Identifying students at risk in Estonian K-12

Technical report
Irene-Angelica Chounta
Tartu, 2020

Table of Contents

1 Abstract

In this document,we report on the processes, analysis and outcomes of the project under CONTRACT FOR SERVICES No. 20-09 on dropout rates for Estonian K-12 Education. The scope of this project was to design computational algorithms to assess the risk that K-12 students may face for dropping out of their studies using students' data as recorded from the digital learning environment eKool in the framework of the project "Saved by the bell: early warning system for dropouts". To this end, we carried out an extensive analysis of existing data (as provided by eKool), engineered features to identify aspects that contribute to unsuccessful study year completion and designed five computational models using three modeling approaches to identify students who may be at-risk of not completing their study year. In the following, we present the analytical and modeling process and a comparison between the proposed modeling approaches and we discuss the use of such an approach as well as potential limitations and pitfalls.

2 Description of Data

The complete dataset was provided on July, 2nd 2020 and it contained data from 47 schools, 33953 unique student ids, and 123381 study year entries (that is, each academic year that a student was enrolled).

2.1 Data format

Data were logged in JSON format: one JSON file was provided for each school. Each JSON file contained information about the school itself (for example, the language of the school), information about the demographics of students enrolled in the respective school (for example, student's id, gender, home language), information about the study years they have attended (for example the class level, and academic year), information about their grades per class level and subject (for example, assessment, course and lesson grades), information about other performance metrics (for example, exams and behaviour assessments), information about their communication with school (that is, notes that the students received from the school), information about their participation (when they were absent and for what reason) and information about decisions or actions the school took related to the students (for example, enrolling or transferring a student).

2.2 Data on dropouts

As dropouts, we define students who do not complete their studies due to reasons that reveal a lack of interest, lack of motivation or unwillingness on behalf of the student, to continue with their studies. However, such kind of behaviour could not be determined based on the provided data. After discussion with the data provider (eKool), we received additional data about "failing" students. In this case, as failing, we identify students who failed to complete the academic year (study year) either because they were expelled from school or because they had to repeat the year due to various reasons. The reasons for which a student had to repeat an academic year were not contained in the provided data.

Overall, 609 students were recorded as failing. This accounts for 1.79% of the total number of students. In other words, 1.79% of students failed to complete the study year successfully either because they were expelled or because they had to repeat the study year. Out of these students, for 246 students there was information only about one study year and for 385 students there was information for two study years. Only for 165 students, there was information for more than 3 study years.

Due to the lack of data for distinguishing dropouts as described above, the analysis goal was adapted to identify factors and to design computational algorithms for assessing students who may be at-risk of completing their study year successfully.

2.3 Demographics

Out of the students who failed to complete the study year due to being expelled or forced to repeat it, 55% of were female, 85% were studying in schools where the primary language was Estonian and 11% came from homes where Estonian was the main (and only) language. Most of the students who failed to complete the study year were expelled or forced to repeat the year for class 10, with classes 11 and 12 following (Figure 1). With respect to the students who completed the most recent study year, as recorded in the dataset, and either have graduated or are continuing with their studies, 51% are female, 70% were studying in schools where Estonian was the primary language and 15% have Estonian as the main and only home language. Table 1 presents additional information regarding the demographic distribution of the initial population.

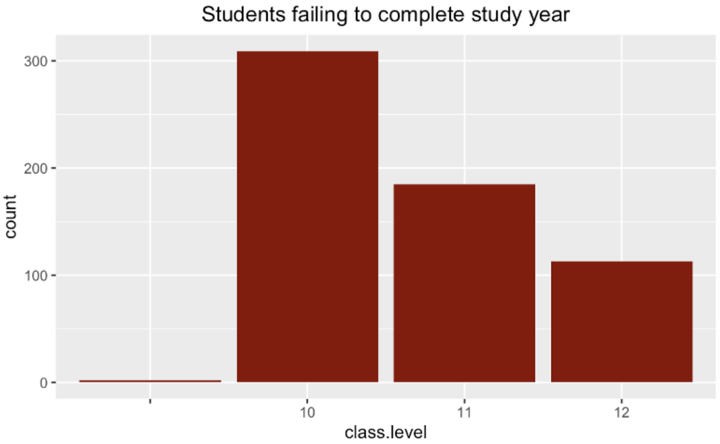

Figure 1: Distribution of students who failed to complete the study year over class levels

