Statistical Learning for the Prediction of School Dropouts

J.H.C. Bunk
Utrecht University

Supervisors:
dr. A.J. Feelders, Utrecht University
prof. dr. J.T. Jeuring, Utrecht University
B.J. Derlagen, MSc, Topicus
R. Heerkens, MSc, Topicus

June 23, 2016

ABSTRACT

Students leaving school without a basic qualification face numerous disadvantages compared to their graduated peers. Therefore, Educational systems in the Netherlands strive to give all students at least a basic qualification. They try to achieve this, among other things, by dropout prevention programs. Although the number of dropouts in the Netherlands is reduced, there are still a large number of dropouts. A significant number of them drop out unexpectedly and, therefore, without intervention of the dropout prevention programs. In this research project, we explore the use of data mining techniques to help identifying the students at risk of dropping out. We will show that data mining has great potential to help schools in this task.

1

INTRODUCTION

Students leaving school without a basic qualification face numerous disadvantages compared to their graduated peers. For instance, research on Dutch dropouts shows that they are more likely to be unemployed [28], are unhealthier [17] and are more likely to exhibit criminal behaviour [28]. Dropping out brings a personal cost to the student, but also a cost to society. The Dutch government subsidises health care, provides support to the unemployed and invests in the prevention of crime.

Unsurprisingly, the EU wants to reduce the student dropout rate. The goal is to reduce it to below 10% before 2020. The Netherlands formulated stronger targets. Vocational education (in Dutch: Middelbaar BeroepsOnderwijs (MBO)) is the education type with the highest percentage of dropouts in the Netherlands. The target is to reduce this percentage to 5% by 2016. Current dropout prevention programs have already successfully reduced the number to 5.2%, but some schools still have a dropout rate of 7% or higher [28].

Early identification of the potential dropouts can contribute to dropout rate reduction, as the potential dropouts can be included in the dropout prevention programs that the schools already have organised. This research project is going to create a model to predict potential dropouts to help in this identification process. The model uses the information that schools have stored in their student information system. In the creation of the model we will pay attention to the dropout indicators found in previous research. They provide us with a strong basis for identifying important factors in our data set. Not all indicators from the literature can be mapped one to one onto the data set. Therefore, the model will include variables that have an indirect relation to these indicators. Furthermore, we included some variables which are not labelled as dropout indicators by previous research. This creates the possibility of finding a new dropout indicator. In the end, we hope that the resulting model will predict dropouts with high accuracy and will help future research by providing the relative importance of the indicators.

Previous research with data mining/machine learning techniques to classify dropouts comes in different flavors. They differ in their definition of dropouts, type of students, type of education, sample size (from fewer than 50 to 100.000), type of technique used and risk factors analysed. For instance, neural networks [25], Classification and Regression Tree (CART) and CHAID analysis [26, 42], Naive Bayes [20] and K-means [11] have been used in previous research to classify dropouts. We see for this research project as an important aspect of the models that they are understandable by the user. Therefore, we avoided "blackbox"-like models such as neural networks.

1.1 EARLY SCHOOL LEAVER

The interest of this study is in the early school leavers (in Dutch: voortijdig schoolverlater) as defined by the Ministry of Education, Culture and Science of the Netherlands. To improve the readability of this paper the term dropout is used exchangeably for early school leaver. We define early school leaver and dropout as:

Definition 1 *Early school leavers are students younger than 23 who stop with their current study without starting with a new study in the next school year and without having achieved a basic qualification.*

The basic qualification can be: a finished secondary vocational education (MBO level 2 or higher), senior general secondary education (HAVO) or pre-university education (VWO).

1.2 WHAT NEXT?

The remainder of this thesis proposal starts by formulating our research questions and our expected result. This is followed by a description of known indicators of student dropout found in previous research. The other chapters discuss the data set, the techniques used, the results, an interpretation of the results and we close off with a discussion and conclusion.

2

METHOD AND HYPOTHESIS

At the moment no automated detection system for potential dropouts is used in the Netherlands. Schools mostly focus on one or two dropout indicators (e.g. absence and (bad) behaviour) to decide which student gets more supervision. We wonder if an automated system can help schools with these decisions. To test this, we formulate our main research question as:

Question 1 *Is it possible for an automated system to detect potential dropouts?*

All schools in the Netherlands use student information systems for their administration. We predict, that these systems provide enough information for automated detection of potential dropouts. Therefore, we specify our first question to:

Question 2 *Do current student information systems provide enough information for an automated system to predict dropouts?*

In particular, we hope that automated systems will contribute to the early detection of unexpected dropouts; early school leavers who did not have poor academic performance before dropping out [16]. To see if this is the case, we try to find an answer to the third question:

Question 3 *Do current student information systems provide enough information for an automated system to predict unexpected dropouts?*

To find answers to these questions, we use the data from EduArte, a student information system created by Topicus. With the data we create multiple dropout prediction models. After the creation of models, we evaluate them and look at the dropout indicators used. In the discussion we look at our questions and the evaluation of our models and try to extract the answers to the questions.

We believe that it is of importance that the users (school staff) can get an understanding of why a certain student has a higher dropout risk, as this can contribute to the dropout prevention and might reduce the negative effects of labelling a student as a potential dropout. We discuss these negative effects in more detail in section 9.1. From this belief we strive for an automated system with as basis a model that is comprehensible for the school staff also.

3

KNOWN INDICATORS

We are limited to the information provided by the student information system EduArte. Still, it is important to have knowledge of previous found dropout indicators. This will help in our exploration of the data set, in selecting the right predictors from the data set and choosing the right prediction method. In this chapter, we describe in each section, expect the last, a category of dropout indicators found in previous research. In the last section, we describe the indicators used by and results of previous research on the prediction of school dropouts and point out the differences and similarities with our research.

3.1 SCHOOL CAREER

Prior education and current position in the students school career can provide information on the dropout risk of a student [28, 36, 10, 7, 39, 23]. This is not limited to their highest qualification and their study progress, but also includes education type, number of unfinished studies, student achievement (e.g. grades), engagement (e.g. absence), grade retention, number of contact hours, being overaged (being one or more years older than the average of the class) and mobility (The number of switches between schools). Some researchers argue that the most accurate indicators are longitudinal trajectories of student achievement (e.g. grades) or engagement in school (e.g. absence) [5]. Others see engagement as one of the most important indicators [36, 41, 10, 16]. Different research use different variables as measure of engagement. The most used measure is the absence of the student, but participating in extracurricular activities, motivation and number of friends on the same school are used as well.

3.2 DEMOGRAPHICS

Demographics such as age, gender, nationality, parent's nationality, family composition, parent's socioeconomic status and neighbourhood of a student has been found to be dropout indicators [28, 36, 10, 19, 38, 39]. Socioeconomic status, in particular, is seen as an important dropout indicator [19, 38] that can explain effects seen in other indicators, such as nationality [36].

3.3 SCHOOL STRUCTURE

Schools differ in many aspects. This influences students' development, but also the type of students that are attracted to the school. Multiple studies show that aspects of the school, such as composition of students, number of students, number of teachers, schools' cost per student, education level of teachers, social climate and the location of the school (urban or countryside), have influence on the dropout of students [14, 6, 36, 39].

3.4 BEHAVIOUR

Student's behaviour can be hard to measure. Still, research has found some behaviours and/or events can be possible dropout indicators. These are bad behaviour, bad relationship with school personnel, being bullied, being sick often, criminal behaviour, working part-time, using drugs and having personal problems at home [36, 2, 28, 1, 39].

3.5 PREDICTION OF SCHOOL DROPOUTS

The dropout definition used in this research project (see section 1.1) is specific for the Netherlands. Although the characteristics of early school leavers have been researched (for instance [39]), to the authors' knowledge no previous research has been done on predicting early school leavers using the definition as provided by the Ministry of Education, Culture and Science of the Netherlands. The definition of dropout used in most papers is: leaving the current study or course without successfully finishing it.

This is not the only facet in which previous research varies. The main components that previous research differs in are the type of students, type/level of education, sample size (from fewer than 50 to more than 100.000), type of technique used for prediction and indicators included in the data set. Despite the differences, there is also common ground.

Predicting dropouts can be done by using a large number of techniques. Not all techniques satisfy our prerequisite that the users can get an understanding of why a certain student has a higher dropout risk. This is not to say that these techniques do not work. For instance, [25] shows that neural networks can be trained to classify dropouts with high sensitivity and accuracy. Other "black box" models, such as random forest, have also been successfully tested on education data sets [9].

In our research project we look at more easily understandable models. Still, not all "white box" models are appropriate. [11] shows with a (small) educational data set that the k-means approach could work to cluster students in categories. Although the decision process is clear in the k-means algorithm, it is hard to distinguish the real risk predictors. In this research project, we will use statistical models and classification trees to predict early school leavers. They both provide a clear decision process in which the effect of the different predictors is visible.

Decision trees have been successfully implemented for data sets in varies fields. In the field of education, they have already been used to predict high-school dropouts [24] and dropouts at a distance-learning institute [21]. The first study shows that decision trees can have a high accuracy when predicting dropouts. The second study has a less positive result, but acknowledges the limits of the data set used.

In educational research, statistical methods are mostly used to map the effects of possible predictors on early school leaving, for instance in Norway [23]. They discuss the resulting model and its predictors, but do not test the resulting model on a test set. We could not find statistical models that were tested on their prediction qualities. Therefore, it seems as if most research with educational data sets, only use statistical methods for data analysis and do not test the resulting model on their prediction abilities. This does not mean there is no potential in using statistical models for the prediction of early school leavers. Statistical models have been proven in other fields to have great prediction abilities.

4

DESCRIPTION OF THE DATA SET

4.1 INTRODUCTION

In this research project we use data from EduArte, a student information system for vocational education created by Topicus. Although the storage of the data is managed by Topicus, the schools hold the rights to the data. Two schools gave us permission to use their data. Not all dropout indicators discussed in the previous chapter can be mapped one to one onto the data set or can be found in the data set at all. In this chapter, we describe the content of the data set and the challenges we faced in collecting and preparing the data.

4.2 DATA SELECTION

The data of one school is stored on an Oracle server and for the other school on a Microsoft SQL server. The databases are designed for administrative purposes and are filled in by hand by the school staff. We wrote queries to access the required information and transform the data in the right form for the analysis.

The definition of an early school leaver (see section 1.1) tells us that a student older than 23 cannot be an early school leaver. Therefore, all students older than 23 at the start of their most recent study were removed.

4.3 DATA SET AND PREPROCESSING

In this section we explain and describe the content of the data set and give the descriptive statistics of some predictors. We make a separation between two types of predictors. First, we discuss the basic information, which contains the information the schools knows at the start of a student's study. Secondly, we describe the trajectory information. This is information gained while a student progresses his/her study (e.g. grades).

4.3.1 *Basic information*

The data set with the basic information contains 25439 students, 3502 of whom are early school leavers. The students in this set started their study between 2009-07-01 and 2014-07-01. The basic information data set with the students who continued their study after one year consists of 14336 students, 1453 of whom are early school leavers.

Age

The date of birth of all subjects is known. We use this to create the following variables: age at the start of the study (m: 18.80; std: 1.81), age when graduating (assuming that the student will have no study delay)(m: 21.59; std: 1.92), years until compulsory education ends (m: 0.40, std: 0.62), years lived in the Netherlands at the start of the study (m: 18.412; std: 2.92) and at which age the subject came to the Netherlands (zero for natives)(m: 0.39; std: 2.38). Table 1 gives the means and standard deviations of these predictors with the data set split between dropouts and non-dropouts.

Early School Leaver	Yes		No	
	mean	std	mean	std
Age at the start of the study	18.45	1.48	18.85	1.85
Age when graduating	20.91	1.62	21.70	1.95
Years until compulsory education ends	0.38	0.54	0.40	0.64
Years lived in the Netherlands at the start of the study	17.80	3.28	18.505	2.85
Age in the Netherlands	0.65	3.02	0.35	2.26

Table 1: The mean and standard deviations of the age predictors split between dropouts and non-dropouts. N=25439

Gender

There are 13247 males (16.43% dropout) and 12192 females (10.88% dropout) in the data set.

