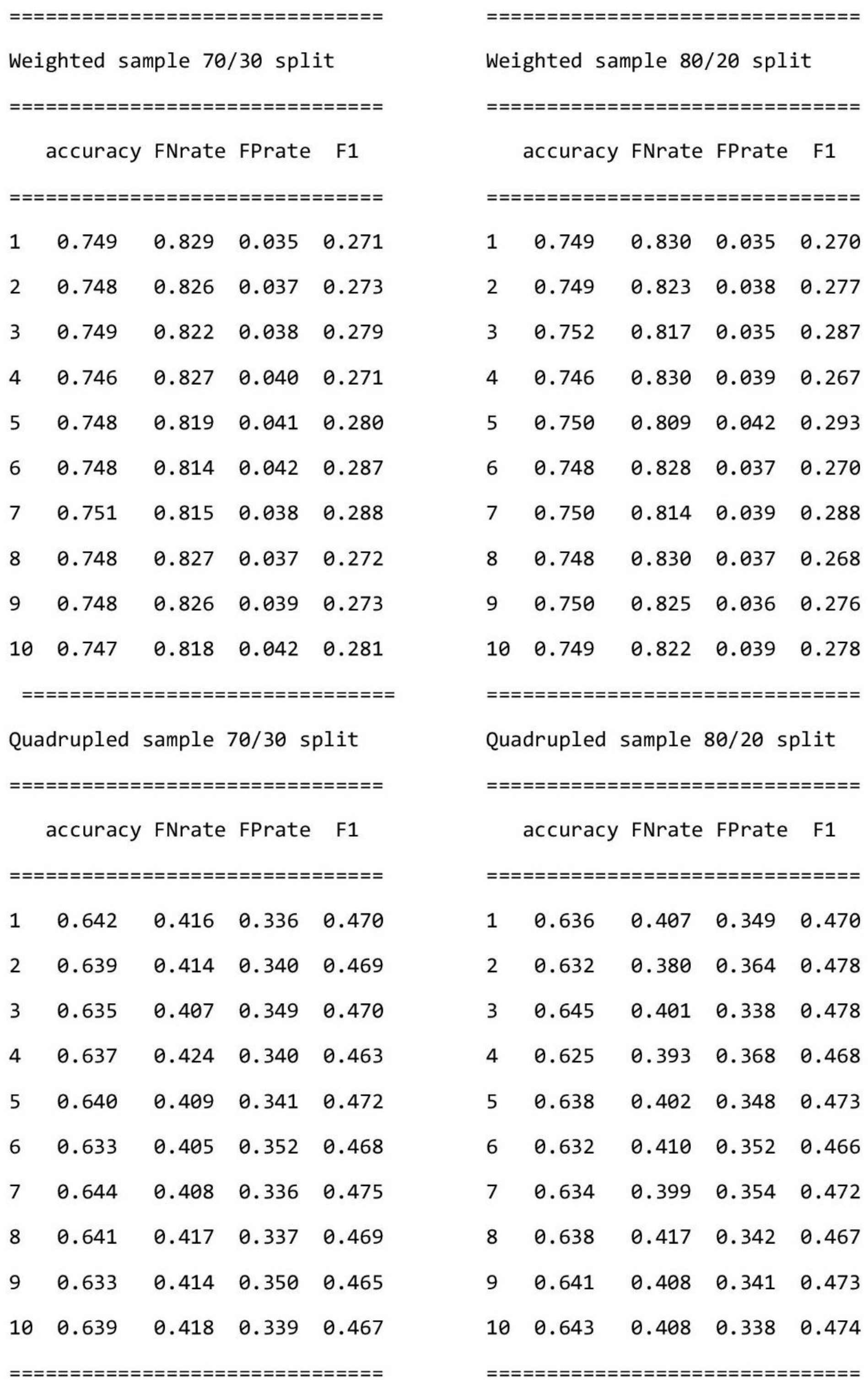
[6] Final sample size for quadrupled 70/30 was N 91392, of which training data N 73938, of which unemployed N 44284

Table 4. Stability tests, weighted and quadrupled samples, CART



```
===============================         ===============================
Weighted sample 70/30 split             Weighted sample 80/20 split
===============================         ===============================
   accuracy FNrate FPrate  F1              accuracy FNrate FPrate  F1
===============================         ===============================
1   0.628   0.401  0.361  0.467         1   0.676   0.496  0.260  0.458
2   0.648   0.434  0.322  0.466         2   0.672   0.471  0.275  0.467
3   0.665   0.468  0.285  0.463         3   0.668   0.469  0.281  0.465
4   0.651   0.459  0.307  0.458         4   0.657   0.484  0.290  0.450
5   0.661   0.471  0.290  0.459         5   0.649   0.462  0.309  0.455
6   0.664   0.489  0.278  0.453         6   0.644   0.449  0.321  0.457
7   0.644   0.419  0.332  0.470         7   0.649   0.439  0.319  0.465
8   0.632   0.421  0.349  0.461         8   0.649   0.489  0.300  0.442
9   0.669   0.493  0.271  0.454         9   0.663   0.478  0.285  0.457
10  0.646   0.451  0.317  0.458         10  0.646   0.453  0.317  0.456
===============================         ===============================
Quadrupled sample 70/30 split           Quadrupled sample 80/20 split
===============================         ===============================
   accuracy FNrate FPrate  F1              accuracy FNrate FPrate  F1
===============================         ===============================
1   0.476   0.170  0.656  0.463         1   0.517   0.212  0.584  0.470
2   0.515   0.217  0.585  0.468         2   0.473   0.168  0.661  0.462
3   0.469   0.178  0.663  0.457         3   0.441   0.150  0.712  0.452
4   0.479   0.173  0.651  0.463         4   0.468   0.162  0.670  0.462
5   0.451   0.133  0.704  0.462         5   0.506   0.206  0.601  0.466
6   0.478   0.183  0.648  0.460         6   0.533   0.245  0.550  0.468
7   0.450   0.137  0.704  0.460         7   0.485   0.173  0.643  0.466
8   0.517   0.238  0.575  0.462         8   0.502   0.229  0.598  0.457
9   0.500   0.204  0.611  0.464         9   0.485   0.186  0.637  0.462
10  0.502   0.214  0.604  0.462         10  0.477   0.191  0.646  0.457
===============================         ===============================
```

Table 5. Stability tests, weighted and quadrupled samples, random forest

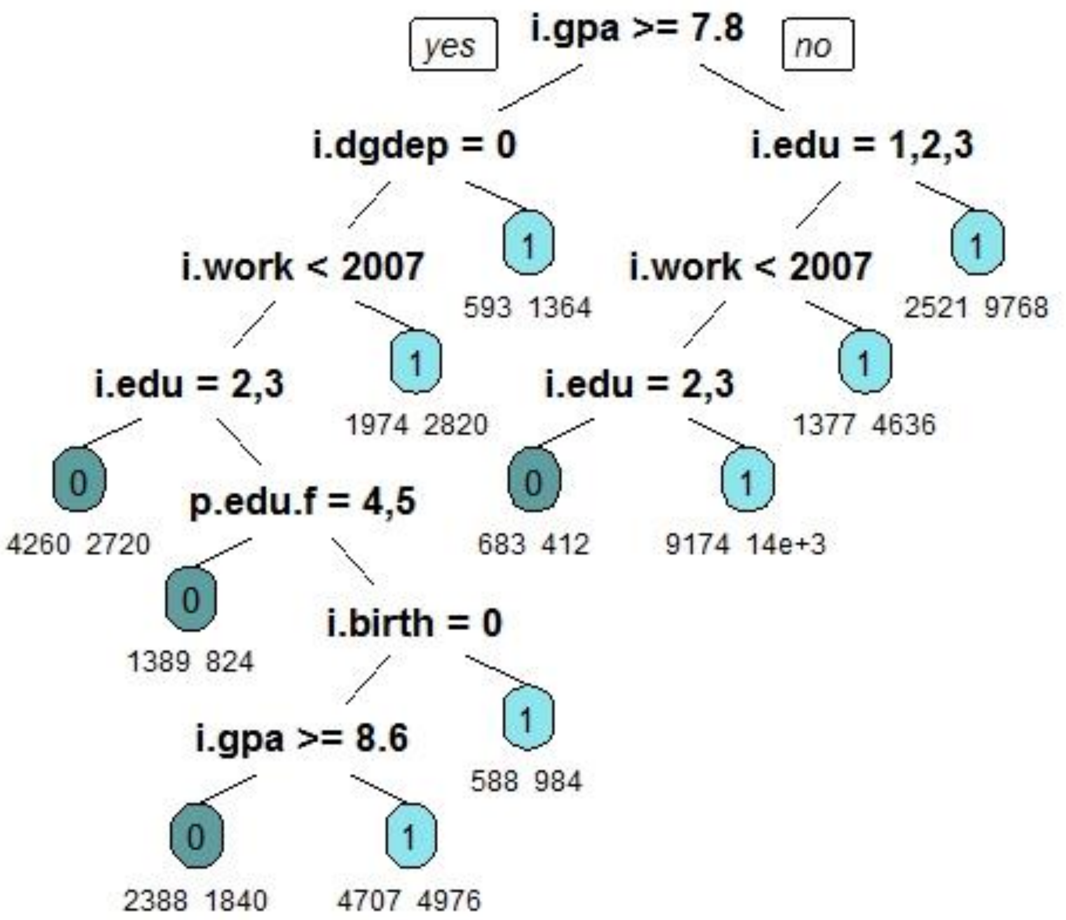


```
===============================       ===============================
Weighted sample 70/30 split           Weighted sample 80/20 split
===============================       ===============================
   accuracy FNrate FPrate  F1            accuracy FNrate FPrate  F1
===============================       ===============================
1   0.749   0.829  0.035  0.271       1   0.749   0.830  0.035  0.270
2   0.748   0.826  0.037  0.273       2   0.749   0.823  0.038  0.277
3   0.749   0.822  0.038  0.279       3   0.752   0.817  0.035  0.287
4   0.746   0.827  0.040  0.271       4   0.746   0.830  0.039  0.267
5   0.748   0.819  0.041  0.280       5   0.750   0.809  0.042  0.293
6   0.748   0.814  0.042  0.287       6   0.748   0.828  0.037  0.270
7   0.751   0.815  0.038  0.288       7   0.750   0.814  0.039  0.288
8   0.748   0.827  0.037  0.272       8   0.748   0.830  0.037  0.268
9   0.748   0.826  0.039  0.273       9   0.750   0.825  0.036  0.276
10  0.747   0.818  0.042  0.281       10  0.749   0.822  0.039  0.278
 ===============================      ===============================
Quadrupled sample 70/30 split         Quadrupled sample 80/20 split
===============================       ===============================
   accuracy FNrate FPrate  F1            accuracy FNrate FPrate  F1
===============================       ===============================
1   0.642   0.416  0.336  0.470       1   0.636   0.407  0.349  0.470
2   0.639   0.414  0.340  0.469       2   0.632   0.380  0.364  0.478
3   0.635   0.407  0.349  0.470       3   0.645   0.401  0.338  0.478
4   0.637   0.424  0.340  0.463       4   0.625   0.393  0.368  0.468
5   0.640   0.409  0.341  0.472       5   0.638   0.402  0.348  0.473
6   0.633   0.405  0.352  0.468       6   0.632   0.410  0.352  0.466
7   0.644   0.408  0.336  0.475       7   0.634   0.399  0.354  0.472
8   0.641   0.417  0.337  0.469       8   0.638   0.417  0.342  0.467
9   0.633   0.414  0.350  0.465       9   0.641   0.408  0.341  0.473
10  0.639   0.418  0.339  0.467       10  0.643   0.408  0.338  0.474
===============================       ===============================
```