	Gender		School Language		Home Language		
	Female	Male	Estonian	Russian	Estonian	Russian	Both/Other
Completed	17018	16326	23228	10116	4977	673	27694
Not Completed	333	276	520	89	69	19	521

Table 1: Demographic information of the initial student population

3 Feature Engineering

To further explore how we can assess and support students who may be at risk of failing their studies, we analyzed the existing dataset on the study year level. This was taken for two reasons: For the majority of failing students in the existing dataset, there is not sufficient information for their school-history. In particular, for more than half of the failing students, recorded information covers, at a maximum, two study years. We want to avoid enforcing bias due to students' prior practice, taking into account that this is preliminary work and the amount of data is limited. To describe students' activity within a study year, we defined five log-based information categories:

- Demographics (6 features):

Feature	**Description**
Person.id (D1)	Personal identification number, unique for every student
Gender (D2)	Male or female
Class.level (D3)	Class level the student was attending
Academic.Year (D4)	The academic year (e.g. 2020-2021)
School.id (D5)	The school in which the student was enrolled
School.language (D6)	The school language (Russian or Estonian)

- School Decision-making (8 features):

Feature	**Description**
Decisions.total (SD1)	Number of school decisions the student received
Accepted (SD2)	Number of acceptances
Expelled (SD3)	Number of expulsions
Repeated (SD4)	Number of decisions to repeat the class
Transferred (SD5)	Number of times the student was transferred
Finished (SD6)	Number of times the student was acknowledged as finishing
Paralled (SD7)	Parallel studies
Other (SD8)	Other decisions

- School communication (4 features):

Feature	**Description**
Notes (SC1)	The number of notes the student received from the school
Positive (SC2)	The number of positive notes
Negative (SC3)	The number of negative notes
Neutral (SC4)	The number of notes tagged as neutral

- Performance (21 features) (see A.1:Table 4)
- Participation (3 features):

Feature	Description
Absences (P1)	All the student's absences during the study year
With.reason (P2)	Number of absences with a reason
Without.reason (P3)	Number of absences without a reason

4 Data Preparation

The data was curated to remove cases with missing information, such as null class levels (20 cases). We removed entries from the dataset for which there was no grade-related information available (6040 cases). For the cases where there was no information regarding school notifications (that is, notes that schools send to students), school decisions, and absences, we assumed - respectively - that the student received no notifications, no school decision was taken, and s/he was not absent during the study year. A set of features was removed from the dataset due to tautology. For example, expelled decisions (SD3) and finished decisions (SD6). The curated dataset contained 117321 data points with each data point representing a unique student study year. The curated dataset was further split into a training set and a test set following a 75% / 25% split with a balanced distribution of rare case. The training set was used for training the predictive models and the test set was used to evaluate the models on unseen data.

The training set consisted of 87990 data points (observations) that were recorded from the activity of 31534 unique students enrolled in 47 schools over 13 class levels and for a time span of 5 academic years. From the 87990 data points, 44512 referred to female students and 43478 to male students, 60407 to schools with Estonian as a primary language and 27583 to schools with Russian as a primary language.

The test set consisted of 29331 data points referring to the activity of 19809 unique students enrolled in 47 schools over 13 class levels during 5 academic years. From the 29331 data points, 14868 referred to female students and 14463 to male students, 20170 to schools with Estonian as a primary language and 9161 to schools with Russian as a primary language.

To avoid singularity and multicollinearity issues, we carried out a correlation analysis in order to remove features that were highly correlated. The final list of the engineered features is presented in Table 2.