Nationality

The nationality and country of birth are grouped in the following categories: Netherlands, Europe (the continent), European Union and western countries[1]. The differences in the number of subjects between the last three categories is minimal. The number of students and the

[1] Western countries are, as defined by the Dutch Central Bureau of Statistics (CBS), countries inside Europe (excluding Turkey), North-America and Oceania, or Indonesia or Japan

nationality	number of students	percentage dropouts
Dutch	23534	13.09%
non Dutch	1905	22.15%
non European Union	1496	22.33%
non Europe	1423	22.21%
non-western	1410	22.13%

Table 2: The number of students and the percentage dropouts for each of the nationality variables.

country of birth	number of students	percentage dropouts
the Netherlands	23534	13.29%
outside the Netherlands	1905	19.63%
outside the European Union	1491	20.05%
outside Europe	1325	20.53%
non-western country	1314	20.40%

Table 3: The number of students and the percentage dropouts for each of the country of birth variables.

percentage dropouts for the nationality categories is listed in table 2. The same information for country of birth can be found in table 3.

Socioeconomic status

Socioeconomic status is not available in the data set. To get an estimated socioeconomic status, we tried to use data about the neighbourhoods from the Dutch *Centraal Bureau voor de Statistiek* (Central Agency for Statistics). The provided information was not complete enough and is therefore not included in the data set.

Education

We extract the most recent study of students from the student information system. The subjects in the data set are following different educational programs. The number of different studies is so large and the number of participants of some studies so small that taking the study itself as predictor in the model is not interesting. Instead, we use the different aspects of the educational programs. The educational programs differ in their level (1 to 4), type of education (combined with a job (BBL)(14.99% dropouts) or not (BOL)(13.34% dropouts)), intensity (full-time (13.78% dropouts), part-time (15.11% dropouts) or exam participant (8.85% dropouts)) and job sector. For the latter we used *kenniscentrum*, which sorts the educational programs into groups based on the job sector the student is educated to work in. There are 17 different job sectors. They vary in their number of students from 23 to 7270 and dropout percentage from 23.58% to 4.62%.

School structure

The data set we use consists of data from only two schools. Therefore, we do not think it is interesting to take school structure into account.

Previous education

The data set provides either the highest qualification or the highest level of enjoyed education (without finishing it) of a student. We create with this information the variable *qualification*, a Boolean indicating whether the subject has a qualification, and the variable *previous education* with their highest level of enjoyed education (whether or not they got their qualification). These levels can be found in table 4. Furthermore, we include in the data set how many vocational studies students started at their current school before their current educational program. The average previous vocational studies is almost the same for non-dropouts (mean: 1.28; std:1.42) and dropouts (mean:1.26; std: 1.46).

	number of students	percentage dropouts
None	103	15.53%
Basisonderwijs	861	22.18%
Basisvorming	4476	23.82%
Havo	601	0.17%
HBO	5	0.00%
MBO level 1 or 2	881	14.07%
MBO level 3 or 4	992	1.11%
Propedeuse HBO	2	0.00%
Vmbo	11031	14.20%
Vmbo-tl	6456	8.07%
VWO	15	20.00%
Unknown	11	27.27%

Table 4: The number of students and dropout percentage for each previous education level in the data set.

Learning disorders

The database contains information on which students have (learning) disorders that can effect the education or the interaction with the student. The different disorders are listed in table 5. We used three variables to group students with a disorder. The first is *number of disorders* (mean: 0.06; std: 0.25;), which contains the number of disorders the student has (there are a few students who have multiple disorders). The other two are *number of disorders as a warning* (mean: 0.03; std: 0.18) and *number of disorders on the student card* (mean 0.05; std: 0.22). These categories were provided by the database. The reason for a disorder to be on the student card is for the student to show he has a certain disorder. This way he or she can show, for instance, that he or she is entitled to more time for his or her exam. When a disorder is in the category of *disorders as a warning*, then the teacher get a heads up when such a student enrol in their class. This allows the teacher to prepare himself.

	Number of students	Percentage dropouts
ADHD	42	35.71%
ADD	8	12.50%
PDD-NOS	51	31.37%
Autism	7	28.57%
Asperger syndrome	17	11.76%
Dyslexia	732	5.87%
Nonverbal learning disorder	3	33.33%
Dyscalculia	22	13.64%
Having fears	6	33.33%
ASS	2	0.00%
Behavioural problems	4	50.00%
Handicap	191	15.71%
ODD	13	38.46%

Table 5: The number of students with a learning disorder and the percentage of those students who became a dropout.

4.3.2 *Trajectory information*

The trajectory information contains the information about the progress of students in their educational program. We believe that using this kind of information will show the real power of automated systems for the prediction of early school leavers. Unfortunately, the schools that gave us permission to work with their data only recently started to use all functions of EduArte. Therefore, most of the trajectory information is not available to us. The number of students in the data set is reduced immensely when we only use students with trajectory information. Still, we use the little data that we have to make a first exploration into the possibilities the trajectory information provides. Note that no real conclusions can be drawn from this exploration, due to the relatively small sample size and the poor quality of the data. The remainder of this section describes the trajectory information that is available in the data set.

Years in educational program

To attain the trajectory information we initially looked at students one year, two years and three years into their study. Even without removing students of which we do not have trajectory information the data set is reduced by each year, because students discontinue their study. There are only 14336 students who continue their study after one year. This reduced even further to 6722 students after two years and to 4104 students after three years.

Grades

Grades are an interesting part of the data set. They can tell more than only the performance of the student. Unfortunately, we do not have enough historical data to create a model that incorporates the full potential grade information has to offer. Of the 14336 students that remain after one year, only 3616 students (including 340 dropouts) have one or more grades. Furthermore, the grade information that is available is of questionable quality. The grade administration function works great for schools, but it is not designed for the purpose we have. The dates of the grades are not always reliable, due to the conversion of schools to different systems. We tried to work around this, but only with little success. The number of grades the 3616 students have after one year range from 1 to 360 (mean: 32.65; std: 43.54).

Besides these problems, one has to consider the fact that these students are in different educational programs. Not only do they get different exams, the grades also may be a grade of something else than an exam. Depending on the educational program, students have to write essays, do an internship or make practical assignments. Representing all these features of the data is not a trivial task. For the exploration of the effects of grades, we only consider the average and the standard deviation of the grades of the students. To measure all grades on the same scale, we divide every grade by the maximum grade the student could score. The mean of the average first year grades of dropouts is 0.55 (std: 0.19) and that of non-dropouts is 0.65 (std: 0.16).

Documents

Teachers have the most knowledge about the students. Some of this knowledge is recorded in the variables mentioned in the other sections, but the expertise of the teacher will not be included by only looking at these variables. Teachers write multiple documents about the student's development and summaries of the meetings with the student or the student's parent(s). This research project aimed to include some of the expertise of the teacher by adding statistical information (e.g. word counts) about these documents to the model. The occurrences of words related to indicators such as bad behaviour, relationship with school personnel, being bullied and the health of the student may provide an estimate of the real indicator. Also, by including these documents it might be possible to capture unexpected events that severely affect a student (e.g. being arrested, using drugs or having serious personal problems at home). This might allow the model to identify "unexpected dropouts"; early school leavers who did not have poor academic performance before dropping out [16]. Unfortunately, not enough documents are available to be able to use them in the model creation process. The option of EduArte to upload these documents is used more and more by the schools. We recommend future research with access to these documents to include them in their data set.

Engagement

The data set contains some indirect information about the engagement: absence/truancy and the number of times late for class. Absence has already been identified as a strong indicator of dropout in American data sets [16]. The schools only recently started to use EduArte to register the absence of students. Because the system was earlier available to them, it is hard to pin-point when they made a full transition to the EduArte-system. Therefore, we only add the absence predictor to the data set with the 3616 students whom had 1 or more first year grades. On average non-dropouts have 10.52 absence notifications (std: 15.76) in their first year. Dropouts have on average 9.48 absence notifications (std: 17.05). The notifications know 11 categories: hospital visit, personal circumstances, appointment, sick, late for class, leave, suspended, internship, specialist, school activity and doctor.

4.4 PERMISSION SCHOOLS

Although the storage of the data is managed by Topicus, the schools hold the rights to the data. To sense the attitude towards our research we contacted a few schools before asking all schools to participate. We sent them an explanation (in Dutch) of the research project and visited the schools. Although we covered this in our explanation[2], we noticed that the schools had a (healthy) concern about the privacy of their students. Speaking in person with the school reassured them we would handle their data with care.

After we improved the explanation of the research, we felt confident to contact the remaining schools. This was not done directly by Topicus, but trough Educus, the company that sells EduArt to the schools. The communication was therefore slower, than would we

[2] We adhere to the Dutch code of conduct for research and statistics (in Dutch: 'Gedragscode voor Onderzoek & Statistiek') as approved by the Dutch Data Protection Authority (in Dutch: 'College Bescherming Persoonsgegevens')

have had direct contact. Some schools did not react to the communication and most schools showed no interest in participating. They provided reasons we could and did refute, but it did not convince them to participate.

One school that we contacted at the beginning was enthusiastic to cooperate at the start, but pulled out in the end. They saw the legal paperwork we provided not as sufficient to protect the privacy of their students. They offered their help in setting up the legal paperwork that would meet their criteria, but only wanted to invest their time if we increased the number of participating schools by contacting the schools trough the MBO-raad, the Netherlands Association of vocational education and training and adult education Colleges. Although we probably would have received more responses to our research request if we would have contacted the MBO-raad, we were running out of time and decided not to do it. Still, we recommend future research into early school leavers in the Netherlands to definitely contact the MBO-raad.

In short the whole process from contacting the school to signing the legal paperwork took several months and resulted in fewer schools than we expected. Furthermore, the two schools who gave us permission did not use all features of EduArte or just recently started using these features. This reduced the dimensions of the data (for instance, the grades) and the quantity even further.

5

CONSTRUCTION OF THE CLASSIFICATION MODELS

5.1 INTRODUCTION

Our goal is to construct a classification model that not only detects potential dropouts, but is also comprehensible and easily interpretable. To facilitate this, we use a logistic regression model. Logistic models are clear in how strongly the different factors influence the classification process. Although this is an important feature of making a comprehensible model, it is not enough. One can quite easily lose the overview when there are dozens of different parameters. To take this into account, to avoid overfitting and to act according to the principles of Occams Razor, we use two heuristics that are designed to make a good trade-off between the performance and complexity of the logistic model.

Logistic regression models are not the only type of models that can facilitate our goals. The if-then structure of classification trees also makes it possible to understand the classification process and see the relevance of the parameters. Creating a different type of model might show different regularities and allows us to compare the performances. In this chapter, we explain logistic regression, the two heuristics used to reduce the complexity of the logistic regression models and C5.0, the algorithm used to create the classification trees.

5.2 LOGISTIC REGRESSION MODEL

Fitting a logistic regression model [8] to the data creates a coefficient vector b that is used in the following (logistic) formula:

$$\widehat{P}(y = 1|x) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \cdots + b_p x_p)}}$$

for p predictors where b_p is the coefficient of predictor x_p, b_0 is known as the intercept and $\widehat{P}(y = 1|x)$ is the estimated probability that $y = 1$ given predictors x. The logistic formula provides us with a curve for which $\widehat{P}(y = 1|x)$ depends on x and is always between 0 and 1. The latter is a desired property, as the true value of y is always one or zero: The student is a dropout or he is not. Furthermore, this property allows us to see $\widehat{P}(y = 1|x)$ as the probability that the student is a dropout. The estimated value $\widehat{P}(y = 1|x)$ will tell how strong the belief is that a sample belongs to one of the two categories. Fitting is done by maximising the likelihood:

$$L(b|y) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

where y_i is the true value of observation i and $p(x_i)$ is $\widehat{P}(y = 1|x)$. To avoid overfitting and to reduce complexity we want b to have as little nonzero values as possible. To find the logistic regression model with the right performance - complexity trade-off we use stepwise regression and Lasso.

