Predicting Dropout in Higher Education based on Secondary School Performance

Marcell Nagy and Roland Molontay
MTA-BME Stochastics Research Group, Hungary
Department of Stochastics, Budapest University of Technology and Economics, Hungary

***Abstract*—Predicting student performance, preventing failure and identifying the factors influencing student dropout are issues that have attracted a great deal of research interest recently. In this study, we employ and evaluate several machine learning algorithms to identify students at-risk and predict student dropout of university programs based on the data available at the time of enrollment (secondary school performance, personal details). We also present a data-driven decision support platform for education directorate and stakeholders.**

The models are built on data of 15,825 undergraduate students from Budapest University of Technology and Economics enrolled between 2010 and 2017 and finished their undergraduate studies either by graduation or dropping out. We handle the problem of missing data by imputation. After performing feature extraction and feature selection, a wide range of classifiers have been trained including Decision Tree-based algorithms, Naive Bayes, k-NN, Linear Models and Deep Learning with different input settings. The methods were tested using 10-fold cross-validation and the AUC of the best models, Gradient Boosted Trees and Deep Learning, were 0.808 and 0.811 respectively.

I. Introduction

Early leaving and retention of students are crucial problems both in secondary and higher education all over the world. The early detection of at-risk students and identifying the main factors of dropping out have been in the focus of research for years [1]–[6]. Although graduation rates vary widely across countries and institutions, generally speaking, roughly every third student is dropped out from undergraduate or bachelor's programs [6]–[9], that is associated with considerable personal and social cost [10], [11]. The dropout rate from higher education in Hungary is particularly high, the rate of graduates of STEM[1] programs is one of lowest in the EU [6], [12].

Understanding, modeling and predicting student performance and academic progress have attracted tremendous amount of research interest in the last few decades [13]–[15]. Although the first related works date back to the '70-'80s [14], [15], the growing amount of data from educational institutes and the rise of data science have led to new directions of research in this field [16]–[18]. These data-driven approaches can be considered as pieces of a recently emerged scientific discipline, Educational Data Mining (EDM) [19], for great reviews of EDM we refer to [20], [21].

In this paper, we model students' final status (graduated or dropped out) using data of 15,285 undergraduate students enrolled at Budapest University of Technology and Economics between 2010 and 2017. We build predictive analytic models with the objective to provide data-driven decision support for education directorate by determining the key factors of dropout and accurately identifying at-risk students. Identifying these students early makes it possible to give them assistance and improve completion rates. In contrast to other works in this field, we have not used data from the students' performance of their first academic year, we have built the prediction only on their achievements from high-school and on some personal details (gender, age etc.). Building the models merely on the data available at the time of enrollment is definitely a more difficult problem, on the other hand, this approach has various advantages. First of all, students, who fail to graduate, usually drop out in their freshman year, that makes it extremely useful to have a predictive model as soon as possible. Furthermore we aspired to study the predictive power of secondary school performance on college graduation. Due to the structure of university admission process in Hungary, we have rich data on secondary school performance. To the best of our knowledge, this is the first predictive analytic study considering dropouts in Hungarian higher education. Our work is based on data of unusually high number of students spanning 7 years.

Our study relates to recent works, that reflects a growing interest in predicting dropout-prone students. Lovenoor et al. used classification methods on a large heterogeneous dataset (more than 32,500 student records) to predict dropout from University of Washington [7]. Erman et al. examined the prediction of dropouts through data mining approaches in an online program [18]. Delen and Thammasiri used machine learning techniques to predict whether a first-year student enrolls for the second semester, moreover they gave a great overview of the related works by comparing classification methods [16]. Neural networks have been tested by Lin et al. on data of 1,508 engineering students, using both cognitive and non-cognitive attributes [4]. Kovacic explored socio-demographic variables, that may influence dropout of students [5]. Anjana et al. used educational data mining techniques to analyze the factors affecting students performance that contributes to the prediction of their dropout [22]. Logistic regression models have been used by Burgos et al. to predict dropout from an e-learning system and designed a tutoring action plan to reduce the dropout rate [23]. Decision tree based classifiers have been also used extensively [24], [25], a review on educational dropout prediction is given in [26].

In this paper, we do not only focus on one method but our

[1]STEM stands for Science, Technology, Engineering and Mathematics.

978-1-5386-1122-7/18/$31.00 ©2018 IEEE

approach involves several predictive models of different types. The commonly used models for this problem in the literature are Decision Trees, Rule Induction, Naive Bayes, and (linear or logistic) regression models and rarely Neural networks [4], [7], [16], [24]–[26]. More complex methods are not preferred since most of the works investigate dropout only in small (usually homogeneous) populations, even the authors of [7], who have the largest known database on higher education attrition, did not use more sophisticated models. Besides the aforementioned models, here we train and test Deep Learning, and advanced Decision Tree based algorithms such as Random Forest and Gradient Boosted trees. After evaluating these models, we select the best two models by comparing their ROC curves and we optimize the hyperparameters (by grid and evolutionary optimization) to increase their performance. It is supposed that the reader is familiar with the mentioned machine learning algorithms and data science concepts, which can be found in [27], [28].