Demographics	School Communication	Participation	School Decision-Making	Performance
Person.id (D1)	Positive (SC2)	With.reason (P2)	Accepted (SD2)	Assessment grades (sufficient/insufficient) (G2/G3)
Gender (D2)	Negative (SC3)	Without.reason (P3)	Paralled (SD7)	Course grades (sufficient/insufficient) (G6/G7)
Class.level (D3)	Neutral (SC4)		Other (SD8)	Behavior assessment (sufficient/insufficient) (G10/G11)
Academic.Year (D4)				Annual exam grade (sufficient/insufficient) (G14/G15)
School.id (D5)				Annual grade (sufficient/insufficient) (G16/G17)
School.language (D6)				

Table 2: The final set of features to be considered during the modeling process

5 Model Training

We modeled the task as a classification problem using three modeling algorithms:

1. Logistic Regression (LMER and GLM);
2. Survival Analysis (Cox Regression);
3. Decision Trees (Random Forest and Tree-based Classification).

The dependent variable (DV) - that is, the unsuccessful completion of the study year - was modeled as a function of the 20 features presented in Table 2. Next, we present the results of the modeling process and the evaluation of each model on the test set.

5.1 Logistic Regression

We used mixed-effects multiple logistic regression (LMER) where the unique identifier of the student, the school, the academic year, and the class level were all modeled as random effects. To cross-validate the findings, we additionally carried out a generalized lineal model logistic regression that does not take into account any random effects that may potentially impact the DV (for example, students belonging in the same school). The impact of each feature on the dependent variable (successful completion of the study year) as expressed by the regression coefficients is presented in section A.2 Table 5. The features that have a statistically significant impact on the study year's successful completion are followed by asterisks. The features that are not present in Table 3, were not found to have an impact on the unsuccessful completion of the study year.

To summarize, the results suggest that students who are frequently absent (either having a reason or not) are more likely to not complete the study year successfully than those who are not. Also, students who receive many notes of neutral or negative nature from the school may also be at-risk of not completing the study year in comparison to those who don't. Unsuccessful completion of the study year negatively relates to the number of school decisions classified as “other” they are involved in. This could potentially be attributed to the fact that the decisions classified as “other” usually relate to school events such as distinctions in sports events or competitions that one would not expect students who struggle to participate in. There was also a positive relation between unsuccessful completion and student enrollments (decisions classified as accepted) but this can be attributed to the fact that when a student gets expelled from one school then they will enroll to another. Unsuccessful completion also relates negatively to grade-related metrics such as the number of courses the student achieved a sufficient assessment in and a sufficient annual grade, which is expected. One surprising result was that unsuccessful completion was positively related with successful exam results - which may indicate that unsuccessful completion may stronger relate to other parameters (social and personal aspects) rather than performance. This is also suggested by the generalized linear model finding that negative behaviour assessments by the teacher have a positive impact on the DV - in other words, students who show behavioral problems tend not to complete the study year. However, the generalized linear model does not fit as well as the mixed-effects model and is thus only considered for validation purposes.

School language and gender did not have a significant impact on the DV.

5.2 Survival Analysis

For the survival analysis, we followed an iterative process of removing features that violated the proportionality assumption. We thus eventually used a set of 9 features, namely: the school's language, a student's gender, the number of unjustified (without reason) absences, the number of notes (tagged as positive, negative and neutral), the number of decisions (tagged in the system as, "paralled" and "other"), and the number of insufficient annual exam grades. The results of the survival analysis partly confirm the results of the regression analysis. Students who have been absent without reason are more likely to fail the study year as the rest of students. Also, students who are not frequently referenced in school decisions (tagged as "other") and students who do not receive positive communication from their school are more likely to fail the study year. Contrary to the regression analysis, the survival analysis showed that school language had a statistically significant impact on unsuccessful completion. In particular, students who are enrolled in schools with Estonian as a main language are more likely to fail the study year compared to students who are enrolled in schools with Russian as a main language (Figure 2). However, the survival analysis model had the worst fit to the data during training in comparison to the other models (section A.2:Table 5: Diagnostics) therefore, it is not advisable to draw conclusions based on these findings