5.2.1 *Stepwise regression*

Stepwise regression, also known as subset selection, is a method to reduce the complexity of a (logistic) regression model. Its aim is to find the model with the lowest Akaike information criterion (AIC) or, as it is in our case, the lowest Bayesian information criterion (BIC). These criteria are a measure of the relative quality of the models and include in their score a punishment for the complexity of the model. Where BIC, given a large enough number of samples, penalises the complexity more heavily than AIC. The definitions of AIC and BIC are:

$$AIC = 2j - 2\ln(L)$$

$$BIC = j\ln(n) - 2\ln(L)$$

where j is the number of predictors, n is the number of samples and L is the maximum likelihood. A lower score on BIC and AIC means a better trade-off between complexity and performance. Finding the model with the lowest BIC or AIC score is NP-hard [27]. Meaning that it is impossible to find the best model in a reasonable amount of time. As a solution, stepwise regression uses local search to find a good (locally optimal) solution.

Stepwise regression starts with the null model. In every search step all neighbouring models, the models that have the same predictors and one extra or have the same predictors except one, are fitted. The neighbouring model with the lowest BIC will be the start of the next search step. The search steps are repeated until we come across a model for which none of the neighbouring models have a lower BIC. This model is the good (locally optimal) solution. The pseudo code of stepwise regression can be found in algorithm 1.

The great feature of stepwise regression in combination with BIC or AIC is that it reduces the complexity of the model by excluding

Algorithm 1 Stepwise regression

```
Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
$i = 0$
while true do
  for every predictor $x$ do
    if predictor is in $\mathcal{M}_i$ then
      Fit the model with all predictors from $\mathcal{M}_i$ except predictor $x$
    else
      Fit the model with all predictors from $\mathcal{M}_i$ and predictor $x$
    end if
    Call the fitted models $\mathcal{M}_{i,x}$
  end for
  Pick the best $\mathcal{M}_{i,x}$ model and call it $\mathcal{M}_{i+1}$. Here, best is defined as having the smallest BIC score
  if BIC_score($\mathcal{M}_i$) < BIC_score($\mathcal{M}_{i+1}$) then
    return $\mathcal{M}_i$
  end if
  $i$ += 1
end while
```

predictors. Still, with a large number of predictors (and their interactions), it can take a long time to find a good solution. Furthermore, as it is a non-continuous process, all the steps made in the search to the solution are quite large (a predictor is included or excluded).

To reduce the running time and to partially incorporate a continuous search process, we first use logistic regression with lasso to select the variables (without interactions) and use stepwise regression with forward and backward selection on the reduced data set (with interactions).

5.2.2 *Logistic regression with lasso*

Lasso (least absolute shrinkage and selection operator) [37] is a method for subset selection that combines favourable features from local search and ridge regression [18]. Just like ridge regression it shrinks the coefficients towards zero, but unlike ridge regression it can also set some coefficients to zero. This makes lasso a stable continuous process that results in a more interpretable model. Lasso is done by fitting the logistic regression model by maximising the joint log-likelihood minus a shrinkage penalty:

$$\ln(L(b|y)) - \lambda \sum_{j=0}^{p} |b_j|$$

where p is the number of predictors, b_0 is known as the intercept, b_j is the coefficient of the jth predictor and $\lambda \geq 0$ is known as the tuning parameter. When the tuning parameter increases, more coefficients will shrink towards zero.

The original lasso algorithm cannot group variables and cannot work with categorical variables. This means that categorical variables need to be transformed to dummy variables; for each category a binary variable is created which indicates whether or not the sample belongs to that category. As these dummy variables cannot be grouped, it is possible that the lasso algorithm selects only a few categories of a categorical variable. This is not always a desired property, as in some practises either a categorical variable is included as whole (all categories) or is excluded as a whole (e.g. we either take into account the previous education of a student or we do not).

Later, grouped lasso, an extension, was proposed [3, 43] to allow for grouping the variables. The extended algorithm still cannot work with categorical variables, but with the extension, the algorithm can group the dummy variables. With the grouping the algorithm selects the important groups rather than the important members within those groups. The members of a group either all have non-zero coefficients or all have a zero coefficient. We explore the use of the original lasso, which we also refer to as ungrouped lasso, and of three penalty functions for grouped lasso: standard, SCAD [12, 13] and minimax concave penalty (MCP) [44].

To find the right value for the tuning parameter lasso was run on the training set with a 10-fold cross validation. The simplest model with the tuning parameter one standard deviation away from the highest AUC (see section 5.5.1) was selected. The selected predictors (the predictors with a nonzero coefficient in the model) and their interactions are the predictors considered in the stepwise regression process.

5.3 C5.0

The C5.0 algorithm [33] is used to create a classification tree. The algorithm is an extended version of the C4.5 algorithm [31], which is in turn an extension of the ID3 algorithm [32]. The details of the extensions of C5.0 are largely undocumented, but seen as an improvement [34, 22]. We start by explaining the C4.5 algorithm, as this is the basis of C5.0. The sections that follow will explain the extensions of C5.0.

5.3.1 *C4.5*

Classification trees try to use the predictors to divide the data set into smaller partitions with a larger proportion of one class. These become the leaves of the decision tree. A new observation will move down the tree to a leaf driven by its own features. The prediction of the new observation's class is based on the majority of samples in the leaf. A leaf with a larger proportion of one class provides more certainty for the prediction. The algorithm C4.5 creates the smaller partitions based on the information statistic (entropy). The information statistic shows how much information is gained when the true class of an observation is revealed. In other words, it tells us how uncertain we are of the true class of a new observation. It decides on where and which split to make based on the information gain (reduced uncertainty) of the potential new leaves. When there are two classes, as in our case, then the information statistic is high for a leaf where the two classes are balanced, and low where they are unbalanced. More formal:

$$info = -(p \log_2 p + (1-p)\log_2(1-p))$$

Where p is the probability of the first class and when $p = 0$ or $p = 1$ it is customary to have $0 \log_2 0 = 0$. When p is close to 0 or 1 (it is unbalanced), then the information statistic is small and when p is close to 0.5 it is high. In C4.5, the value of p is the proportion of the class in the leaf. This gives us the following formula:

$$info(\text{leaf}) = -\left(\frac{n_1}{n}\log_2(\frac{n_1}{n}) + \frac{n_2}{n}\log_2(\frac{n_2}{n})\right)$$

where n are the number samples in the leaf, n_1 are the number of samples of the first class and n_2 the number of samples of the second class. Note that as we only have two classes $1 - \frac{n_1}{n}$ is equal to $\frac{n_2}{n}$.

The information after the split is calculated in a similar way, but weighed by the number of samples in each of the new leaves. With two new leaves the information statistic after the split will be:

$$info(\text{after split}) = \frac{n_{leaf\_1}}{n_{total}} info(\text{leaf\_1}) + \frac{n_{leaf\_2}}{n_{total}} info(\text{leaf\_2})$$

This formula can be generalised to splits with l leaves:

$$info(\text{after split}) = \sum_{i=1}^{l}\left(\frac{n_{leaf\_i}}{n_{total}} info(\text{leaf\_i})\right)$$

With these formulae the algorithm can calculate if there are splits that lead to less uncertainty and, if so, how much information they give. Furthermore, these formulae allows us to calculate the information gain of a split by simply subtracting the information after the split from the information before the split:

$$gain(split) = info(\text{before split}) - info(\text{after split})$$

A strategy to build the tree could be to go over all possible splits and expand the tree with the split that has the most gain in information. However, this tactic will be strongly biased towards multi-valued categorical variables, as multi-way splits are likely to have more information gain [22] and the splits of the numeric variables considered are always binary splits. To overcome this bias, the information gain ratio is used. This is calculated by dividing the information gain by its intrinsic value, the entropy of distribution of instances into branches or in other words the amount of information we need to tell which branch an instance belongs to.

$$intrinsic\_value(split) = -\sum_{i=1}^{l} \left(\frac{n_{leaf_i}}{n_{total}} \log_2 \left(\frac{n_{leaf_i}}{n_{total}} \right) \right)$$

$$gain\_ratio(split) = \frac{gain(split)}{intrinsic\_value(split)}$$

Using the gain ratio the algorithm searches for the best split. This process is repeated until there is no split left that will improve the tree, for which all leaves have more samples than the set minimum amount. The resulting tree is large and likely to over-fit the data. The next step in C4.5 is to prune the tree.

The pruning phase consists of eliminating sub-trees and replacing nodes by raising sub-trees. To decide if and what to prune C4.5 makes an estimation of the error rate. The estimation is made by calculating the error rate on the training samples and taking the upper limit of the error's confidence interval. The pruning action is based on the estimation of the error rate with and without the node. When the tree without the node has a lower error rate estimation, then the node is pruned. The upper limit of the error's confidence interval can be reduced by increasing the confidence factor, a parameter of C4.5. A lower confidence factor results in more leaves being pruned. The default value of the confidence factor is 0.25.

5.3.2 *Extensions of C5.0*

There is little literature about the extensions of C5.0. In 2011 the source code became available to the public. The description here is based on the evaluation of the source code by [22]. That said, the main improvements seem to be:

- a cost-complexity approach to the pruning phase.
- a boosting procedure.

Cost-complexity pruning

C5.0 adds a final global cost-complexity pruning procedure. After the pruning as described in section 5.3.1, sub-trees are removed until the error rate is one standard deviation away from the tree without pruning.

Boosting

C5.0 has a boosting procedure similar to ADABoost [15]. In ADABoost, multiple, but weaker, classifiers are trained. After the generation of a classifier the weights of incorrectly classified training samples are increased and the weights of the correctly classified training samples are reduced. The next classifier is trained with these weights on the samples. This process creates multiple classifiers that each focus on a different region of the data set. During the classification process, all the trained classifiers are used and their influence on the prediction is based on their training set error rate.

The boosting procedure in C5.0, just like in ADABoost, builds multiple classifiers with weighted samples. Unlike ADABoost, it restricts the size of the classifiers to be the same as the first one and all the classifiers have the same influence on the classification process. It also adds two evaluations to the process. The first evaluation is the effectiveness of the current model. If this is very high or very low the boosting procedure will stop automatically. The other evaluation is only performed once at half of the number of boosting iterations and checks if it is even possible to correctly classify all training samples. If not, then the not classifiable training samples are removed.

5.3.3 *Parameter tuning*

We have three parameters to tune for the C5.0 algorithm: the number of trials for the boosting procedure, the maximum number of splits and the confidence factor. The first classification tree that we train will not use boosting, as boosted trees are less comprehensible. To see the difference in performance and the full potential of C5.0, we train a second classification tree with boosting. The parameters for the C5.0 algorithm are tuned separately for both models. To tune these parameters we use 10-fold cross validation and select the parameters with the highest average AUC (see section 5.5.1). During each fold of the 10-fold cross validation $\frac{9}{10}$th of the total training set is used to train the model. This is why we use the maximum number of splits as tuning parameter and not the minimum number of samples in a leaf. The minimum number of samples in a leaf is calculated by dividing the number of samples used during the training by the maximum number of splits:

$$\text{minimum number of samples in a leaf} = \frac{\text{number of samples in set}}{\text{maximum number of splits}}$$

where *maximum number of splits* denotes the maximum number of possible splits and is used as the tuning parameter instead of the minimum number samples in a leaf.

5.4 DATA IMPUTATION

A few samples in the basic information data set had missing values for some predictors. In total 87 samples had a missing value for either kenniscentrum (36), previous education qualification (11), highest level enjoyed previous education (11), intensity (66), study level (32) or age at end of study (9). In the data set with the trajectory information we have 3 samples with missing values for either intensity (2), previous education qualification (1), highest level enjoyed previous education (1), age at end of study (1). Logistic regression models do not work with missing values. To still be able to use these samples we use data imputation to fill the missing values.

One has to be careful with data imputation. The imputed values are an (educated) guess of the true value, but the classifiers will process them as observed values. When working with a data set with a large number of missing values it is good practice to investigate the effects of (different) imputation(s) on the data set and classification. Luckily, we have a relatively small amount of missing values. The imputation has therefore no significant impact on the classification, besides the positive effect that we do not have the remove the 87 students and we can keep the population complete.