4.3. Empirical results

Proceeding to interpret the final results, it is done in terms of visual tree and importance scores for CART (see Figure 4 and Table 6) as well as importance scores for random forest (Table 7). As one tree in CART is sensitive to the specific sample of training data, one needs to be careful when interpreting one single tree model. Due to the previous test phase, I have at my disposal ten trees constructed with the same data but with different sampling splits: I decided to interpret the results by comparing all trees looking for consistency across them, to gain more robust results. The other 9 trees are included in Appendix A-I and their importance scores in Appendix J-K. For random forest, only the importance scores for the final model are displayed. Importance scores for all forests are robust and thus the chosen model is enough to draw conclusions from. Moreover, no visual tree representation can be drawn from the forest model because it builds on and averages across 100 different trees. In a visual CART model, leaf node value 0 = employed while 1 = unemployed. A branch going left from a split indicates the threshold condition being fulfilled, while right side remains unfulfilled. Numeric values below the leaves indicate the number of observations from each class (left: employed; right: unemployed) in the training data, that ended up in the leaf nodes while training the model: the bigger the difference, the better the split predicts. Importance scores are interpreted according to their relative standing across models (CART and RF). The magnitude can be compared across one type of model (here: all different CART models) but not between CART and RF, due to their differing ways of model construction (one tree vs. 100 trees) and thus different way of calculating the importance.

Figure 4. CART tree number 5



I use visual trees and importance scores simultaneously to interpret the CART models. Across all CART models, GPA was the most important variable for predicting long-term unemployment, setting itself apart from the rest of the variables also by its magnitude: it is the single most predictive variable in all CART models. It was also used as the first split criteria in every tree, with a threshold varying between 7.5-7.8: GPA lower than this value is a strong predictor for unemployment as the trained model has 9768 unemployed in this leaf node as opposed to 2521 employed. The amount of employed remaining in this leaf node is the result from optimizing the model to prioritize correctly classifying unemployed rather than overall accuracy of the model. GPA sometimes reappears on the left side of the trees, often predicting high GPA (here: >= 8.6) to indicate employment. The second most important variable across all CARTs is ego's education. It often appears several times, separating between having obtained at least a secondary degree, having a high degree or having a secondary degree. In tree 5, for the individuals with GPA lower than 7.8, having not obtained a secondary degree strongly predicts unemployment. Lower on the left main branch (after early work history start) having a tertiary education predicts employment while having secondary predicts unemployment. Both variables GPA and ego's education exist in same trees, suggesting that their high correlation (.47, see Figure 3) is not an issue for rpart in this case.

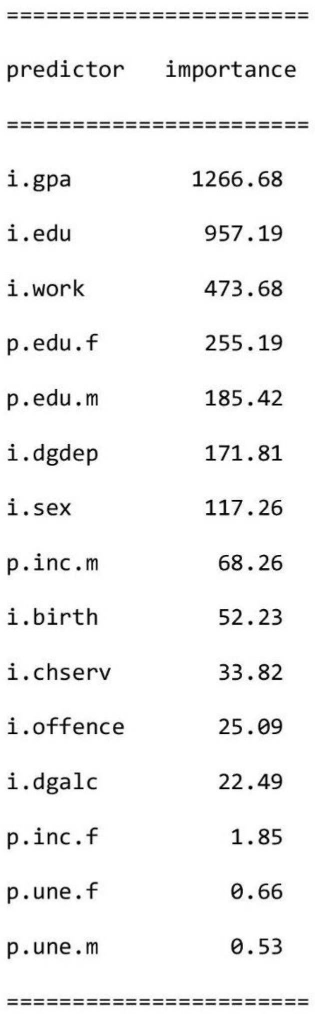
Identifying the two most important variables in the CART models was straightforward, but the models begin introducing more variance in important variables from the third place onwards. Across all CARTs, the top-6 most important variables often set themselves apart from the rest in terms of importance magnitude: I focus the interpretation to those variables. The third place in importance rankings is dominated (in frequency) by work history start (5), 4th place father's education (6), 5th place mother's education (6) and 6th place, rather equally shared, mood diagnosis (4), sex (3) and number of offences (3). The variables (with frequencies) that appear in top-6 importance scores in general are GPA (10), ego's education (10), father's education (7), mother's education (7), mood diagnosis (7), sex (6) and work history start (5). In the following these variables are interpreted as they appear in the tree branches.

In tree number 5, three more of the most important variables appear for splitting the nodes: work history start year, mood diagnosis and father's education. Work history appears on both sides of the tree with the same threshold split: not having had a first job latest in 2006, the year for the cohort to graduate from secondary education, is an important indicator of unemployment. Mood diagnosis is chosen as a split among the group with minimum GPA of 7.8. Given that an individual has reasonably high GPA, mood diagnosis come into play: if such a person has ever had e.g. a depression diagnosis, it is a strong predictor for unemployment. Similar combination appears in several trees, sometimes adding having a low education level to the mix. Simultaneously, for individuals who do not have high GPA, having a mood diagnosis is not important. In fact, only low education level and late work history start are important predictors combined with low GPA. Father's education level appears as the fifth split for individuals with high GPA, no mood diagnosis, who have started working early and who have maximum secondary education level: such individuals with fathers who do not have very high education level (master or higher), are at risk for unemployment. Similar patterns can be observed in other trees as well, father's low education level often following more splits with GPA or parental unemployment or income level.

Despite their importance, the three other top-6 variables that do not appear in tree 5 rarely appear in other trees either: sex appears in four trees, but neither criminal offence nor mother's education appear even once. Mother's education level missing in the trees is likely due to the fact that it correlates highly with father's education and they can be assumed to produce similar results: because of multicollinearity, the algorithm chooses only one of them. Following the same logic, also parental income level or unemployment could be interpreted as being important for both parents while only one appears in the tree. Interpreting number of offences leaves more room for guessing. One possible explanation is that since it correlates to some extent with ego's education level, which is such an important predictor in all models, it does not appear in the splits because it does not provide enough new information, but this is difficult to confirm. Sex typically appears after several splits: the models seem to predict unemployment for males with a GPA ~7.5-8.7, secondary education, no depression, who started working early and whose mother has max 1,5 years of accumulated unemployment spells. Due to the operationalization, giving birth before 25 years of age partly captures the effect of sex: as seen in tree 5 as well as some other trees, giving birth predicts unemployment for women in combination with high GPA, no depression, early work history start, low education and father's high education level.