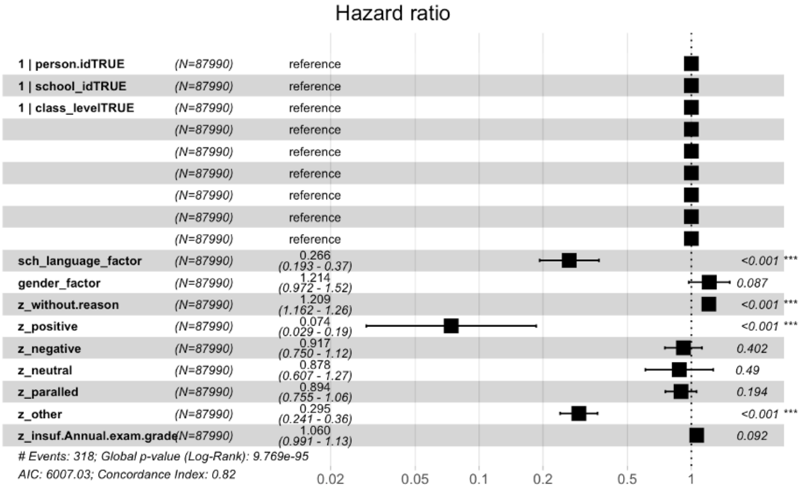

Figure 2: Forest plot of the Cox regression model

5.3 Decision Trees

For the Random Forest model, we used the same set of features as presented in Table 2 to train a binary classifier. During the training phase, 317 cases were miss-classified leading to an error rate of 36%. The error rate did not change significantly for lower and higher number of trees. The importance of the features regarding the mean decrease of accuracy and the mean decrease of the Gini coefficient are presented in Figure 3. The results indicate agreement with the findings of the logistic regression to some extent, showing that the students' absences, assessments (in terms of grades) and school communication (in terms of notes) are important factors that impact study year completion. Similar results were achieved by fitting a decision tree as a binary classifier, both in terms of classification error and in terms of features' importance. The decision tree binary classifier is presented in section A.3 Figure 4.

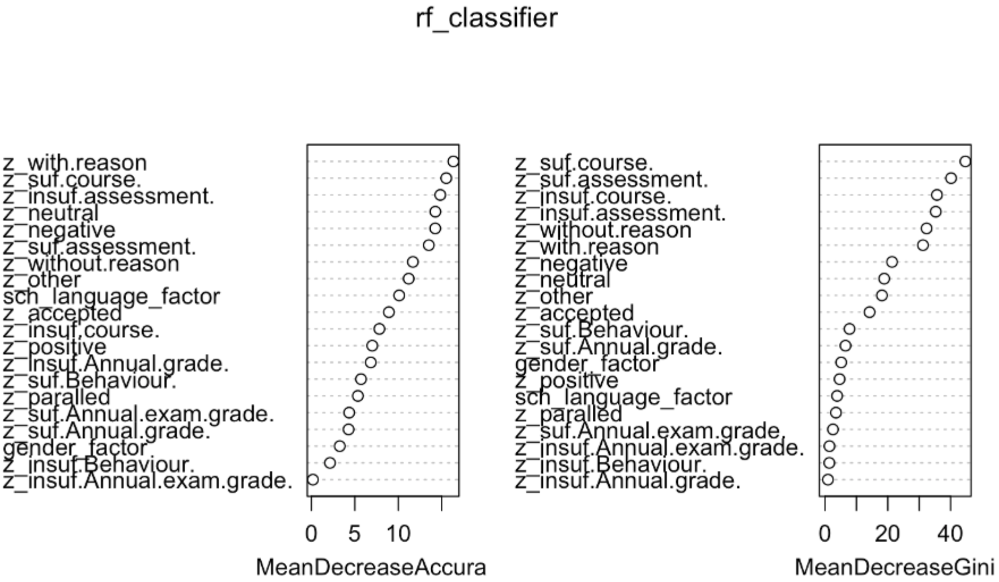

Figure 3: Features importance for the Random Forest Classifier

5.4 Feature Importance

Based on the outcomes of the model training, the following features were identified as informative or important to assess students' risk of failing their study year:

- absences: the number of students' absences (either with or without a reason) relates positively to the risk of failing to complete the study year; that is, the more absences for a student, the higher the risk;
- non-positive school notes: the more non-positive school notes a student receives, the higher the risk of failure. However, schools do not consistently classify the notes they

send to students. Therefore, extensive qualitative analysis of the notes' contents is required to explore this feature further;

- non-specific school decisions: the less a student is involved in non-specific ("other") school decisions, the higher the risk of failure. Usually, that kind of decisions relate to school events, contests and awards, therefore one may argue that students who struggle do not usually participate in that kind of events. As in the case of notes, schools do not consistently categorize these decisions. Therefore, further qualitative analysis of the decision-making processes is needed;
- insufficient course and annual grades: the lower the course and annual grades of students the higher the risk of failing. This is rather an unsurprising finding that could however be used towards supporting the accuracy of the predictive model.