For the imputation process we use the R package MICE [40] for Multivariate Imputations by Chained Equations to impute missing values. MICE uses Gibbs sampling over the set conditional distributions of the missing values. A data set with imputed missing values is returned after a few, in our case 5, iterations. MICE offers the possibility to create multiple data sets so the effects of different draws from the distributions can be investigated. As this not necessary for us, we only create one data set with imputed values and use it to create a training and test set.

5.5 TRAINING AND TESTING PROCEDURE

We train the models on two different data sets. The first only contains the information that is available at the start of a study (The variables described in section 4.3.1) and contains 25439 students. The second data set contains the basic information as well as the trajectory information. The second data contains 3616 students.

To be able to train and test the models we randomly divide the data sets into a training set, with 75% of the samples, and a test set, with 25% of the samples. During the training of a model, the model does not see samples from the test set. The test set provides, therefore, a set which we can use to test the performance of the model.

5.5.1 *ROC and AUC*

The number of samples of each class is unbalanced in the data sets. The data set with only the basic information has 3502 students that are early school leavers of the in total 25439 students. In the data set with the trajectory information there are 340 dropouts in the total set of 3616 students. If we would use the misclassification rate as performance measure for the models, than a model that would classify all students as non-dropouts would seem to perform quite well; only 3502 would be misclassified in the first data set, which gives an error rate of only 13.80%. As this is not the kind of performance we want of the model, we use the receiver operating characteristic (ROC) curve and the area under the curve (AUC) as performance metric. Furthermore, this provides a better way to compare our results with the results of other studies [5].

Deciding upon the decision threshold of the models is mostly a trade-off between the false-positive rate, the percentage of non-dropouts classified as dropouts, and the true positive rate, the number of dropouts correctly classified as dropouts. This trade-off is illustrated with the ROC curve. The AUC is the area under the ROC curve. With the AUC we can speak about the performance of the model without specifying the decision threshold. One way to understand the AUC is as the chance that when we take a random dropout and a random non-dropout that, according to the model, the dropout has a higher estimate probability that he is a dropout than the non-dropout. The AUC is a value between 0 and 1, where a value closer to one indicates better performance, a value of 0.5 represents random classification and everything below 0.5 indicates a performance worse than random.

6

RESULTS; BASIC INFORMATION

Using the data set with the basic information, We constructed classification models as described in chapter 5. In this chapter, we describe the resulting models and their performance on the test set.

6.1 LOGISTIC REGRESSION

6.1.1 *Grouped Lasso*

We have the choice out of three different penalty functions, standard, SCAD and MCP, for the lasso algorithm. We created a logistic regression model with each of them. The penalty functions SCAD and MCP result, for our data set, in the same predictor selection. In the following sections we use the ":"-sign to indicate the interaction of two predictors.

Standard

The grouped lasso selection with standard penalty function selects the following predictors to be used by stepwise regression:

- Kenniscentrum
- Fear
- ADHD
- Western country of birth
- Age end of study
- Born in the Netherlands
- Western nationality
- Gender
- Born in EU
- Previous education qualification
- Age start study
- Handicapt
- Asperger syndrome
- Years compulsory education
- NLD
- Autism
- Disorders as warning
- Intensity
- PDD-NOS
- Highest level enjoyed previous education
- Number of previous studies
- Disorders on student pass
- EU nationality
- Dutch nationality
- Years in Netherlands start study
- ODD
- Education level
- ADD
- Dyslexia
- ASS

The stepwise regression started with the "empty" model and the most complex model in the search space included all pairwise interactions. The predictors and the coefficients of the model found by stepwise regression can be seen in table 6. The ROC curve of the model on the test set can be found in figure 1. The AUC is 0.78.

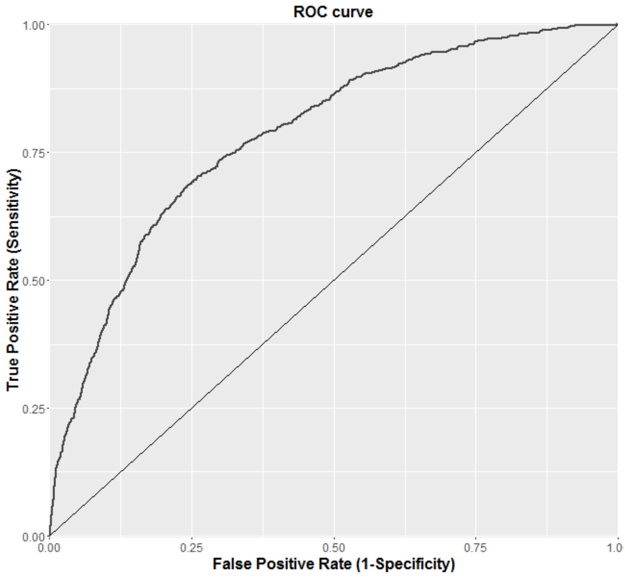

Figure 1: The ROC curve of the logistic regression model found by stepwise regression with a selection on the predictors by grouped lasso with the standard penalty function.

SCAD and MCP

The SCAD en MCP penalty functions for grouped lasso selection both select the same predictors. These are:

- Kenniscentrum
- Disorders as warning
- Intensity
- Fear
- Age end of study
- Gender
- Number of previous studies
- Western country of birth
- Dutch nationality
- Previous education qualification
- Highest level enjoyed previous education
- Age start study
- Handicap
- Study level
- Years compulsory education
- Dyslexia

The selected predictors and the coefficients of the model resulting from stepwise regression can be found in table 7. Interesting is to see that the model found here is the same as the model found by the standard grouped lasso penalty, except for the use of *disorders as warning* instead of *disorders on student pass*. The ROC curve, with an AUC of 0.78, can be found in figure 2.

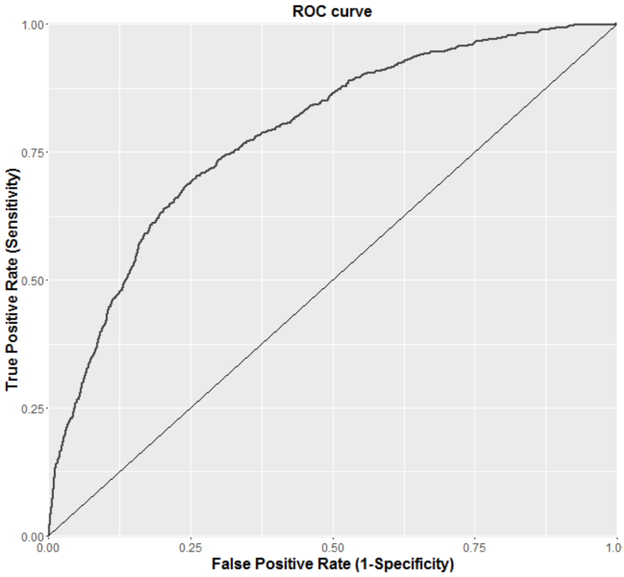

Figure 2: The ROC curve of the logistic regression model found by stepwise regression with a selection on the predictors by grouped lasso with the SCAD penalty function.

MCP; fewer age predictors

There are multiple predictors that involve the age of the student. These are:

- Age start study
- Age at end of study
- Years compulsory education
- Age in the Netherlands
- Years in the Netherlands at start study

All these predictors are derived from three variables: the date of birth of the student, the date the student came to the Netherlands and the start date of the students most recent study. These predictors overlap in context. To see what the model would look like when we do not use all the predictors, grouped lasso selection and stepwise regression were run on the data set without *Age end of study*, *Years compulsory education* and *Years in the Netherlands at start study*. The resulting model has an AUC of 0.77. The coefficients can be found in table 8. It is interesting to see that there is only a minor drop in AUC, it uses fewer predictors and *Dutch nationality* is used by the model.

6.1.2 *Ungrouped*

In the ungrouped variant of lasso, the levels of the categorical predictors are not linked. This way, one or some levels of a categorical variable might be selected while others are not. The ungrouped lasso selects the following predictors:

- Woman
- Disorders as warning
- Dyslexia
- Exam-participant
- Handicap
- Education level 2
- Kenniscentrum Innovam groep
- Kenniscentrum SH&M
- Previous education qualification
- No previous education
- MBO 1 or 2
- Dutch nationality
- Years compulsory education
- Number of previous studies
- ADHD
- ODD
- Full-time student
- Education level 1
- Education level 4
- Kenniscentrum Fundeon
- Kenniscentrum Handel
- Kenniscentrum Stichting Kenwerk
- Basisvorming
- Havo
- MBO 3 or 4
- Age start study
- Years in Netherlands start study

The stepwise regression started with the "empty" model and the most complex model in the search space included all pairwise interactions. The model, found by stepwise regression, has an AUC of 0.77. The ROC curve is found in figure 3, the coefficients in table 9.

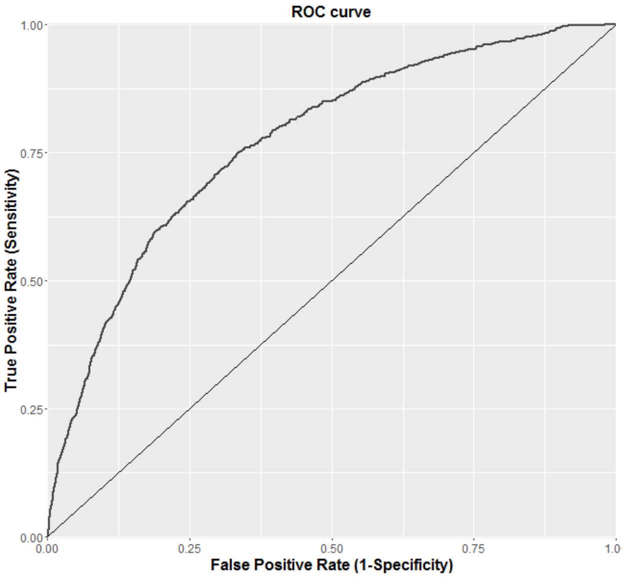

Figure 3: The ROC curve of the logistic regression model found by stepwise regression with a selection on the predictors by ungrouped lasso.

6.2 CLASSIFICATION TREE

6.2.1 *Without Boosting*

Parameter tuning

The parameter values for the classification tree without boosting were tuned by 10-fold cross validation. The maximum number of splits considered are all tens between 10 and 200. For the confidence factor the cross validation considered: 0.05, 0.10, 0.15, 0.20, 025, 0.30 and 0.40. The parameter values with the highest average AUC over the 10 folds are:

- maximum number of Splits = 100
- Confidence factor = 0.05

Model

The trained classification tree has an AUC of 0.71 on the test set. The corresponding ROC curve can be found in figure 4. The trained model has only 5 decision nodes and is visualised in figure 5. We listed in table 10 for each predictor the percentage of samples that had the predictor in their path to their leaf.

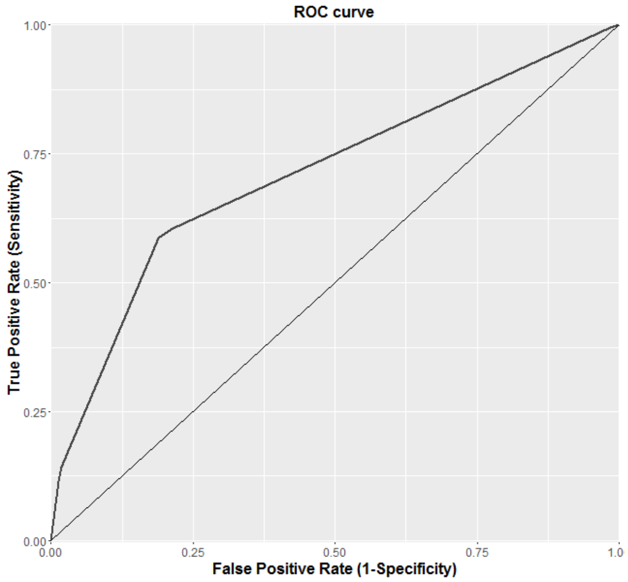

Figure 4: The ROC curve of the classification tree, created by the C5.0 algorithm without boosting, on the test set with basic information.