II. Background on the education system of Hungary

To follow the rest of the paper, it is important to understand the education system in Hungary, in particular the university admission process, in this section we give a brief overview.

In Hungary, secondary education consists of 4 (sometimes 5) years of schooling, preceding 8 years of elementary education and followed by higher education. Like several other countries, a five-point grading scale is used for grading, where (1) is the failing grade and (5) corresponds to excellent. At the end of the high school studies, students take a centralized exit exam called "érettségi", means maturity diploma or matura, that consists of five exams of core subjects: Mathematics, Hungarian Language and Literature, History, a chosen Foreign Language and one subject of the students' choice that they have been learning for at least 2 years. The students can decide whether they take the exam of a subject at normal or at advanced level.

The admission to higher education in Hungary mostly rely on the secondary school performance of the students and in particular the results of their matura. Students applying to colleges in Hungary gain an admission points score (APS) based on three factors: grades in secondary school and matura results (study points - SP), results of the maturity exams from two subjects required by the given university program (matura points - MP) and extra points (EP) for additional achievements (e.g. taking advanced-level matura exams, having certificate of a foreign language, earning a prestigious place in sport, art or academic competitions) and equal opportunity points (having disability, disadvantage, being on child care). Every bachelor's program requires matura exams of specified subjects, thus the aforementioned subject of students' choice may depend on the desired program (e.g. engineering bachelor programs usually require maths and a science subject).

There are two ways to calculate the admission point score, and the system automatically takes into account the one that is more advantageous for the student. The first way is $APS = SP + MP + EP$, and the other is the "doubling method": $APS = 2 \cdot MP + EP$. The composition of APS is as follows:

1) Study points (SP):
   - Two times the sum of the grades of the core subjects and a chosen science subject regarding the last two academic years when the subjects were studied. At most $2 \cdot 2 \cdot 5 \cdot 5 = 100$ points
   - The average results (in percentage) of the five matura exams. At most 100 points.
2) Matura points (MP):
   - Sum of the results (in percentage) of two certain matura exams required by the bachelor's program. At most 200 points.
3) Extra points (EP):
   - Advanced-level matura exam (+50 points per subject, only if the subject is used for calculating matura points)
   - Certificate of foreign language (a B2 certificate is worth 28 points, while a C1 certificate is worth 40 points)
   - Equal opportunities: (disadvantaged background, disability, being on child care) +40 points
   - Higher-level vocational training: depending on the results it may be worth 32, 20 or 10 points
   - Prestigious place in sport, art or academic competitions may be worth 10-100 points

   At most 100 extra points[2] can be gained.
   The maximum acquirable admission point score hence is at most 500 points.

Each university defines a minimal admission point score (MAPS) to its programs, and accepts those students whose APS are greater than the MAPS of the desired program. The admission program is based on the Gale-Shapely matching algorithm [29].

In the rest of this paper, we analyze the predictive power of the admission point score and its components on the college performance, in particular on dropping out.

III. Data preparation

The data have been provided by the Central Academic Office of Budapest University of Technology and Economics stored in the data warehouse of Neptun educational administration system. We received anonymized data of 15,825 undergraduate students enrolled between 2010 and 2017 regarding both their secondary school and university performance. Due to an upgrade of the administration system in 2012 some data were not restored before that, therefore we had to deal with a great amount of missing data. Some of the dropout prediction studies have to deal with the problem of imbalanced classification, meaning that there are much less available data of dropped out students than graduated ones [16], [30]. We do not face this problem due to the high dropout rate and due to fact that students who have started their university studies after

[2]Before 2012, advanced-level exams gave only 40 extra points, and maximum additionally acquirable points were 80, and APS was at most 480 points.

2014 could not graduate yet, meaning that from this student population we have more data of dropped out students.

After pivoting and joining the received sheets of student data into one single table and filtering students who graduated or dropped out, the main data preprocessing and cleaning tasks were as follows:

A. Handling missing data
B. Attribute transformation and generation
C. Dimension and redundancy reduction

For both data preparation and modeling, we used RapidMiner, a visual workflow designer data science tool [31].

A. Missing data

The absence of data occurred due to the aforementioned system upgrade and the multiple (outer) joins induced deficient rows. There are various possible ways to handle missing data, the simplest one is deletion and a more sophisticated method is imputation [32]. We have tested both methods, for imputation the k-NN algorithm was used with previously optimized parameters, and only the numeric variables were imputed. We trained and tested models using both approaches creating three datasets: one by deleting not complete records (filtered), one by imputing missing data (imputed) and one by imputing missing data but weighting the rows that were complete originally (weighted), for the results see Section IV.