Other potentially important features to further investigate are school language and behaviour/diligence assessments. School language was suggested as an important feature from the survival analysis: students who study in Estonian-language speaking schools are more likely to fail their study year than those who study in Russian-language speaking schools. However, this indication requires further investigation. Additionally, the analysis showed that students who do not complete their study year, have sufficient exam results. This could either indicate some discrepancy or inconsistency in the data or it could suggest that the main reason behind failure to complete the study year is social or personal. Therefore, further investigation of behavior and diligence assessments is required. From the existing data, it was evident that behavior and diligence assessments correlate with course assessments - in other words, students who perform well in their studies also get good behavior assessments. Even though this is expected to some extent, it does not allow insights regarding students' social and personal characteristics. Thus, additional data collection and analysis may be needed towards this direction.

6 Model testing

Here, we report the results from the evaluation of the models that were trained in the previous section. To evaluate the models, we used them to predict unsuccessful study year completion (that is, “positive” class = ‘1’) on unseen data (the test set). The evaluation metrics we use are the following: accuracy, precision, recall, F-value and Area Under the Curve (AUC) (Table 3). Along with these metrics, we provide the confusion matrices for the binary classifiers and the ROC curves as part of the project’s outputs. Overall, a model with high precision and recall would be the optimal solution. However, in this case (taking into account that the objective is to support students who may not complete the study year), good recall should be considered more important since the task is to identify students who may not complete the study year successfully. This means that the cost of missclassifying students who may fail as "safe" (false negative) is higher than missclassifying students who are "safe" as "failing" (false positive). Taking the evaluation metrics into account, the mixed-effects logistic regression model (LMER) outperforms the rest, although it still presents low recall. The mixed-effects logistic regression model, missclassified the majority of the relevant cases (62 out of 107) but classified correctly with good accuracy students who were able to finish the year successfully. On the other hand, the survival analysis model was able to identify the most relevant cases (students who did not complete the year successfully) and missclassified only a few of them (7 cases out of 107 in total). That being said, an ensemble of modeling approaches (for example, a rule-based, weighted combination of survival analysis and logistic regression) could be further explored as the optimal solution.

Model	Accuracy	Precision	Recall	F1-score	AUC
LMER	0.997	0.999	0.421	0.592	0.71
GLM	0.996	1.00	0.019	0.037	0.509
COX	0.496	0.494	0.935	0.647	0.714
RF	0.996	1.00	0.00	0.00	0.5
DT	0.997	1.00	0.18	0.302	0.589

Table 3: Comparison of the models prediction performance on unseen data

7 Limitations

We came across four main limitations regarding the data that may negatively affect feature engineering and modeling:

- the imbalanced data in terms of the DV (unsuccessful completion): Out of the 33953 unique students present in the data, only 609 students were recorded as failing (1.79%). Therefore, the outcomes of this study can provide some suggestions regarding the importance of log-based features in terms of computationally assessing students at-risk. However, do not allow the training a robust predictive model for generalized use;
- missing information: For the majority of the students who were identified as not successfully completing their studies, information was available for less than three academic years. This prevents longitudinal analysis of students' data. In 6040 cases the grades of students were completely missing, and state subjects were not categorized as required;
- inconsistent information between schools: the notation used to record student grades in the data is not consistent between schools. For example, some schools assign numerical grades (1 to 5) while other schools assign letters (A to F). Some schools use arithmetic signs (+) and (-), others use specific notation (MR, VH). Some schools use a special symbols ("x"), some schools leave blank fields (" "). This means that in order for a thorough analysis of the student grades over all the student population to take place, we need to establish a common notation after consulting with e-Kool and school stakeholders. Due to the limited time and resources at the current point, we resided to designing metrics based on "sufficient" and "insufficient" assessments. This inconsistency extends to the codes of state subjects and electives. Some schools provide codes for the state subjects and some others leave them blank. The same is the case with electives. Therefore, it is practically impossible without the existence of a universal scheme, to determine grades per subject and to provide a detailed assessment plan for schools altogether. The notation used to record notes and decisions is not consistent between schools. For example, some schools classify the notes using the "positive", "negative" and "neutral" categories while other schools place all notes as "neutral". The same stands for the "decisions" categories. Some schools classify the logged decisions using predefined tags while others use only the general category "other".
- language limitations: Further analysis of content (regarding notes, comments and decisions usually accompanying data) was not possible due to language limitations.