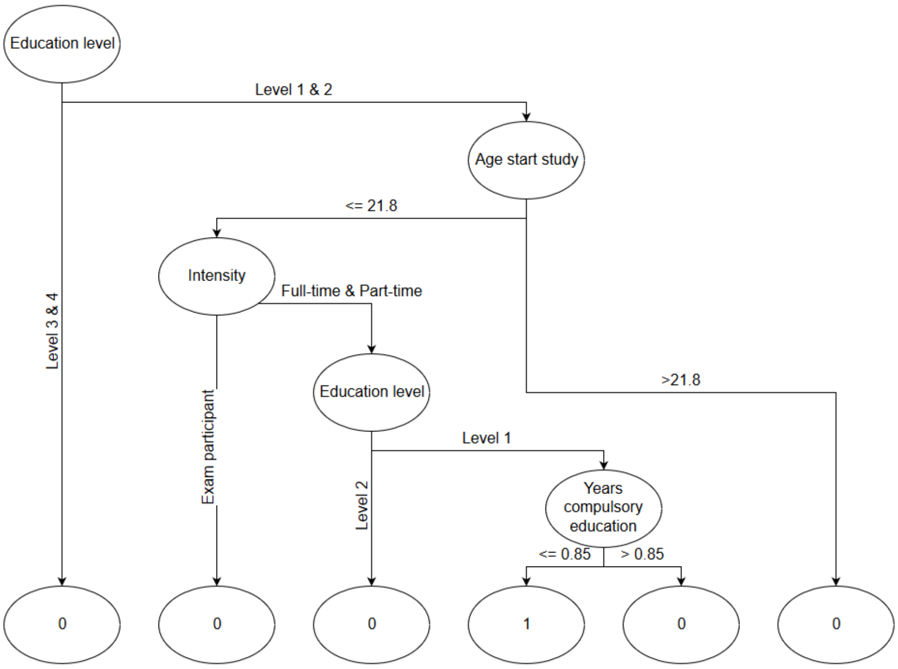

Figure 5: The classification tree created by the C5.0 algorithm without boosting on the data set with basic information.

6.2.2 *Boosting*

Parameter tuning

The model below are created by the C5.0 algorithm with the parameter values found using 10-fold cross validation. The cross validation considered for the maximum number of splits all tens between 10 and 200, a confidence factor of 0.05, 0.10, 0.15, 0.20, 025, 0.30 and 0.40 and the number of boosting trials of 2, , 5, 10, 15, 20, 30, 40, 60, 80 and 100. The parameter values found by the cross validation are:

- Trials = 15
- Splits = 100
- Confidence factor = 0.25

Although the model had the possibility to have 15 trials for the boosting procedure, it used only eight, because of the evaluation at halfway the boosting procedure.

Model

The classification tree created by the C5.0 algorithm has an AUC of 0.77. The ROC curve can be seen in figure 6. The predictors used in the classification and their importance, the percentage of training set samples that have the predictor in their path to their end node, can be found in table 11.

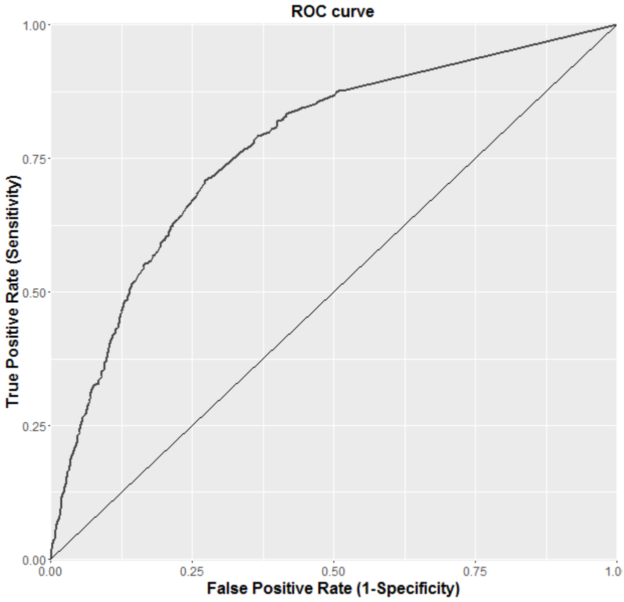

Figure 6: The ROC curve of the classification tree, created by the C5.0 algorithm with boosting, on the test set with basic information.

		Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)		-36.3997	5.7468	-6.33	0.0000
Education level	Education level 2	-2.1794	1.1465	-1.90	0.0573
	Education level 3	0.5700	1.2817	0.44	0.6565
	Education level 4	-8.3141	1.4524	-5.72	0.0000
Intensity	Exam participant	-2.5696	0.4092	-6.28	0.0000
	Full-time	0.2320	0.2205	1.05	0.2926
Highest level enjoyed previous education	Basisvorming	0.1798	0.1139	1.58	0.1142
	No previous education	-0.5346	0.3674	-1.46	0.1457
	Havo	-14.0906	109.4035	-0.13	0.8975
	HBO	-13.4808	1183.4264	-0.01	0.9909
	MBO 1 or 2	-0.2348	0.1584	-1.48	0.1383
	MBO 3 or 4	-1.8231	0.3764	-4.84	0.0000
	propedeuse HBO	-13.8499	1686.3204	-0.01	0.9934
	VMBO	-0.0795	0.1122	-0.71	0.4788
	VMBO-TL	-0.0289	0.1257	-0.23	0.8182
	VWO	-0.7682	0.9224	-0.83	0.4049
Years compulsory education		-7.2708	1.4256	-5.10	0.0000
Age start study		1.8670	0.3202	5.83	0.0000
Gender	Woman	-0.2302	0.0483	-4.76	0.0000
Dyslexia		-1.2402	0.2327	-5.33	0.0000
Disorders on student pass		0.5352	0.1247	4.29	0.0000
Age end of study		1.8801	0.2480	7.58	0.0000
Education level : Intensity	Education level 2 : Exam participant	2.5071	0.4405	5.69	0.0000
	Education level 3 : Exam participant	3.4424	0.4570	7.53	0.0000
	Education level 4 : Exam participant	1.8383	0.5262	3.49	0.0005
	Education level 2 : Full-time	0.0843	0.2268	0.37	0.7100
	Education level 3 : Full-time	0.0141	0.2400	0.06	0.9531
	Education level 4 : Full-time	-0.2949	0.2748	-1.07	0.2832
Years compulsory education : Age start study		0.4423	0.0873	5.07	0.0000
Intensity : Years compulsory education	Exam participant : Years compulsory education	-6.4603	2.8488	-2.27	0.0233
	Full-time : Years compulsory education	-0.2788	0.0990	-2.81	0.0049
Age start study : Age end of study		-0.0961	0.0137	-7.02	0.0000
Education level : Age end of study	Education level 2 : Age end of study	0.0543	0.0561	0.97	0.3332
	Education level 3 : Age end of study	-0.1269	0.0618	-2.05	0.0402
	Education level 4 : Age end of study	0.2540	0.0677	3.75	0.0002

Table 6: The coefficients of the logistic regression model found by stepwise regression with as criteria BIC. The predictors considered by stepwise regression were selected by grouped lasso with the standard penalty function.

		Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)		-36.3972	5.7462	-6.33	0.0000
Education level	Education level 2	-2.1494	1.1455	-1.88	0.0606
	Education level 3	0.6085	1.2810	0.48	0.6347
	Education level 4	-8.2642	1.4516	-5.69	0.0000
Intensity	Exam participant	-2.5695	0.4091	-6.28	0.0000
	Full-time	0.2334	0.2204	1.06	0.2895
Highest level enjoyed previous education	Basisvorming	0.1793	0.1139	1.57	0.1153
	No previous education	-0.5330	0.3676	-1.45	0.1471
	HAVO	-14.0919	109.3944	-0.13	0.8975
	HBO	-13.4823	1183.3493	-0.01	0.9909
	MBO 1 or 2	-0.2332	0.1584	-1.47	0.1409
	MBO 3 or 4	-1.8231	0.3764	-4.84	0.0000
	propedeuse HBO	-13.8509	1686.2862	-0.01	0.9934
	VMBO	-0.0800	0.1122	-0.71	0.4761
	VMBO-TL	-0.0296	0.1257	-0.24	0.8137
	VWO	-0.7386	0.9132	-0.81	0.4186
Years compulsory education		-7.2899	1.4253	-5.11	0.0000
Age start study		1.8654	0.3201	5.83	0.0000
Gender	Woman	-0.2303	0.0483	-4.77	0.0000
Dyslexia		-0.9695	0.2100	-4.62	0.0000
Disorders as warning		0.4811	0.1181	4.07	0.0000
Age end of study		1.8800	0.2480	7.58	0.0000
Education level : Intensity	Education level 2 : Exam participant	2.5095	0.4405	5.70	0.0000
	Education level 3 : Exam participant	3.4434	0.4570	7.54	0.0000
	Education level 4 : Exam participant	1.8381	0.5261	3.49	0.0005
	Education level 2 : Full-time	0.0855	0.2267	0.38	0.7061
	Education level 3 : Full-time	0.0153	0.2399	0.06	0.9490
	Education level 4 : Full-time	-0.2961	0.2747	-1.08	0.2811
Years compulsory education : Age start study		0.4434	0.0873	5.08	0.0000
Intensity : Years compulsory education	Exam participant : Years compulsory education	-6.4591	2.8477	-2.27	0.0233
	Full-time : Years compulsory education	-0.2778	0.0990	-2.81	0.0050
Age start study : Age end of study		-0.0960	0.0137	-7.01	0.0000
Education level : Age end of study	Education level 2:Age end of study	0.0527	0.0560	0.94	0.3469
	Education level 3:Age end of study	-0.1288	0.0618	-2.08	0.0371
	Education level 4:Age end of study	0.2517	0.0676	3.72	0.0002

Table 7: The coefficients of the logistic regression model found by stepwise regression with a selection on the predictors by grouped lasso with the SCAD penalty function.

		Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)		-0.3563	0.8847	-0.40	0.6871
Education Level	Education level 2	0.3223	0.9431	0.34	0.7325
	Education level 3	2.4650	1.0283	2.40	0.0165
	Education level 4	-0.0185	1.0762	-0.02	0.9863
Intensity	Exam participant	-3.2487	0.3876	-8.38	0.0000
	Full-time	0.1172	0.2025	0.58	0.5630
Highest level enjoyed previous education	Basisvorming	0.3233	0.1157	2.79	0.0052
	No previous education	-0.4267	0.3448	-1.24	0.2158
	Havo	-13.8532	112.0556	-0.12	0.9016
	HBO	-15.9582	1098.7028	-0.01	0.9884
	MBO 1 or 2	-0.0558	0.1571	-0.36	0.7225
	MBO 3 or 4	-2.1047	0.4283	-4.91	0.0000
	Propedeuse HBO	-13.5097	1696.6024	-0.01	0.9936
	VMBO	0.0975	0.1143	0.85	0.3936
	VMBO-TL	0.0599	0.1276	0.47	0.6391
	VWO	-0.4045	0.8663	-0.47	0.6405
Gender	Woman	-0.7360	0.1455	-5.06	0.0000
Age start study		0.0296	0.0452	0.65	0.5128
Dyslexia		-1.1909	0.2139	-5.57	0.0000
Disorders as warning		0.5715	0.1153	4.96	0.0000
Dutch nationality		-0.5114	0.0957	-5.34	0.0000
Education level : Intensity	Education level 2 : Exam participant	3.0295	0.4208	7.20	0.0000
	Education level 3 : Exam participant	3.9492	0.4401	8.97	0.0000
	Education level 4 : Exam participant	2.7997	0.4875	5.74	0.0000
	Education level 2 : full-time	0.0313	0.2135	0.15	0.8836
	Education level 3 : full-time	-0.0037	0.2271	-0.02	0.9870
	Education level 4 : full-time	-0.3675	0.2618	-1.40	0.1605
Gender : Dutch Nationality	Woman : Dutch nationality	0.5152	0.1535	3.36	0.0008
Education level : Age start study	Education level 2 : Age start study	-0.0596	0.0494	-1.21	0.2277
	Education level 3 : Age start study	-0.2314	0.0534	-4.33	0.0000
	Education level 4 : Age start study	-0.1192	0.0548	-2.18	0.0295

Table 8: The coefficients of the logistic regression model found by stepwise regression with as criteria BIC. The predictors considered by stepwise regression were selected by grouped lasso with the MCP penalty function. The predictors *age end of study*, *years in Netherlands start study* and *years compulsory education* were removed from the used data set to create this model.

		Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)		4.6799	0.5601	8.36	0.0000
Education level	Education level 4	-0.1607	0.1235	-1.30	0.1930
	Education level 2	-1.1926	0.5698	-2.09	0.0363
	Education level 1	2.2317	0.1217	18.33	0.0000
Intensity	Exam participant	0.1933	0.1223	1.58	0.1139
	Full-time	0.2259	0.0562	4.02	0.0001
Highest level enjoyed previous education	Havo	-13.8840	107.6531	-0.13	0.8974
	MBO 3 or 4	-1.6400	0.3625	-4.52	0.0000
Age start study		-0.3442	0.0285	-12.08	0.0000
years compulsory education		-8.6394	1.4253	-6.06	0.0000
Kenniscentrum	Kenniscentrum Fundeon	-0.8876	0.1436	-6.18	0.0000
	Kenniscentrum Handel	0.5154	0.0729	7.07	0.0000
	Kenniscentrum Innovam Groep	0.4087	0.0983	4.16	0.0000
Gender	Woman	-0.3257	0.0500	-6.51	0.0000
Qualification		-0.1670	0.0592	-2.82	0.0048
Dyslexia		-1.1957	0.2167	-5.52	0.0000
Disorders as warning		0.7056	0.1151	6.13	0.0000
Handicap		-0.8893	0.2452	-3.63	0.0003
Education level 1 : Exam participant		-3.1479	0.4064	-7.75	0.0000
Age start study : Years compulsory education		0.4783	0.0872	5.49	0.0000
Exam participant : Years compulsory education		-8.4173	3.6418	-2.31	0.0208
Education level 4 : Qualification		-0.5908	0.1310	-4.51	0.0000
Years compulsory education : Kenniscentrum Fundeon		0.7381	0.1793	4.12	0.0000
Education level 1 : Years compulsory education		-0.5951	0.1424	-4.18	0.0000
Education level 2 : Age start study		0.1200	0.0305	3.94	0.0001
Education level 4 : Kenniscentrum Handel		-1.2013	0.2103	-5.71	0.0000

Table 9: The coefficients of the logistic regression model found by stepwise regression with as criteria BIC. The predictors considered by stepwise regression were selected by lasso with the standard penalty function.

	Overall
Education level	100.00
Age start study	29.18
Intensity	27.54
Years compulsory education	3.54

Table 10: The predictors used in the classification tree, created by the C5.0 algorithm without boosting. Importance is the percentage of subjects in the training set, with basic information, that have the predictor in their path to the end node.

	Overall
Education level	100.00
Highest level enjoyed education	100.00
Age end study	100.00
Years in the Netherlands	93.57
Dyslexia	93.41
Disorders as warning	90.39
Handicap	88.33
Years compulsory education	87.03
Number of previous studies	81.06
Intensity	80.43
Kenniscentrum	51.40
Age start study	29.38
Gender	26.81
Born in western country	26.32

Table 11: The predictors used in the classification tree, created by the C5.0 algorithm with boosting. Importance is the percentage of subjects in the training set, with basic information, that have the predictor in their path to the end node.

7

RESULTS; TRAJECTORY INFORMATION

Using the data set with the basic and trajectory information, We constructed classification models as described in chapter 5. In this chapter, we describe the resulting models and their performance on the test set.

7.1 logistic regression

7.1.1 *Grouped lasso*

Using the data set with basic and trajectory information, we created a model for each penalty function. Both MCP and SCAD selected only education level and average grade to be used by stepwise regression. The grouped lasso selection with the standard penalty function selected the following predictors:

- ODD
- Autism
- Kenniscentrum
- Highest level enjoyed previous education
- ADD
- Disorders as warning
- Intensity
- Hospital visit
- Born in EU
- Appointment
- suspended
- Number of previous studies
- Behavioural problem
- Sick
- Number of grades
- Late for class
- Age end of study
- Specialist
- Standard deviation grades
- Age in the Netherlands
- Fear
- Dutch nationality
- Leave
- Age start of study
- Handicap
- Years compulsory education
- Dyscalculia
- Gender
- Western nationality
- Dyslexia
- ADHD
- Education level
- Average grade

Although grouped lasso with the standard penalty function selects far more predictors for stepwise regression, the end results of stepwise regression starting from the "empty" model is for all the penalty functions the same. Stepwise regression ends with a small logistic regression model using only the level of education and the average grade. The ROC curve of this model on the test set can be found in figure 7. The AUC of this curve is 0.75. The coefficients can be found in table 12. Note when interpreting these results that the data set with the trajectory information is small and the quality of the data available is questionable.

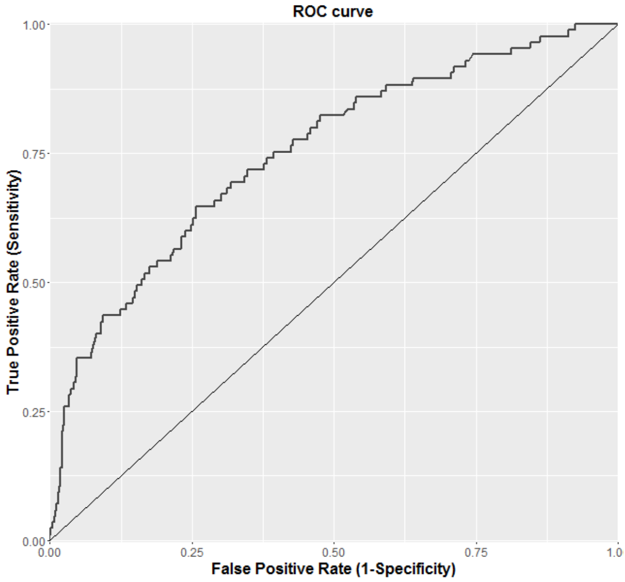

Figure 7: The ROC curve of the logistic regression model found by stepwise regression, with a selection on the predictors by grouped lasso with either the SCAD, MCP or standard penalty function. The model is trained using the train set containing basic and trajectory information.

7.1.2 *Ungrouped lasso*

The ungrouped lasso selects the levels of the same predictors as grouped lasso with the SCAD or MCP penalty function, but not all levels. The selected predictors are:

- Education level 1
- Education level 2
- Average grade

The model found by stepwise regression drops the average grade as predictor. The ROC curve of the model can be found in figure 8 and has an AUC of 0.70. The coefficients of the model can be found in table 13.

7.2 classification tree

7.2.1 *Without boosting*

Parameter tuning

The parameters for the classification tree without boosting were tuned by 10-fold cross validation. The parameters with the highest average AUC over the 10 folds are.

- Splits = 160
- Confidence factor = 0.4

Model

The tree trained on the training set with the above parameters has an AUC of 0.71 on the test set. The tree consists out of four decision nodes and can be seen in figure 10. The ROC curve can be found in figure 9. The predictors used and their importance to the classification can be found in table 12.

7.2.2 *Boosting*

For the classification tree with boosting we tuned the parameters by 10-fold cross validation. The parameters with the highest average AUC over the 10 folds are:

- Splits = 280
- Trials = 100
- Confidence factor = 0.5

The C5.0 used in during the boosting only 9 of 100 iterations. The resulting classification tree has an AUC of 0.77. The ROC curve can be found in figure 11. The used predictors and their importance can be found in table 15

		Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)		2.4885	0.7346	3.39	0.0007
Education level	Education level 2	-1.5140	0.6939	-2.18	0.0291
	Education level 3	-3.1384	0.7061	-4.44	0.0000
	Education level 4	-3.5502	0.7025	-5.05	0.0000
Average grade		-3.5583	0.4190	-8.49	0.0000

Table 12: The coefficients of the logistic regression model found by stepwise regression with a selection on the predictors by grouped lasso with either the SCAD, MCP or standard penalty function. The model is trained using the train set containing basic and trajectory information.

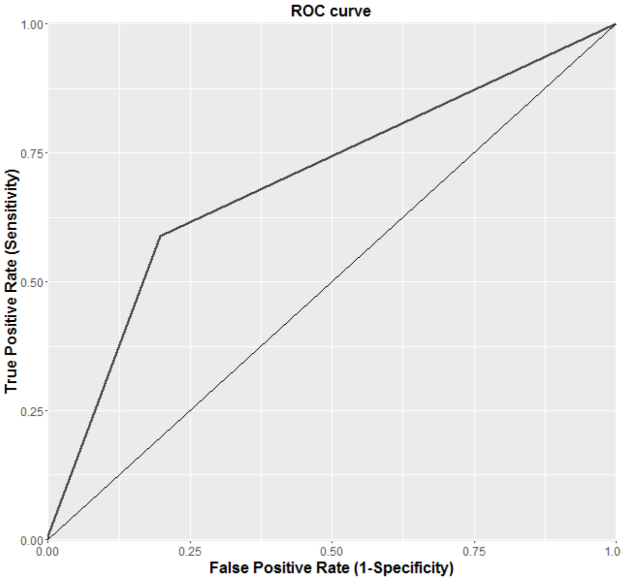

Figure 8: The ROC curve of the logistic regression model found by stepwise regression with a selection on the predictors by ungrouped lasso. The model is trained using the train set containing basic and trajectory information.

		Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)		-3.0029	0.1040	-28.87	0.0000
Education level	Education level 2	1.8142	0.1393	13.02	0.0000
	Education level 1	3.4084	0.6538	5.21	0.0000

Table 13: The coefficients of the logistic regression model found by stepwise regression with a selection on the predictors by lasso. The model is trained using the train set containing basic and trajectory information.

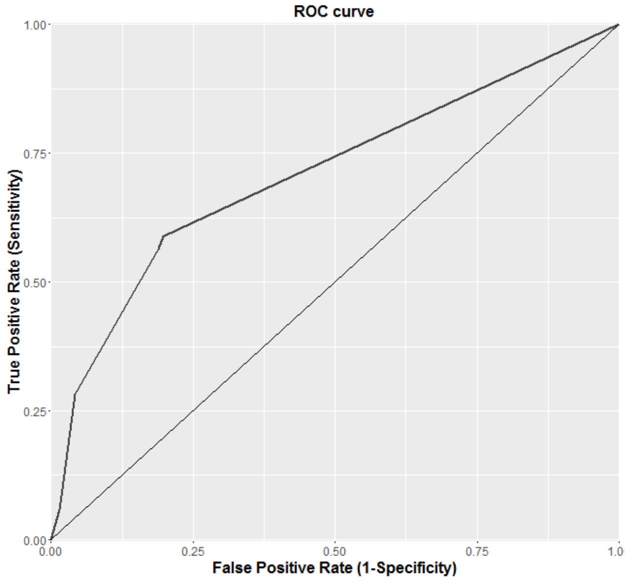

Figure 9: The ROC curve of the classification tree, created by the C5.0 algorithm without boosting, on the test set with trajectory information.

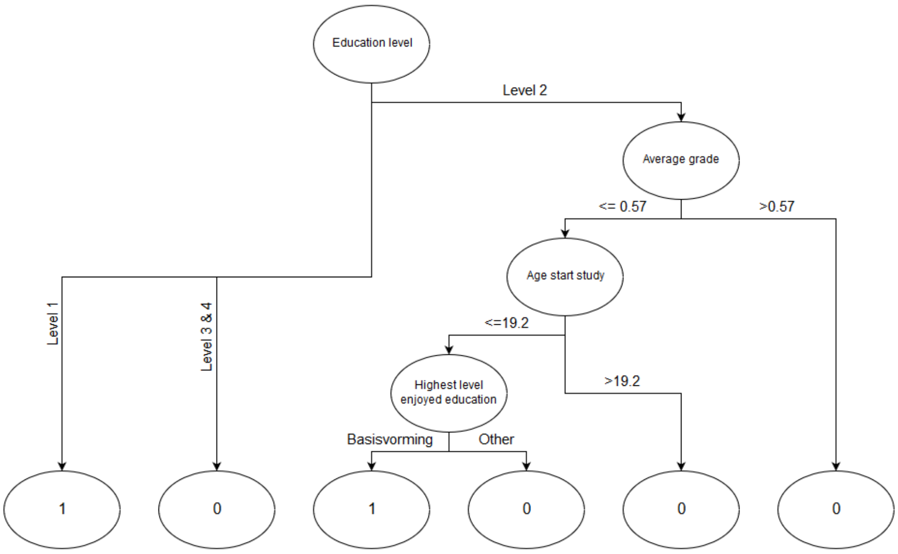

Figure 10: The classification tree created by the C5.0 algorithm on the data set with trajectory information.