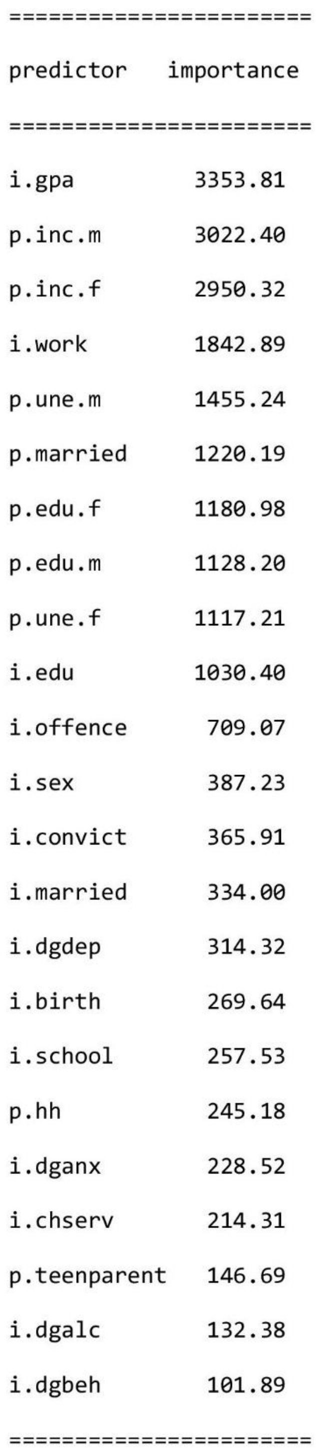
To conclude, there are patterns that repeat in several trees in terms of variable combinations and ranges that predict unemployment: cross-comparing the trees seems to provide rather robust results. Not all predictors in the data appear in one tree or have importance scores, due to the optimization process of the algorithm: some variables that do not contribute enough for the model (importance converges to 0) are left out. Moreover, not all variables that appear in the importance scores for a CART model appear in the visual tree. While potentially scoring high importance, a variable might not appear in the trees and thus produce uncertainty for interpretation: mother's education is missing most likely due to father's education being present, but number of offences is more difficult to track down. The further down in the importance score a variable lands, the more careful one should be when making conclusions about that variable.

Table 6. Importance scores for CART model



```
=======================

predictor   importance

=======================

i.gpa         1266.68

i.edu          957.19

i.work         473.68

p.edu.f        255.19

p.edu.m        185.42

i.dgdep        171.81

i.sex          117.26

p.inc.m         68.26

i.birth         52.23

i.chserv        33.82

i.offence       25.09

i.dgalc         22.49

p.inc.f          1.85

p.une.f          0.66

p.une.m          0.53

=======================
```

Table 7. Importance scores for random forest model



```
=======================
predictor   importance
=======================
i.gpa         3353.81
p.inc.m       3022.40
p.inc.f       2950.32
i.work        1842.89
p.une.m       1455.24
p.married     1220.19
p.edu.f       1180.98
p.edu.m       1128.20
p.une.f       1117.21
i.edu         1030.40
i.offence      709.07
i.sex          387.23
i.convict      365.91
i.married      334.00
i.dgdep        314.32
i.birth        269.64
i.school       257.53
p.hh           245.18
i.dganx        228.52
i.chserv       214.31
p.teenparent   146.69
i.dgalc        132.38
i.dgbeh        101.89
=======================
```

In order to compare CART and random forest models, Table 7 displays importance scores for RF. What is important to note is that the RF model produced evaluation measures differing from those of CART: while lowering the predictive power for unemployment (59%, CART 87%), the predictive power for employment rose (65%, CART 30%) and alongside with it the accuracy of the full model (63%, CART 45%). This has some consequences for interpreting the importance scores: while the random forest correctly predicts a fairly high share of unemployed, the importance scores of variables can now also suggest towards predicting the opposite, employment. Any differences compared to the CART models can thus reveal important variables that would correctly predict the employed. Another noteworthy aspect is that due to the smoothed predictions that derive from several trees trained with different subsets of the training data, random forest can reveal important variables predicting unemployment that are overlooked by CARTs which are forced to find good predictions for one outcome class and that do not allow for correlating predictors to exist in the model. Across all RFs, the ranking of variables and roughly also the absolute importance scores remained unchanged: the results are robust.

Unlike in CART, all predictors in the data gain some importance in random forest, varying in strength. GPA remains at the top of the ranking, confirming that it is the single most predictive variable for unemployment across models. Ego’s education drops in the ranking while being left with ~1/3 of the importance magnitude to that of GPA. However, the differences in magnitude of importance between ego’s education level, parental education levels and father’s unemployment are not very large, suggesting to a rather equal ranking status between these predictors. Work history start, while being an important variable in CARTs, has now gained an even higher ranking. Previously important sex and mood diagnosis have dropped in ranking. Income level of both parents have significantly raised their importance rankings compared to CART models, from being at the low end of ranking to being at the top. This suggests that parental income plays a large role in processes of (un)employment. Parental unemployment, particularly mother’s, has increased its ranking as well. Parental marital status is also now among the most important variables.

In conclusion, GPA is a strong and robust predictor for unemployment based on both CART and RF models. Work history start is equally among the most important variables in both models. Ego’s education level does not hold its ranking as well, but can still be interpreted as remaining among the important ones. A pattern emerging from random forests is the importance of parental variables for socioeconomic status and family situation (income, education, unemployment, marital status). CART and RF together capture factors that seem to interplay in processes of labor market attachment: individual level factors related to e.g. school achievement and early work life attachment as well as parental variables related to socioeconomic status and family situation.

5. Discussion

The research questions of this study aimed to find out which variables and their ranges are together predictive of long-term unemployment in early adulthood, and how CART and random forest algorithms compare to each other in such a research task. The results revealed that classification tree based models can be useful in distinguishing important variables and their ranges for life outcomes. In accordance to previous research predicting life outcomes, high predictive power in terms of accuracy for the whole model was troublesome to obtain. This is likely due to individual life course being inheritably heterogenous and complex, and not necessarily the method itself. Nevertheless, the CART model was able to correctly classify 87% of the unemployed, with the trade-off of misclassifying 70% of the employed. Random forest managed to correctly predict 59% of the unemployed and 65% of the employed, raising the model accuracy to ~63%: for social sciences, these are untypically high prediction rates. If we were to take any randomly selected individual, and let her fall down the trees in the random forest based on her measurable characteristics, and take an average of the results, in 63 percent of the cases we would predict her (un)employment status correctly.

Grade point average remained as the single most predictive variable across all models, suggesting that unemployment trajectories start to formulate already during school years. This is in line with previous research: using the same data source, Lallukka et al. (2019) also found GPA to be the most predictive variable. Ego's education level revealed to be another important individual level predictor: having not obtained a secondary degree is a strong predictor for unemployment. This was also concluded by Caspi et al. (1998). Having only secondary level degree is also a risk factor in terms of education. As GPA and education correlate, they can be seen to some extent to interplay in a chain reaction manner: while one's GPA is low, it is more difficult to acquire human capital in terms of high educational degrees. Work history start remained as one of the most predictive variables in both models: when one's first paid job occurs after the year of graduating from secondary education, it serves as a strong predictor for unemployment. Across all trees it operates for both groups, GPA lower or higher ~7.5, suggesting that it serves as a strong predictor regardless of school performance. Precarious work arrangements during studies, e.g. summer jobs and part-time work which are a typical form of double status in the context of Finland (Wolbers, 2003), seem to work as a predictor for employment in this case. However, the quality of employment cannot be confirmed with this analysis: it remains unknown whether precarious work arrangements help finding better job arrangements in the future (de Graaf-Zijl et al., 2011). Early work history start is also important from the perspective of scarring effects: when transitioning to work life, employers might statistically discriminate against individuals with no previous work experience, thus contributing to diminished chances of later employment (Helbling et al., 2019). Combinations and ranges of the three variables mentioned above, as indicated by the visual tree models and importance scores, is a result which would have been difficult to obtain only using traditional methods, speaking for the usefulness of classification tree based methods.

In addition to individual level, family level predictors matter as well. Socioeconomic status of the parents, in terms of income level, education level and unemployment predict work life trajectories: according to the results, highly educated parents with high income levels, who have not had to rely on unemployment benefits as a source of long-term income, can protect the child from unemployment, and vice versa. These results illustrate intergenerational transmission of (dis)advantage, where family background matters for individuals' life trajectories (e.g. Ganzeboom et al., 1991). Family situation in terms of parental marital status serves as an important predictor as well. However, it did not appear in any tree model, and cannot be interpreted in terms of higher-order interactions. Referring to the transmission of parental human capital through family social capital (Coleman, 1988), a link between parental and individual predictors can be drawn: family structure in terms of parents being separated or single parenting, as well as stress brought upon by unemployment can operate by reducing the intergenerational flow of human capital, playing a part in children's lowered school achievements, thus contributing towards unemployment trajectories.

In CART, focusing on correctly predicting the unemployed instead of aiming for accuracy in the whole model lead to interesting results in terms of complementing deductive research approaches: the models seem to capture similar factors to previous research within the same topic. In addition to validating previous research, the results provide some new insight: the higher-order interactions between variables and their ranges captured with CART would have been very difficult to obtain using e.g. linear or logistic regression models. The downside is that while overfitting the model to correctly label most cases of unemployed and to find combinations of predictive variables, CART mislabeled many cases of employed as being unemployed. It can be that there are more factors, which are not captured in the data used here, that help distinguishing the unemployed and employed that are predicted as unemployed. This can also be interpreted in terms of risk groups: the employed individuals who were mislabeled potentially belong to a risk group for unemployment, but simultaneously bear some resilience factors that helped them acquire and maintain a job despite the risk. This is an interesting direction for future research.