Further investigation and actions towards these four directions is expected to provide additional insights regarding context and rationale behind the factors that affect students' academic career and to contribute towards the development of generalizable algorithmic solutions for identifying students at-risk.

8 Highlights

- We focused on students who did not complete successfully the school year either because they were expelled or because they were forced to repeat it.
- The level of analysis in this study was the study year - overall, we analyzed 123381 cases that spread over 47 schools and 5 academic years.
- An initial descriptive analysis of the dataset showed that about 1.8% of the students are likely to fail completing the study year. This population is mostly comprised of female students who come from Estonian-speaking schools.
- To model students' success of completing the study year, we engineered 42 features representing 5 dimensions: demographics, school communication, school decision-making, participation and performance.
- We designed and trained five binary classifiers using machine learning to assess whether a student will complete the study year successfully.
- The features identified as important in terms of assessing study year completion were the amount of absences, the amount of school notes and the amount of relevant school-decisions involving the student.
- Student grades were associated with unsuccessful completion to some extent. In particular, course grades and annual grades indicated successful completion of the study year.
- A logistic regression binary classifier outperformed the other models in terms of accuracy and F1-score but survival analysis performed better in terms of recall.

A Appendix

A.1 Performance-related features

Feature	Description
All.grades (G1)	Number of grades that the student received
Assessment grades (sufficient/insufficient) (G2/G3)	Number of sufficient and insufficient assessment grades
Lesson grades (sufficient/insufficient) (G4/G5)	Number of sufficient and insufficient lesson grades
Course grades (sufficient/insufficient) (G6/G7)	Number of sufficient and insufficient course grades
Term grades (sufficient/insufficient) (G8/G9)	Number of sufficient and insufficient term grades
Behavior assessment (sufficient/insufficient) (G10/G11)	Number of sufficient and insufficient behavior assessments
Diligence assessment (sufficient/insufficient) (G12/G13)	Number of sufficient and insufficient diligence assessments
Annual exam grade (sufficient/insufficient) (G14/G15)	Number of sufficient and insufficient exam grades
Annual grade (sufficient/insufficient) (G16/G17)	Number of sufficient and insufficient annual grades
State-subjects grades (sufficient/insufficient) (G18/G19)	Number of sufficient and insufficient grades on state subjects
Elective-subjects grades (sufficient/insufficient) (G20/G21)	Number of sufficient and insufficient grades on elective subjects

Table 4: Features regarding students' performance during the study year

A.2 Regression coefficient for Logistic Regression models and Survival Analysis

	LMER	GLM	Cox
(Intercept)	-37.79	-10.82	
School language	0.46	-0.04	-1.32 ***
Gender	0.06	0.14	0.19
[absences]without.reason	0.14 **	-0.01	0.19 ***
[notes]positive	0.10	-2.00 ***	-2.61 ***
[notes]negative	0.18 *	-0.11	-0.09
[notes]neutral	0.91 ***	0.00	-0.13
[decision]paralled	-0.19	-0.28 **	-0.11
[decision]other	-0.61 ***	-1.47 ***	-1.22 ***
insuf.Annual.exam.grade.	-0.02	-0.04	0.06
[absences]with.reason	0.27 ***	0.14 **	
[decision]accepted	0.23 **	0.27 ***	
insuf.Behaviour.	0.03	0.12 **	
suf.Behaviour.	-0.12	-0.76 ***	
suf.Annual.exam.grade.	0.64 ***	0.15	
suf.assessment.	-0.12	-0.35 *	
insuf.assessment.	-0.06	0.08 *	
suf.course.	-1.99 ***	-0.26 *	
insuf.course.	-0.02	0.12 ***	
insuf.Annual.grade.	-7.69	-6.48	
suf.Annual.grade.	-1.25 **	-2.24 ***	
Diagnostics	AIC: 1521.9	AIC: 2654.6	AIC: 6009