	Importance
Education level	100.00
Average grade	24.00
Age start study	7.63
Highest level enjoyed previous education	6.08

Table 14: The predictors used in the classification tree, created by the C5.0 algorithm without boosting with the trajectory information. Importance is the percentage of subjects in the training set that have the predictor in their path to the end node.

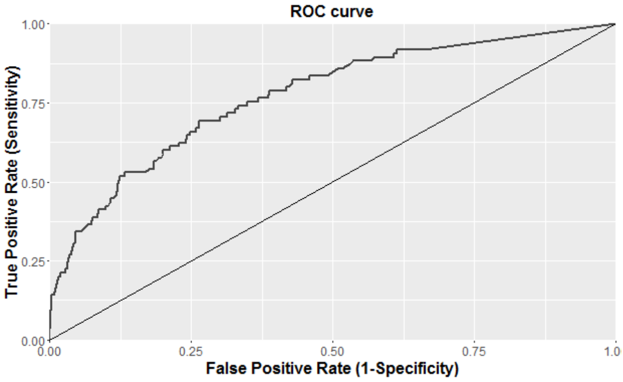

Figure 11: The ROC curve of the classification tree, created by the C5.0 algorithm with boosting, on the test set with trajectory information.

	Overall
Education level	100.00
Highest level enjoyed previous education	100.00
Age end study	100.00
Age start study	100.00
Number of grades	97.23
Specialist	97.09
Appointment	96.28
Years in the Netherlands	93.66
Average grade	92.37
Handicap	92.00
Late for class	90.49
Kenniscentrum	89.53
Type of education	88.79
Total absence	88.02
Hospital visit	87.87
standard deviation grades	87.50
School activity	84.81
Number of previous studies	84.70
leave	84.18
Autism	64.90
Dyscalculia	60.73
Personal circumstances	60.40
Born in the Netherlands	46.68
Dyslexia	45.21
Suspended	43.47
Doctor	42.63
Gender	37.83
Years compulsory education	36.28
Western nationality	35.62
Disorders on the student card	35.03
Age in the Netherlands	32.37
Intensity	23.34
Sick	16.04

Table 15: The predictors used in the classification tree, created by the C5.0 algorithm with boosting with the trajectory information. Importance is the percentage of subjects in the training set that have the predictor in their path to the end node.

8

INTERPRETATION OF THE LOGISTIC REGRESSION MODEL

The logistic formula used by the logistic regression models makes it hard to interpret the effect of the predictors on the estimated probability. Some predictors are also part of an interaction predictor. This makes it even harder to interpret the effects of these predictors. In this section, we try to interpret some predictors by looking at their marginal effects on the predicted probability.

The marginal effect is calculated by keeping all predictors of the students unchanged except the predictor of which we want to know the marginal effect. The value of that predictor is set to the same value for all the students. For all students we calculate the estimated probability with the set value and with the value of the predictor increased with one unit. The difference between the estimated probabilities is the marginal effect. Example, for the calculation of the marginal effect of dyslexia we set the predictor dyslexia for all students on false (as if none of the students has dyslexia) and calculate the estimated probability for each student. We also calculate the estimated probability of each student with the predictor dyslexia on true (as if all of the students have dyslexia). For each student we calculate the difference between their two estimated probabilities, giving us the marginal effect. We average this to get the average marginal effect of the predictor. The average marginal effect found for dyslexia is -0.0747. Meaning that with all else equal, on average the probability that a student with dyslexia drops out is approximately 7.5 percentage points less than for a student that does not have dyslexia (more on the interpretation dyslexia in section 8.1).

Besides the average marginal effect, we report the median, min, max and standard deviation of the marginal effects. With these descriptive statistics we try to interpret the effects of the predictors. The interpretation in this section is about the models trained with the basic information set. The marginal effects reported here are that of the logistic regression model found by stepwise regression with a selection on the predictors by grouped lasso with the SCAD penalty function (table 7).

8.1 DYSLEXIA

The predictor dyslexia has a negative coefficient in the logistic model and is not part of an interaction. This indicates that, according to the model, students with dyslexia have a lower probability to drop out. The average marginal effect is -0.0747. Meaning that with all else equal, on average the probability that a student with dyslexia drops out is approximately 7.5 percentage points less than for a student that does not have dyslexia. On average a small difference, but for some students the effect is quite large: the maximum marginal effect is a reduction of 23.8 percentage points. Still, for most students it has a small effect. The median marginal effect is -0.0531 and the standard deviation is 0.0624. A box plot of the marginal effect can be found in figure 12.

The negative effect on the dropout probability is surprising. Having dyslexia makes it harder to learn and to make exams. These are enough reasons to expect that these students are vulnerable to be early school leavers. From the data available to us it is hard to give a reason why the opposite is true. Possible explanations are that students with dyslexia get more attention from teachers, the benefits (e.g. more time for exams) outweigh the negative effects or these students have a different attitude towards their study having to deal with dyslexia all their life.

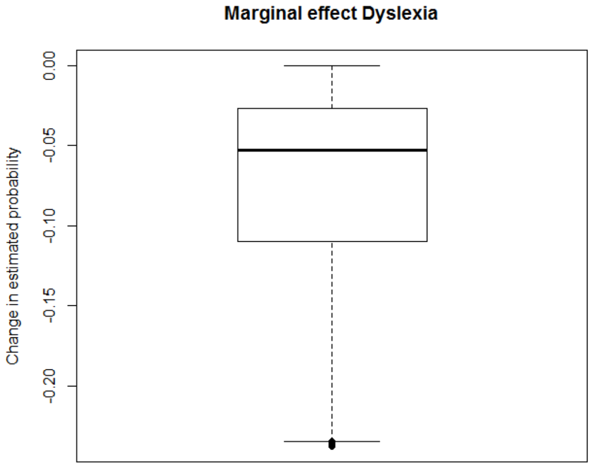

Figure 12: Boxplot of the marginal effect of dyslexia.

8.2 GENDER

The predictor woman has a negative coefficient in the logistic model, which indicates that, according to the model, women have a lower probability to be early school leavers than men. This is also visible in the marginal effects, but the effect of the predictor is small. The average marginal effect is -0.0230, or a reduction of 2.3 percentage points, with a standard deviation of -0.0168. The maximum marginal effect is -0.0575. The boxplot of the marginal effects can be found in figure 13.

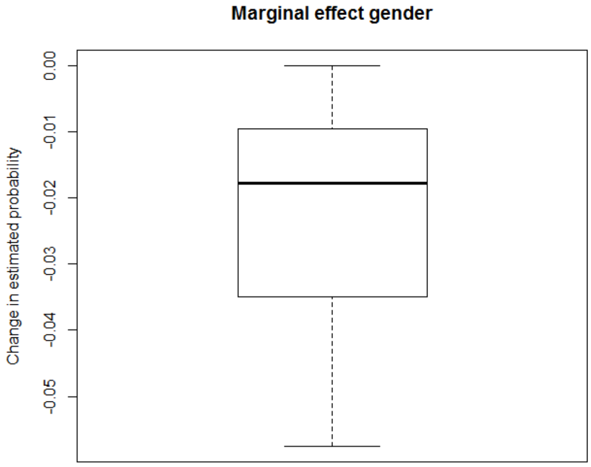

Figure 13: Boxplot of the marginal effect of gender.

8.3 AGE

The age of a student is incorporated trough multiple predictors in the model. In the above calculations we only change one predictor. As it makes no sense to set, for example, the age a student starts with his study to 20, but keep the number of compulsory education years of the student to be 2 years, we manipulated all predictors containing the age. The three predictors that contain the age of the student are: Age end study, Age start study and years compulsory education. These predictors are also part of some interactions. Boxplots of the marginal effect of age can be found in figure 15 and the boxplots of the estimated probabilities of the students with their age set to a specific value can be found in figure 14. In the latter, we see that risk of dropping out increases when the students is older when he or she starts with his or her study until the age of 18. After the age of 18 the risk decreases with each year.

The national statistics [28] tell us that most students drop out after their 18th birthday. Can we give an explanation to the decrease in estimated probability after the age of 18 in our model? first, note that the variable age in figures 14 and 15 is the age the student starts his study and not their current age. This together with the definition of an early school leaver (see section 1.1) provide the first explanation. Students who start at a later age can drop out during his study, but not be an early school leaver. For example, a student that starts his study at the age 21 will only be labelled as an early school leaver if he drop outs within a year. Another explanation can be that students that start a study at a later age have made a more thought-out decision are therefore more likely to finish their study.

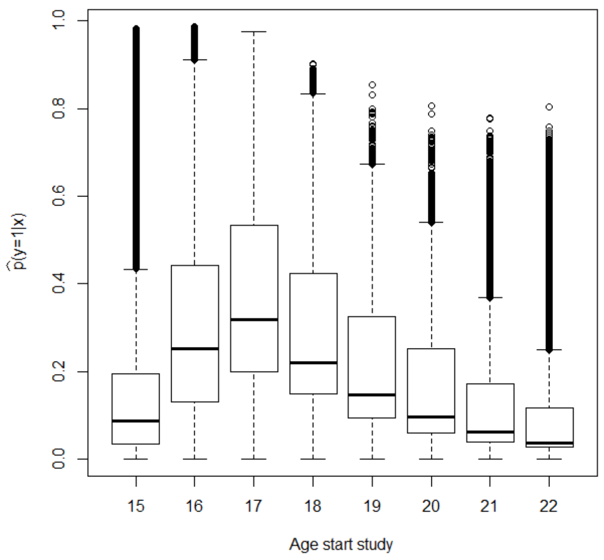

Figure 14: Boxplots of the estimated probabilities of the students in the dataset with their age set to the value on the x-axis.

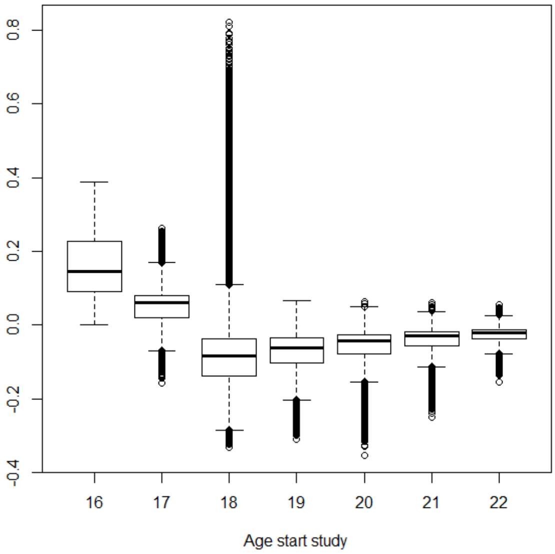

Figure 15: Boxplots of the marginal effect of the predictor age. The boxplot x shows the change in estimated probability when the predictor age changes from $x-1$ to x.

9

DISCUSSION

We started our venture wondering if is possible for an automated system to detect potential dropouts. We specified this question in two ways.