In terms of method validation, the testing phase revealed that even after optimal parameters have been distinguished, there was some uncertainty and instability in the models: this is the trade-off of a black-box algorithm guiding the analysis, as well as studying processes with plenty of complexity. Nevertheless, a good strategy proved to be relying on stability tests to address the known disadvantages of the algorithms. For CARTs, interpreting not only one randomly selected model but across models that are trained with different subsets of data (somewhat similarly to what RF does), revealed consistent patterns across them and gave more robustness for the models and leverage for interpretation. Moreover, comparing tree and forest models, mirroring to their known features, provides more insight. The known feature of CART, overfitting, was made to work in advantage of the wanted prediction outcome, unemployed. Complementing CARTs with random forest, CART results could be cross-checked: both individual and family related variables were found meaningful in RF, while parental socioeconomic status and family situation were more pronounced than in the CARTs. The known feature of RF, robustness, was confirmed in the analysis, and the results were found to complement and validate those of CART models.

To gain more robustness and interpretability in the models, different approaches can be further tested. To be able to rely more on one single CART model, different balancing options could be explored further. Quadrupled sample is sensitive to the cases sampled for training, magnifying those cases 4-times and leaving less room for variation. ROSE oversampling was not found very efficient in these tests. However, it produced equal class sizes in the tests: using ROSE or another similar approach for synthetic minority over-sampling (e.g. SMOTE) for creating unbalanced classes in favor of the minority, can potentially increase predictive power. Mean imputation technique was used for some variables (GPA, parental income levels, parental marital status), and while unlikely in this case with such small numbers of cases where the technique was used, it can introduce some bias in the models: individuals with missing GPA are likely to have low GPA thus not applying for secondary education alongside with the rest of the cohort. More precise imputation values can be obtained using e.g. a regression approach.

There are likely to be unobserved factors for processes of becoming unemployed, such as macro conditions or individual attitudes and values. Combining register data with survey data can provide a more comprehensive understanding of the topic. Feature selection techniques (e.g. LASSO or elastic net) can prove useful to bypass the potential bias produced by the researcher selecting meaningful variables based on known predictors: potentially predictive factors might be overlooked in the data available. The problem with this approach using classified big data with restricted access is that computational boosting with a super computer or cloud computing solutions are not applicable. Using small number of cases can be handled with a powerful stationary computer, but would also likely reduce the generalizability of the results studying complex topics with heterogeneity.

Validating these results further will provide more leverage for these results. In terms of machine learning approaches, out-of-sample validation is important for generalizability of the results: as is now, they provide information about a specific cohort, but not necessarily about other cohorts. The comparability of this exploratory approach to traditional parametric approaches can also be tested. Using the same data as this analysis, a comparative analysis using linear predictive modelling (see e.g. Mood, 2010) is proposed. If these approaches were to produce similar results in this type of research task, while CART providing higher-order interactions that traditional methods struggle to produce, this would be a strong incentive for further life course research using register data and CART based methods. Libraries exist for comparing classification tree algorithm results and providing more interpretability for the black box: *IML* and *DALEX* are examples of packages that produce more comparable results from classification and regression tree based models, including how much of a variable's importance is due to interactions. This could give an even more precise estimate of how much this kind of method adds to our understanding compared to e.g. a standard logistic regression where all interactions would need to be explicitly modelled.

6. Conclusion

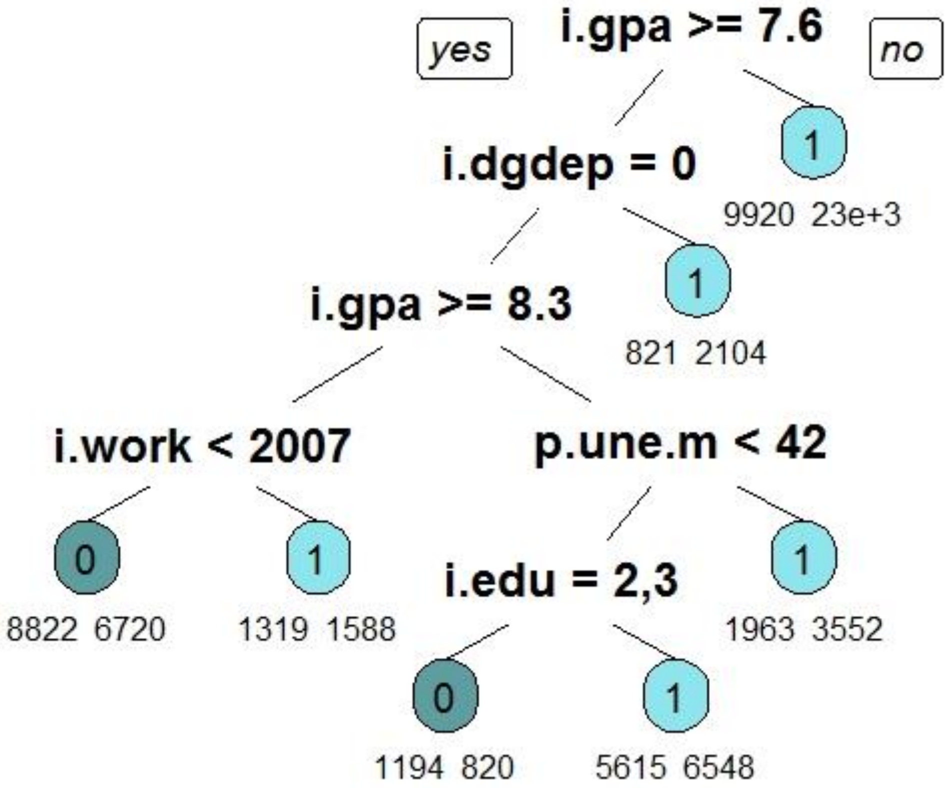
This study revealed individual and family level predictors important in trajectories for unemployment, which are in accordance to previous research. They can operate e.g. through processes of accumulating human capital and intergenerationally transferring it through family social capital. The contribution of this study is pinpointing critical factor combinations for unemployment in a data-driven way: allowing associations to arise from the data used in the analysis. CART and random forest algorithms used in this study were able to produce high prediction rates for unemployed: CARTs required a trade-off of misclassifying many employed, while random forest managed to balance prediction rates for both, simultaneously maintaining them rather high.

While being aware of the shortcomings of this fairly novel approach, some careful policy suggestion guidelines can be drawn based on the results. As school performance in terms of grade point average remained highly predictive across all models, interventions for the school learning environment to benefit the most poorly performing pupils is important. Providing more opportunities to gain work life experience before ending secondary school is encouraged. Support for families with low socioeconomic status in terms of e.g. family cohesion can decrease the negative effects for children. This can also create a more stress-free and fertile learning environment for the children, not only in schools but at home as well.

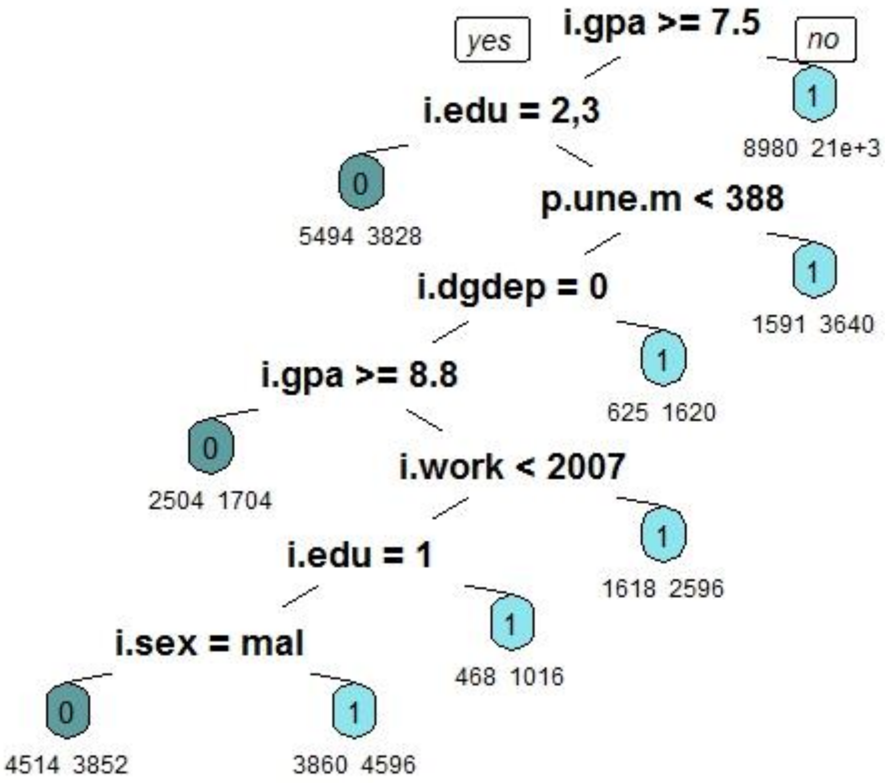
Several approaches mentioned in the discussion were left outside of the scope of this thesis due to time constraints. Further research in terms of the above mentioned is, however, highly encouraged. The contribution of this thesis is shedding more light in the field of life course and life outcomes research using non-traditional approaches utilizing machine learning techniques. In this subfield where little research has been conducted so far, this thesis also provides a suggested workflow for such a research project. Building on previous studies, this thesis gives further promise for using machine learning approaches in social research. There has been recent discussion about “failed” life outcome prediction and what it can teach us (Garip, 2020). The results of this study complement as well as nuance this discussion about using supervised machine learning in aiming to predict life outcomes: we should not give up just yet.