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\cdot p < 0.1$

Table 5: Regression models' coefficients for mixed-effects logistic regression (LMER), generalized logistic regression (GLM) and Survival Analysis (Cox). The statistically significant coefficient are indicated by asterisks.

A.3 Pruned Decision-Tree Classifier

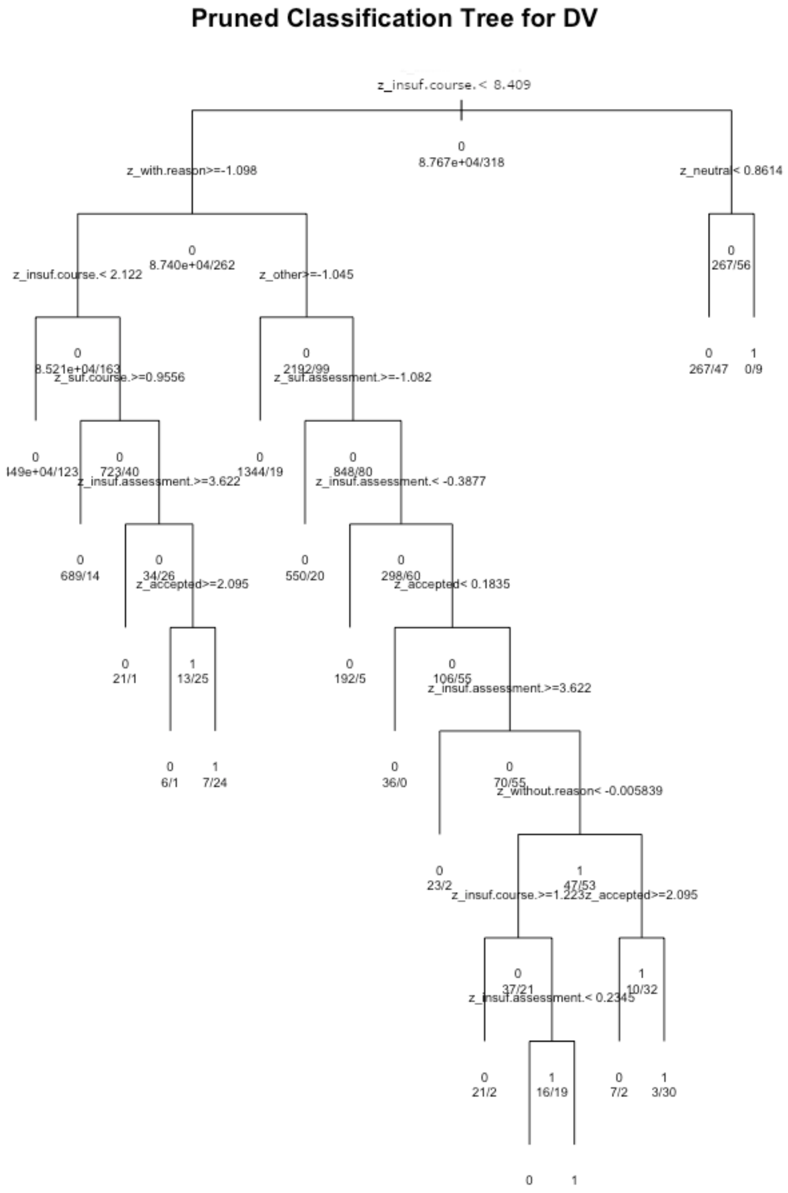

Figure 4: The pruned tree-based classifier for identifying students who may be at-risk of not completing the study year