Firstly, we wanted to look if this is possible with the information in the current student information systems. The AUC scores show that early school leavers can be detected reasonably well with information in the student information system that is available at the start of students study. However, when we look at the confusion matrix of, for instance, one of the logistic models (see table 16), we see that only a small number of the early school leavers are correctly classified (sensitivity). If we set the decision threshold lower to increase the sensitivity, then the number of misclassified non-dropouts (false positives) increases (see table 17 and see table 18 for a confusion matrix with a higher threshold). One might argue that it is worth to have more false positives, if this increases the number of correctly classified dropouts. In some situations this might hold true, but schools have only limited resources to spend on potential dropouts. Putting non-dropouts in dropout prevention programs would be a waste of these resources. The increase of false positives is too large to use this as a solution. This sounds like we are heading towards a negative conclusion about our second question; current student information systems will not provide enough information for an automated system to predict dropouts. But we see in our results that there is a strong potential for these kind of systems. We feel encouraged by our results to say that a future model that incorporates more information (e.g. the grades of the students) will reach the performance needed to be of real use to schools. Furthermore, we speak here of the performance of the isolated system. The models could also be used as a recommendation system for a human, who makes the classification decision. Such a machine-human combination might outperform the individual system and human.

Secondly, at the start we stated our belief that automated systems can contribute to the detection of unexpected early school leavers. Unfortunately, we do not have knowledge about which students in our data set dropped out unexpectedly. We do not expect that the current system as an isolated system would detect the unexpected

		Early school leavers	
		0	1
Prediction	0	5419	763
	1	65	112

Table 16: The confusion matrix for the logistic regression model found by stepwise regression with a selection on the predictors by grouped lasso with the SCAD penalty function. The decision threshold is 0.5.

		Early school leavers	
		0	1
Prediction	0	4973	525
	1	511	350

Table 17: The confusion matrix for the logistic regression model found by stepwise regression with a selection on the predictors by grouped lasso with the SCAD penalty function. The decision threshold is 0.3.

early school leavers, but as we said before we believe that the system can be improved by incorporating more information. This would also improve the detection of the unexpected early school leavers.

Besides answers to our questions there is more information to extract from the results. The models decision process is comprehensible and we can see which indicators are important factors. This can provide schools with valuable insights on which students might be a potential dropout and this can provide future research a starting point to create improved models. Furthermore, previous research had already shown that although nationality can be a dropout indicator, this can be explained by other factors [36]. Our results confirm this as nationality is not included in any of our decision models.

9.1 ETHICS

Using technology in education has a lot of potential. It opens ways for improving and personalising study paths and study material. But using (big data) technology in educational settings also comes with a risk. Current objections to the use of big data focus mostly on privacy and confidentiality issues. Although this is a major concern, there are also other concerns that come with the use of these technologies in education.

An important question is to whom and how the gained information from educational data mining is shared. It is seen by experts as an ethical obligation to share the knowledge in a way that it bene-

		Early school leavers	
		0	1
Prediction	0	5465	836
	1	19	39

Table 18: The confusion matrix for the logistic regression model found by stepwise regression with a selection on the predictors by grouped lasso with the SCAD penalty function. The decision threshold is 0.6.

fits the stakeholders (e.g. the student, the teacher and/or the school management) [4]. But sharing the knowledge can also do harm.

For instance, the student labels created by educational data mining can create deterministic student paths. This can happen in two ways: by influencing the teachers and by limiting the choice of the student. Research has shown that the teacher's expectations affect student performance. Learners who are expected to be smart, whether or not this is true, will have a higher learning rate [35]. Labels create such expectations. The use of data mining techniques to identify the less potential can increase the chance of these students to become one, as teachers will (unconsciously) treat them as such. Also sharing this kind of information with the student (e.g. you have 20% chance to successfully finish this course) can affect the student behaviour. This could be positive (the student invests more time in the course) or negative (the student is demotivated by the information).

The deterministic student path can also be caused by a limited choice in student path. Educational institutes already set preconditions on course and study enrolment or use methods as numerus clausus to limit the number of students. If data mining techniques are used to determine who is allowed to enrol in a course/study (e.g. only the students who are certain to successfully finish the course/study), then students might be limited by facts out of their control. Xavier Prats Monn, the Director General of the Directorate-General for Education and Culture of the European Commission, writes in the European journal of education:

> "Technology and big data also bring new risks — not just for privacy, as often stated, but, more importantly, for the temptation of determinism: since ICTs forget nothing, learners could be bound by their own past or denied from an early age the recognition of their ability to improve; and they could be limited in their own choice and freedom to learn by institutions playing with statistics and predictive algorithms" [30].

Furthermore, the students do not know on which factors the (data mining) system categorises/judges them, which creates a Big Brother

[29] like situation. The students need to behave on their best in all possible fields, to make sure they keep access to courses. To prevent and/or reduce the negative effects of the knowledge gained, research on how the information is shared is needed.

In short, we need to be careful as even though a technology is implemented with the right intentions it can have unwanted side-effects. The remainder of this section reviews the model created in this project with these unwanted side-effects in mind.

9.1.1 *Ethics & automated dropout detection*

The model created in this project will help students who are detected correctly as early school leavers (true positives), but can have negative effects for other students. The labelling is disadvantageous for students who are not and will never be early school leavers, but are wrongly classified as one. These false positive classified students have nothing to gain from the label. The dropout prevention program will not help them as they were not going to drop out anyway[1], but the label can negatively affect their study path, because teachers and schools will have different expectations. This not the only reason to focus on a low false positive rate. As said before, there is only a limited amount of resources to spend on the help of potential dropouts. Reducing the number of false positives would mean that more recourse are spend where they are needed. Still, the system would be most beneficial if it would identify the students who now drop out of school unexpectedly. As these are hard to identify, the system might need a focus on high precision. This would probably happen at the expense of the false positive rate.

No matter which decision the schools makes about the precision and accuracy trade-off, it is important that the schools think about who can access the information gain from the system and know why a student will be classified as a potential dropout. As long as we keep realising that the classification influences a real human being and we are aware of the positive and negative effects of the classification, we are really helping the students.

[1] We assume the most negative situation: the dropout prevention program only effects the chance of being an early school leaver. Although, it might be that the program also effects grades or have other positive effects.

10

CONCLUSION

We explored the use of data mining to identify early school leavers without the schools collecting extra information. We can conclude that data mining on the information from the student information system offers great potential. Our results do not provide an automated system that is completely up to the task, but the results are promising. We expect that with more historical data and/or more features of the students information system (e.g. grades) available an automated system to detect potential dropouts can be created. We suggested that future research continues this exploration when more data is available.

BIBLIOGRAPHY

[1] Henrietta Ijeoma Alika. Bullying as a correlate of dropout from school among adolescents in delta state: implication for counselling. *Education*, 132(3):523, 2012.

[2] Jim Allen, Christoph Michael Meng, et al. *Voortijdige schoolverlaters: Aanleiding en gevolgen*. Research centre for education and the Labour Market, School of Business and Economics, Maastricht University, 2010.

[3] Sergey Bakin. *Adaptive regression and model selection in data mining problems*. PhD thesis, School of Mathematical Sciences, Australian National University, 1999.

[4] Marie Bienkowski, Mingyu Feng, and Barbara Means. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology*, pages 1–57, 2012.

[5] Alex J Bowers, Ryan Sprott, and Sherry A Taff. Do we know who will drop out?: A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 96(2):77–100, 2013.

[6] Gytha Burman, Carl Lamote, Karin Hannes, and J van Damme. Waarom verlaten jongeren vroegtijdig het secundair onderwijs. *Impuls*, 43:131–138, 2013.

[7] Sofie J Cabus and Kristof De Witte. Does school time matter?on the impact of compulsory education age on school dropout. *Economics of Education Review*, 30(6):1384–1398, 2011.

[8] Jan Cramer. The origins of logistic regression. Tinbergen Institute Discussion Papers 02-119/4, Tinbergen Institute, 2002.

[9] Gerben W Dekker, Mykola Pechenizkiy, and Jan M Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009.

[10] Hetty Dekkers and Geert Driessen. An evaluation of the educational priority policy in relation to early school leaving. *Studies in educational Evaluation*, 23(3):209–230, 1997.

[11] Şenol Zafer Erdoğan and Mehpare Timor. A data mining application in a student database. *Journal of aeronautics and space technologies*, 2(2):53–57, 2005.

[12] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[13] Jianqing Fan, Heng Peng, et al. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

[14] OECD (Organisation for Economic Co-operation and Development). Education at a glance 2012.

[15] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156, 1996.

[16] Philip Gleason and Mark Dynarski. Do we know whom to serve? issues in using risk factors to identify dropouts. *Journal of Education for Students Students Placed At Risk*, 7(1):25–41, 2002.

[17] Wim Groot and Henriëtte Maassen van den Brink. The health effects of education. *Economics of Education Review*, 26(2):186–200, 2007.

[18] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[19] Matthijs Kalmijn and Gerbert Kraaykamp. Dropout and downward mobility in the educational career: An event-history anaylsis of ethnic schooling differences in the netherlands. *Educational Research and Evaluation*, 9(3):265–287, 2003.

[20] Sotiris B Kotsiantis, CJ Pierrakeas, and Panayiotis E Pintelas. Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 267–274. Springer, 2003.

[21] Zlatko J Kovačić. Early prediction of student success: mining students enrolment data. In *Proceedings of Informing Science & IT Education Conference (InSITE)*, pages 647–665. Citeseer, 2010.

[22] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer, 2013.

[23] Eifred Markussen, Mari W. Frseth, and Nina Sandberg. Reaching for the unreachable: Identifying factors predicting early school leaving and non-completion in norwegian upper secondary education. *Scandinavian Journal of Educational Research*, 55(3):225–253, 2011.

[24] C. Marquez-Vera, C. R. Morales, and S. V. Soto. Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 8(1):7–14, Feb 2013.

[25] Valquiria Ribeiro De Carvalho Martinho, Clodoaldo Nunes, and Carlos Roberto Minussi. An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. In *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*, pages 159–166. IEEE, 2013.

[26] Mohammad Nurul Mustafa, Linkon Chowdhury, and MS Kamal. Students dropout prediction for intelligent system from tertiary level in developing country. In *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, pages 113–118. IEEE, 2012.

[27] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[28] Dutch Ministry of Education. New early school leavers [nieuwe voortijdig schoolverlaters]. 2015. `http://www.aanvalopschooluitval.nl/userfiles/file/cijferbijlage/VSV-Cijferbijlage_2015.pdf`.

[29] George Orwell. *Nineteen eighty-four*. Secker Warburg, 1949.

[30] Xavier Prats Monné. What is learning for? the promise of a better future. *European Journal of Education*, 2015.

[31] J Ross Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014.

[32] J Ross Quinlan et al. *Discovering rules by induction from large collections of examples*. Expert systems in the micro electronic age. Edinburgh University Press, 1979.

[33] Ross Quinlan. Data mining tools see5 and c5.0. 2004. `http://www.rulequest.com`.

[34] Ross Quinlan. Is see5/c5.0 better than c4.5? 2012. `http://www.rulequest.com`.

[35] Robert Rosenthal and Lenore Jacobson. Teachers'expectancies: Determinants of pupils'iq gains. *Psychological reports*, 19(1):115–118, 1966.

[36] Russell W Rumberger and Sun Ah Lim. Why students drop out of school: A review of 25 years of research. Technical report, California Dropout Research Project Report, 2008.

[37] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[38] Nicole Tieben and Maarten Wolbers. Success and failure in secondary education: socio-economic background effects on secondary school outcome in the netherlands, 1927–1998. *British Journal of Sociology of Education*, 31(3):277–290, 2010.

[39] Tanja Traag. *Early school leaving in the Netherlands: a multidisciplinary study of risk and protective factors explaining early school-leaving*. PhD thesis, Maastricht university, 2012.

[40] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 2011. Open Access.

[41] Peer van den Bouwhuijsen. Schooluitval aanpakken in zorg en welzijn. *Onderwijs en gezondheidszorg*, 35(2):12–14, 2011.

[42] William Robert Veitch. Identifying characteristics of high school dropouts: Data mining with a decision tree model. *Paper Presented at Annual Meeting of the American Educational Research Association, San Diego, CA, 2004 (ERIC Document No. ED490086).*

[43] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[44] Nengfeng Zhou and Ji Zhu. Group variable selection via a hierarchical lasso and its oracle property. *arXiv preprint arXiv:1006.2871*, 2010.