Appendices

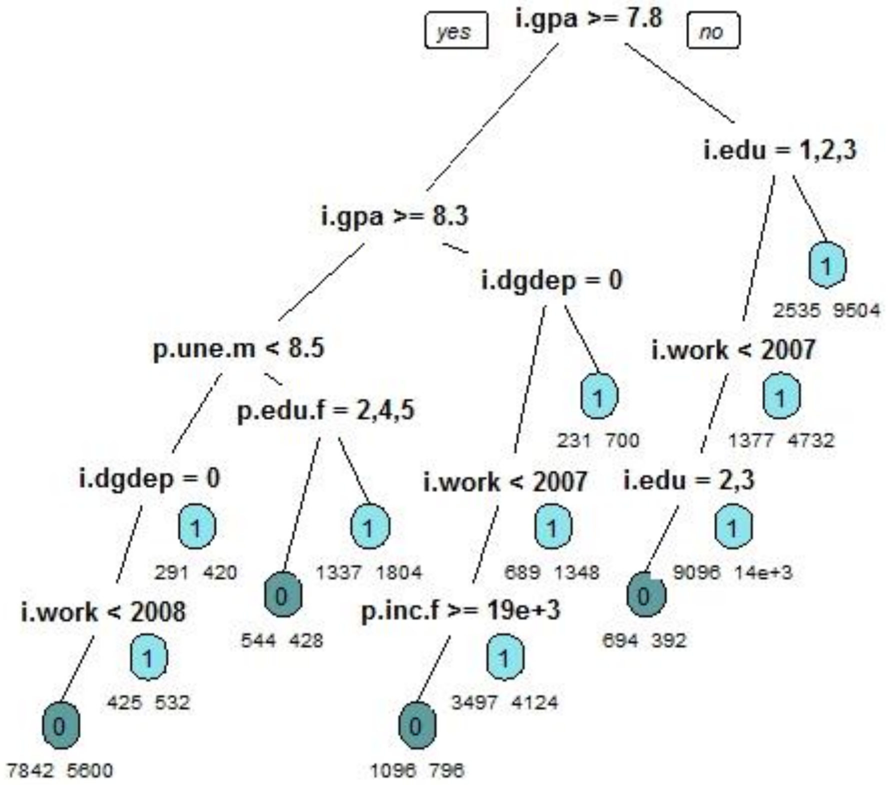
Appendix A. Tree #1



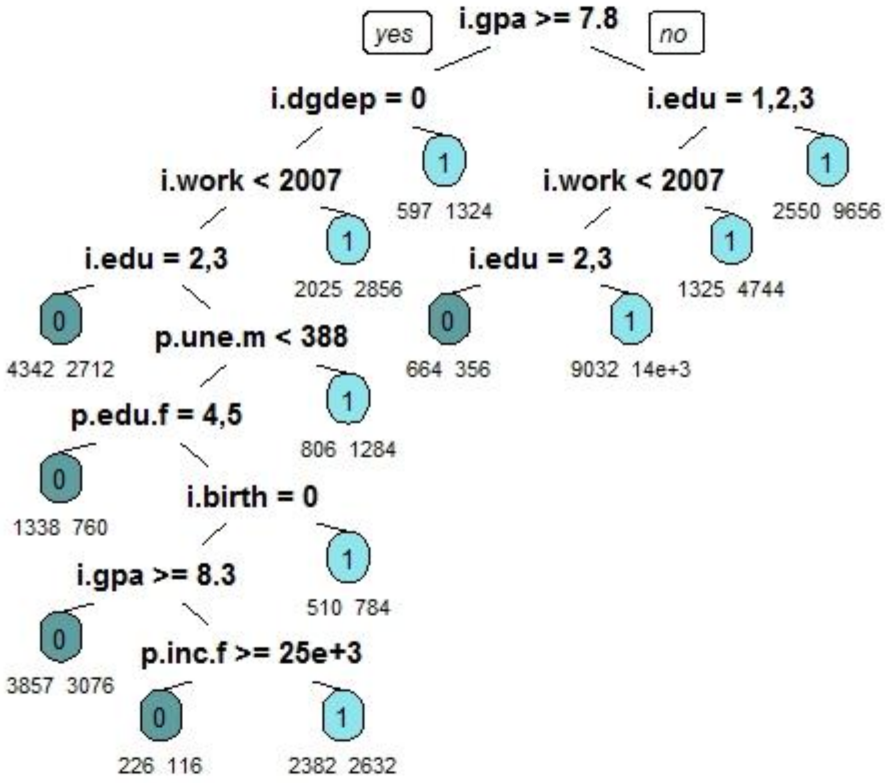
Appendix B. Tree #2



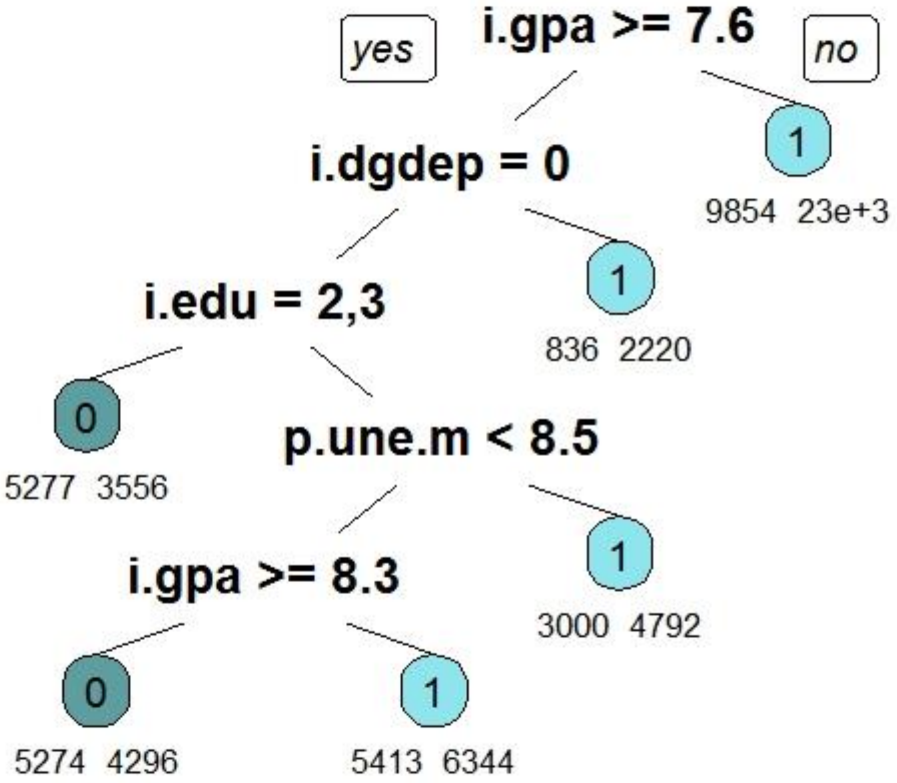
Appendix C. Tree #3



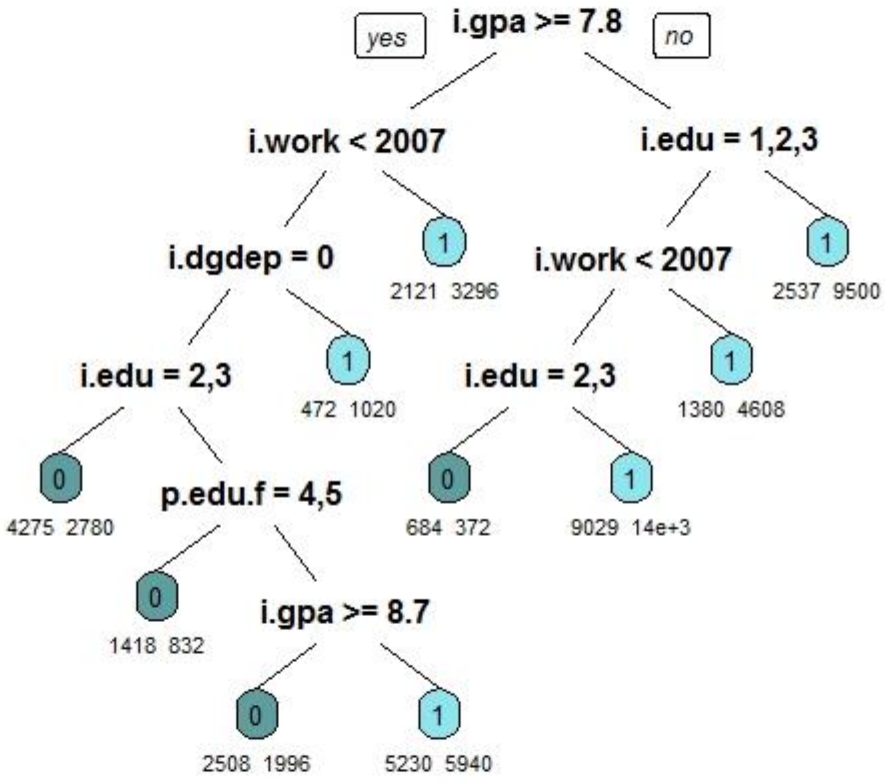
Appendix D. Tree #4



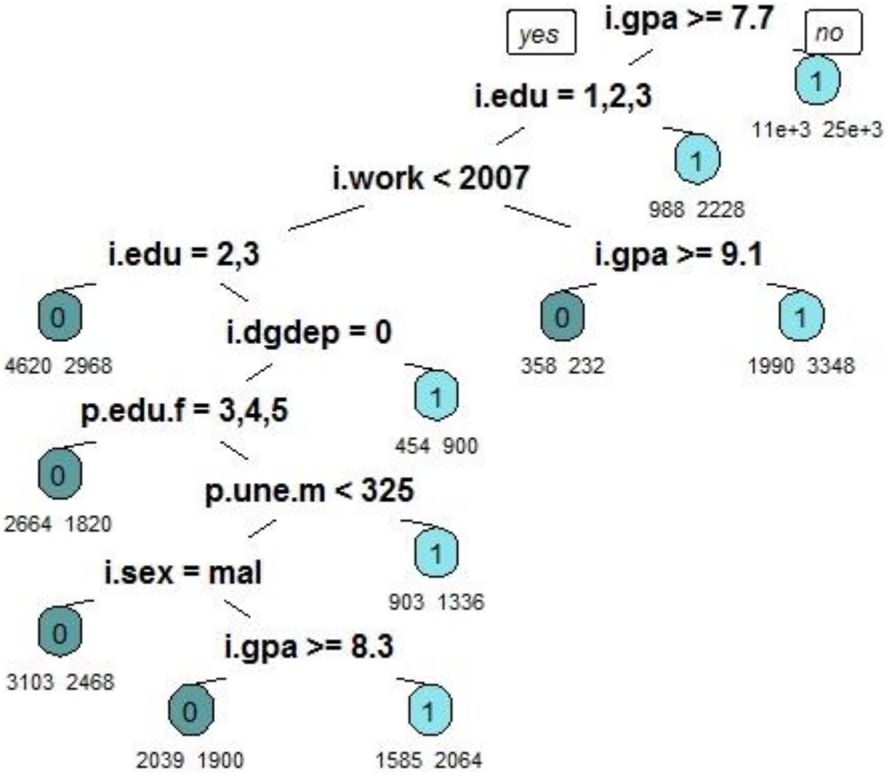
Appendix E. Tree #6



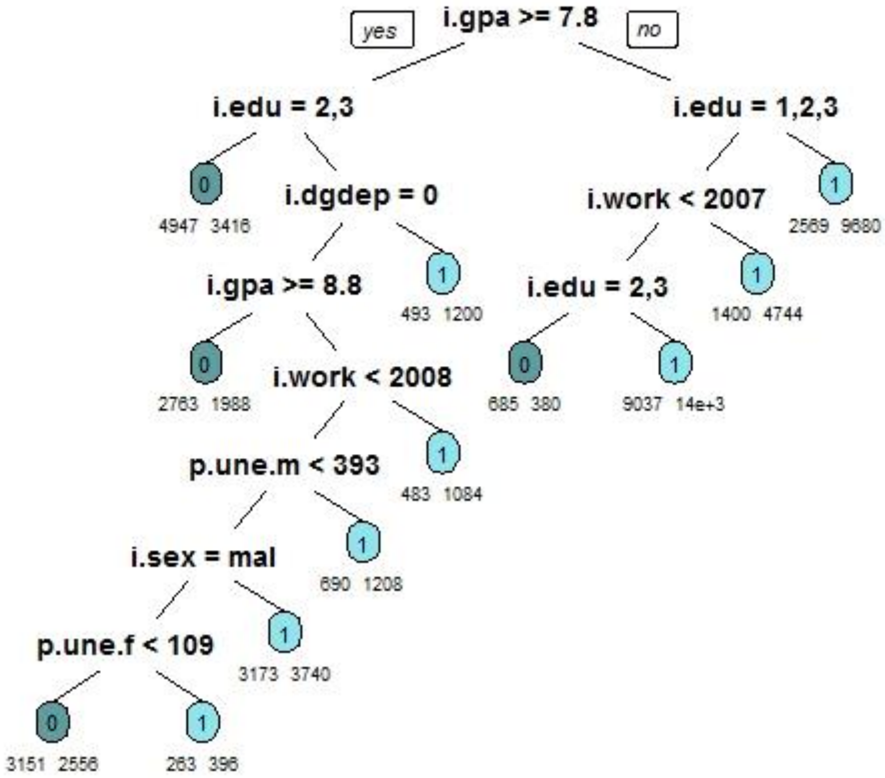
Appendix F. Tree #7



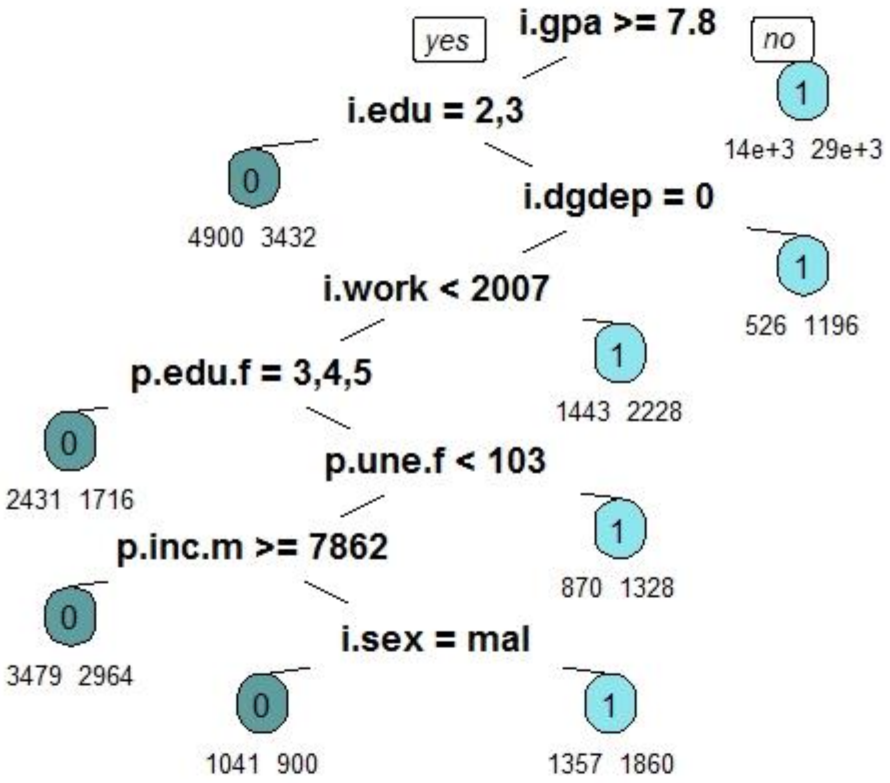
Appendix G. Tree #8



Appendix H. Tree #9



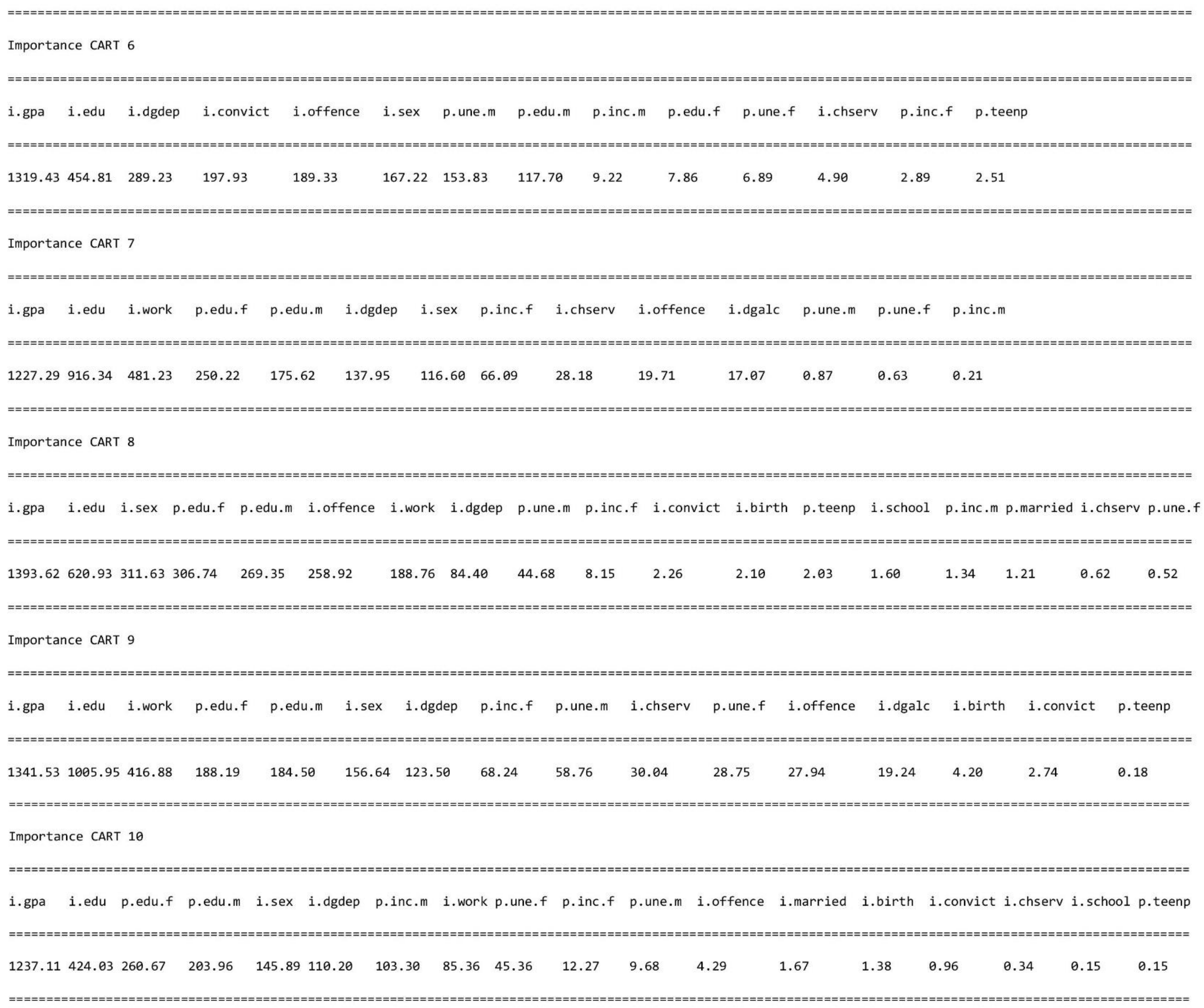
Appendix I. Tree #10



Appendix J. Importance scores CARTs 1-4

```
==============================================================================================================================================================
Importance CART 1
==============================================================================================================================================================
i.gpa   i.edu   i.dgdep   i.convict   i.sex   i.offence   p.edu.m   p.une.m   i.work   p.edu.f   p.inc.m   p.une.f   i.chserv    p.teenp
==============================================================================================================================================================
1373.84 321.86  251.59    190.91      181.77  174.40      132.04    122.78    63.53    26.82     24.27     8.42      4.83        2.12
==============================================================================================================================================================
Importance CART 2
==============================================================================================================================================================
i.gpa   i.edu   p.une.m   i.convict   i.dgdep   i.offence   i.work   p.une.f   i.sex   i.birth   p.edu.f   p.edu.m   i.chserv   p.teenp p.inc.f p.inc.m
==============================================================================================================================================================
1278.72 602.25  309.19    178.33      176.86    149.59      67.22    66.56     58.06   10.02     5.85      5.81      4.16       1.01         0.12    0.10
==============================================================================================================================================================
Importance CART 3
==============================================================================================================================================================
i.gpa   i.edu   i.work   p.edu.f   p.edu.m   i.sex   i.dgdep   p.une.m   p.inc.m   p.inc.f  i.chserv  i.offence  i.dgalc  i.convict  p.une.f  i.school p.teenp
==============================================================================================================================================================
1339.24 829.56  451.13   208.14    179.42    127.32  112.24    78.01     66.02     45.30    25.87     23.35      15.85    5.07       3.45     2.03     0.50
==============================================================================================================================================================
Importance CART 4
==============================================================================================================================================================
i.gpa   i.edu   i.work   p.edu.f   p.edu.m  i.dgdep  i.sex  p.inc.f  p.une.m  i.birth  i.chserv  i.offence  i.dgalc  p.une.f  i.convict  i.school  p.inc.m
==============================================================================================================================================================
1333.56 957.12  492.74   258.85    209.56   159.95   131.81 98.76    77.55    40.77    29.48     21.81      19.55    1.56     0.98       0.49      0.30
==============================================================================================================================================================
```

Appendix K. Importance scores CARTs 6-10



```
==========================================================================================================================================================================
Importance CART 6
==========================================================================================================================================================================
i.gpa   i.edu   i.dgdep   i.convict   i.offence   i.sex   p.une.m   p.edu.m   p.inc.m   p.edu.f   p.une.f   i.chserv   p.inc.f   p.teenp
==========================================================================================================================================================================
1319.43 454.81  289.23    197.93      189.33      167.22  153.83    117.70    9.22      7.86      6.89      4.90       2.89      2.51
==========================================================================================================================================================================
Importance CART 7
==========================================================================================================================================================================
i.gpa   i.edu   i.work   p.edu.f   p.edu.m   i.dgdep   i.sex   p.inc.f   i.chserv   i.offence   i.dgalc   p.une.m   p.une.f   p.inc.m
==========================================================================================================================================================================
1227.29 916.34  481.23   250.22    175.62    137.95    116.60  66.09     28.18      19.71       17.07     0.87      0.63      0.21
==========================================================================================================================================================================
Importance CART 8
==========================================================================================================================================================================
i.gpa   i.edu  i.sex  p.edu.f  p.edu.m  i.offence  i.work  i.dgdep  p.une.m  p.inc.f  i.convict  i.birth  p.teenp  i.school  p.inc.m p.married i.chserv p.une.f
==========================================================================================================================================================================
1393.62 620.93 311.63 306.74   269.35   258.92     188.76  84.40    44.68    8.15     2.26       2.10     2.03     1.60      1.34    1.21      0.62     0.52
==========================================================================================================================================================================
Importance CART 9
==========================================================================================================================================================================
i.gpa   i.edu   i.work   p.edu.f   p.edu.m   i.sex   i.dgdep   p.inc.f   p.une.m   i.chserv   p.une.f   i.offence   i.dgalc   i.birth   i.convict   p.teenp
==========================================================================================================================================================================
1341.53 1005.95 416.88   188.19    184.50    156.64  123.50    68.24     58.76     30.04      28.75     27.94       19.24     4.20      2.74        0.18
==========================================================================================================================================================================
Importance CART 10
==========================================================================================================================================================================
i.gpa   i.edu  p.edu.f  p.edu.m  i.sex  i.dgdep  p.inc.m  i.work p.une.f  p.inc.f  p.une.m  i.offence  i.married  i.birth  i.convict i.chserv i.school p.teenp
==========================================================================================================================================================================
1237.11 424.03 260.67   203.96   145.89 110.20   103.30   85.36  45.36    12.27    9.68     4.29       1.67       1.38     0.96      0.34     0.15     0.15
==========================================================================================================================================================================
```

References

Abebe, D. S., & Hyggen, C. (2019). Moderators of Unemployment and Wage Scarring During the Transition to Young Adulthood: Evidence from Norway. In B. Hvinden, J. O'Reilly, M. A. Schoyen, & C. Hyggen (Eds.), *Negotiating Early Job Insecurity - Well-being, Scarring and Resilience of European Youth*. Edward Elgar Publishing Ltd.

Arnett, J. J. (2000). Emerging Adulthood: A Theory of Development From the Late Teens Through the Twenties. *American Psychologist*, *55*(5), 469–480. https://doi.org/10.1037//0003-066X.55.5.469

Athey, S., & Wager, S. (2019). *Estimating Treatment Effects with Causal Forests: An Application*. arXiv.

Becker, G. S. (1975). *Human Capital: a Theoretical and Empirical Analysis, with Special Reference to Education* (G. S. Becker (ed.)). National Bureau of Economic Research. https://doi.org/10.2307/1401709

Black, S. E., & Devereux, P. (2010). Recent Developments in Intergenerational Mobility. In *NBER Working Paper Series* (No. 15889). https://doi.org/10.1093/acprof:osobl/9780199587377.003.0014

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2005). Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital. *The American Economic Review*, *95*(1), 437–449. http://www.nber.org/papers/w10066%5Cnpapers://e09fda77-1450-4449-8ecf-5a9bb72f5b0a/Paper/p3107

Boelaert, J., & Ollion, E. (2018). The Great Regression. Machine Learning, Econometrics, and the Future of Quantitative Social Sciences. *Revue Française de Sociologie*.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–215.

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2018). *randomForest* (4.6-14). https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Burkov, A. (2019). *The Hundred-Page Machine Learning Book* (1st ed.). Andriy Burkov.

Cambridge Dictionary. (2020). Algorithm. In *Cambridge Dictionary* (Online). Cambridge Univerisity Press. https://dictionary.cambridge.org/dictionary/english/algorithm

Caspi, A., Wright, B. R. E., Moffitt, T. E., & Silva, P. A. (1998). Early Failure in the Labor Market: Childhood and Adolescent Predictors of Unemployment in the Transition to Adulthood. *American Sociological Review*, *63*(3), 424–451.

Clark, A. E., Georgellis, Y., & Sanfey, P. (2001). Scarring: The Psychological Impact of Past Unemployment. *Economica*, *68*, 221–241. https://doi.org/10.1111/1468-0335.00243

Clark, A. E., & Lepinteur, A. (2019). The Causes and Consequences of Early-Adult Unemployment: Evidence from Cohort Data. *Journal of Economic Behavior and Organization*, *166*, 107–124. https://doi.org/10.1016/j.jebo.2019.08.020

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *The American Journal of Sociology*, *94*.

Coleman, J. S. (1994). Social Capital, Human Capital, and Investment in Youth. In A. C. Petersen & J. T. Mortimer (Eds.), *Youth Unemployment and Society*. Cambridge University Press. https://doi.org/10.1017/CBO9780511664021

Currie, J. (2009). Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development. *Journal of Economic Literature*, *47*(1), 87–122. https://doi.org/10.1257/jel.47.1.87

Daoud, A., & Johansson, F. (2019). *Estimating Treatment Heterogeneity of International Monetary Fund Programs on Child Poverty with Generalized Random Forest*. 1–52. https://doi.org/10.31235/osf.io/awfjt

de Graaf-Zijl, M., van den Berg, G. J., & Heyma, A. (2011). Stepping Stones for the Unemployed: The effect of Temporary Jobs on the Duration Until (Regular) Work. *Journal of Population Economics*, *24*, 107–139. https://doi.org/10.1007/s00148-009-0287-y

Dohmen, T., & Van Landeghem, B. (2019). *Numeracy and Unemployment Duration* (No. 12531; IZA Discussion Paper Series).

Doku, D. T., Acacio-Claro, P. J., Koivusilta, L., & Rimpelä, A. (2018). Health and Socioeconomic Circumstances over Three Generations as Predictors of Youth Unemployment Trajectories. *The European Journal of Public Health*, *29*(3), 517–523. https://doi.org/10.1093/eurpub/cky242

Erikson, E. (1968). *Identity, Youth and Crisis*. W. W. Norton Company.

Eurostat. (2020). *Unemployment Statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php/Unemployment_statistics#Youth_unemployment

Ganzeboom, H. B. G., Treiman, D. J., & Ultee, W. C. (1991). Comparative Intergenerational Stratification Research: Three Generations and Beyond. *Annual Review of Sociology*, *17*(1), 277–302. https://doi.org/10.1146/annurev.so.17.080191.001425

Garip, F. (2020). What Failure to Predict Life Outcomes Can Teach Us. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8234–8235. https://doi.org/10.1073/pnas.2003390117

Hao, K. (2019, January). AI is Sending People to Jail —and Getting it Wrong. *MIT Technology Review*. https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference and Prediction* (2nd ed.). Springer.

Helbling, L. A., Sacchi, S., & Imdorf, C. (2019). Comparing Long-term Scarring Effects of Unemployment Across Countries: the Impact of Graduating During an Economic Downturn. In B. Hvinden, J. O'Reilly, M. Schoyen, & C. Hyggen (Eds.), *Negotiating Early Job Insecurity - Well-being, Scarring and Resilience of European Youth*. Edward Elgar Publishing Ltd. https://doi.org/10.4337/9781788118798.00011

Hess, L. E., Petersen, A. C., & Mortimer, J. T. (1994). Youth, Unemployment and Marginality: The Problem and the Solution. In A. C. Petersen & J. T. Mortimer (Eds.), *Youth Unemployment and Society*. Cambridge University Press. https://doi.org/10.1017/CBO9780511664021

Hobcraft, J. N., & Sigle-Rushton, W. (2009). Identifying Patterns of Resilience Using Classification Trees. *Social Policy & Society*, *8*(1), 87–98. https://doi.org/10.1017/S1474746408004612

Hogan, D. P., & Astone, N. M. (1986). The Transition to Adulthood. *Annual Review of Sociology*, *12*, 109–130.

Holte, B. H., Swart, I., & Hiilamo, H. (2019). The NEET Concept in Comparative Youth Research: the Nordic Countries and South Africa. *Journal of Youth Studies*, *22*(2), 256–272. https://doi.org/10.1080/13676261.2018.1496406

Jahoda, M. (1981). Work, Employment, and Unemployment: Values, Theories, and Approaches in Social Research. *American Psychologist*, *36*(2), 184–191. https://doi.org/10.1037/0003-066X.36.2.184

Lallukka, T., Kerkelä, M., Ristikari, T., Merikukka, M., Hiilamo, H., Virtanen, M., Øverland, S., Gissler, M., & Halonen, J. I. (2019). Determinants of Long-Term Unemployment in Early Adulthood: A Finnish Birth Cohort Study. *SSM - Population Health*, *8*. https://doi.org/10.1016/j.ssmph.2019.100410

Levels, M., van der Velden, R., & Di Stasio, V. (2014). From School to Fitting Work: How Education-to-Job Matching of European School Leavers is Related to Educational System Characteristics. *Acta Sociologica*, *57*(4), 341–361. https://doi.org/10.1177/0001699314552807

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *The R Journal*, *6*(1), 79–89.

Mai, J. E. (2016). Big Data Privacy: The Datafication of Personal Information. *Information Society*, *32*(3), 192–199. https://doi.org/10.1080/01972243.2016.1153010

Menardi, G., & Torelli, N. (2014). Training and Assessing Classification Rules with Imbalanced Data. *Data Mining and Knowledge Discovery*, *28*, 92–122. https://doi.org/10.1007/s10618-012-0295-5

Molina, M., & Garip, F. (2019). Machine Learning for Sociology. *Annual Review of Sociology*, *45*.

Mood, C. (2010). Logistic regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, *26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Mood, C. (2017). More than Money: Social Class, Income, and the Intergenerational Persistence of Advantage. *Sociological Science*, *4*, 263–287. https://doi.org/10.15195/v4.a12

Mousteri, V., Daly, M., & Delaney, L. (2018). The Scarring Effect of Unemployment on Psychological Well-Being Across Europe. *Social Science Research*, *72*, 146–169. https://doi.org/10.1016/j.ssresearch.2018.01.007

Müller, W., & Gangl, M. (2003). *Transitions from Education to Work in Europe: The Integration of Youth into EU Labour Markets*. Oxford University Press. https://doi.org/10.1093/0199252475.001.0001

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (2nd ed.). Broadway Books.

Official Statistics of Finland. (2015). *Employment Bulletin November 2015*. https://tem.fi/en/year-2015

Paananen, R., & Gissler, M. (2012). Cohort profile: The 1987 Finnish Birth Cohort. *International Journal of Epidemiology*, *41*(4), 941–945. https://doi.org/10.1093/ije/dyr035

Rigobon, D. E., Jahani, E., Suhara, Y., AlGhoneim, K., Alghunaim, A., Pentland, A., & Almaatouq, A. (2018). *Winning Models for GPA, Grit, and Layoff in the Fragile Families Challenge*. arXiv.

Rinne, R., & Järvinen, T. (2010). The "Losers" in Education, Work and Life Chances - the Case of Finland. *Zeitschrift Für Pädagogik*, *56*(4), 512–530.

Safran, C., Bloomrosen, M., Hammond, E., Labkoff, S., Markel-Fox, S., Tang, P. C., & Detmer, D. E. (2007). Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, *14*(1), 1–9. https://doi.org/10.1197/jamia.M2273

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-ghoneim, K., Baer-bositis, L., Moritz, B., Chung, B., Eggert, W., Faletto, G., Fan, Z., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E. H., Leizman, B., Mercado-garcia, D., … Wang, Z. (2020). Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Sampson, R. J., & Laub, J. H. (1994). Urban Poverty and the Family Context of Delinquency: A New Look at Structure and Process in a Classic Study. *Child Development*, *65*, 523–540.

Sanford, M., Offord, D., McLeod, K., Boyle, M., Byrne, C., & Hall, B. (1994). Pathways into the Work Force: Antecedents of School and Work Force Status. *Journal of the American Academy of Child & Adolescent Psychiatry*, *33*(7), 1036–1046.

Shanahan, M. J. (2000). Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective. *Annual Review of Sociology*, *26*, 667–691.

Täht, K., & Reiska, E. (2016). *Institutions and the Youth Labor Market Exclusion and Insecurity in Europe : A Literature Review* (No. 4; EXCEPT Working Papers). http://www.except-project.eu/working-papers/

TE-services. (2018). *A Jobseeker's Independent Study Can Be Supported by Unemployment Benefit*. http://www.te-palvelut.fi/te/en/jobseekers/career_education_training/independent_study/index.html

Therneau, T., Atkinson, B., & Ripley, B. (2019). *rpart* (4.1-15). https://cran.r-project.org/web/packages/rpart/rpart.pdf

Vayena, E., & Blasimme, A. (2018). Health Research with Big Data: Time for Systemic Oversight. *Journal of Law, Medicine and Ethics*, *46*(1), 119–129. https://doi.org/10.1177/1073110518766026

Wadsworth, M. E., Raviv, T., Compas, B. E., & Connor-Smith, J. K. (2005). Parent and Adolescent Responses to Poverty-Related stress: Tests of Mediated and Moderated Coping Models. *Journal of Child and Family Studies*, *14*(2), 283–298. https://doi.org/10.1007/s10826-005-5056-2

Westerinen, H. (2018). *Prevalence of Intellectual Disability in Finland* [University of Helsinki]. https://doi.org/10.1080/13668259700033311

Wolbers, M. H. J. (2003). Learning and Working: Double Statuses in Youth Transitions. In W. Müller & M. Gangl (Eds.), *Transitions from Education to Work in Europe: The Integration of Youth into EU Labour Markets*. Oxford University Press. https://doi.org/10.1093/0199252475.001.0001

World Health Organization. (2020). *Classification of Diseases*. https://www.who.int/classifications/icd/icdonlineversions/en/