B. Attributes

In general, there are four types of attributes by their reference: university program related (e.g. ID, faculty, date of application), high school related (name, ID, city), high school performance related (e.g. results of matura exams), personal details (address, gender, age). In addition, there is the binary attribute with the prediction role, the target variable, namely the final status that expresses whether the student graduated or dropped out. The attributes are summarized in Table I. In this section we overview the non-trivial attribute transformation steps and we discuss the most important attribute definitions and transformations. In order to obtain better performance we have generated new variables (e.g. 'Freshman or re-enrolled' attribute - for more details see III-B1 or the age of the students at time of enrollment). In Table I at the Sum of grades section, HS stands for High School, which distinguishes these variables from the matura exam results.

1) Student ID: Although, there is a student ID attribute in the original data sheets, we have generated a new ID because the student ID does not identify the records uniquely, since a student can be enrolled in different bachelor's programs, moreover the number of re-enrolling students are on the increase. Re-enrolling students are students who have already been dropped out of a program at least once, but successfully enrolled again [33]. For unique identification, we concatenated the student ID, program ID and the year of enrollment. In order to help the classification of the models, we generated three new binary attribute handling re-enrolling students, indicating if a student has ever been enrolled in the same program or in the same faculty or in any programs in the university.

TABLE I
SUMMARY OF DATA FIELDS

Feature Class	Feature name	Type
University program related	Student ID	Nominal
	Program ID	Nominal
	Year of enrollment	Date
	Faculty of program	Categorical
	Field of program	Categorical
	Financial Status	Binary
	Re-enrolled (3 versions)	Binary
Special roles: Label	Final status	Binary
ID	Concatenated ID	Nominal
High school related	High-school ID	Nominal
	Name of High School	Nominal
	Type of previous educational institute	Categorical
	City and county	Categorical
High school performance related	Admission points score	Real
	Study points	Real
	Matura points	Real
	Extra points	Real
	Base points (APS minus EP)	Real
Matura exam results (multiplicative and additive)	Hungarian Language and Literature	Real
	Mathematics	Real
	History	Real
	Foreign Language	Real
Sum of grades	Mathematics HS	Real
	Hungarian Language HS	Real
	Hungarian Literature HS	Real
	History HS	Real
	Foreign Language HS	Real
	Chosen science subject HS	Real
Certificate of foreign language	Language	Categorical
	Level	Categorical
	Consortium	Nominal
	Competition achievement	Binary
	APS calculating method	Binary
Personal details	Gender	Binary
	Date of birth	Date
	Address	Nominal

2) Advanced-level matura exams: Students can take normal level or advanced level matura exams, and as we mentioned before, advanced-level exams are rewarded with additional 50 points if the subject is used for calculating matura points (see Section II). The problem is, the percentage results of the normal level and advanced level exams are represented in the same way, but a student who achieves 70% in an advanced exam knows much more than a student reaches 70% in a normal level exam of the same subject. We introduced two new corrected attributes for each subject exam result:

- Additive attribute: +50 points for advanced level exam (this is how it is calculated in the admission process)
- Multiplicative attribute: 1.5 times the percentage result for advanced-level exams

We believe that the multiplicative method is a better representation of the students' real knowledge. In the dimension reduction and modelling period, the multiplicative solution turned out to be indeed more useful.

3) Foreign language attributes: There are two types of foreign language attributes: matura exams and state approved language certificates. Regarding matura exams, we have combined the different foreign language results into one single attribute (if a student took multiple exams of different foreign languages, then we used the average result). Regarding language certificates, since the majority of the students learn English or German language, we have categorized the rest of the languages into 'Other'.

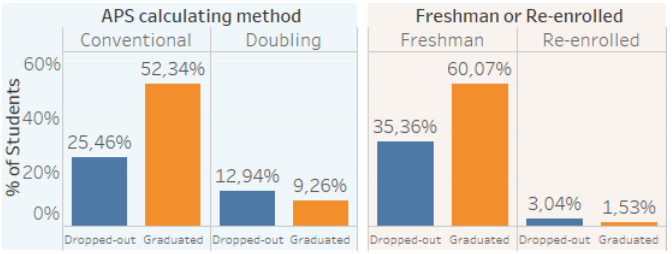


Fig. 1. Dropout rates of students enrolled between 2010 and 2012. Figure was created in Tableau.

C. Dimension reduction and feature importance

Predicting student failure at school is also known as the "one thousand factors problem", due to the large amounts of risk factors or characteristics of the students that can influence school failure [17], [34], thus the gathered data from this field usually have high dimensionality. On the other hand, the high number of attributes are usually redundant. After converting the non-numerical variables to numerical, we performed Principal Component Analysis (PCA) and we observed that the 95% of the variation of the dataset can be explained by 7 principal components indicating significant redundancy among the attributes. Since it is difficult to interpret the principal components, we performed Feature Selection (FS) algorithms (forward selection, backward elimination, evolutionary) to reduce the dimensionality [35]. Table II shows the most important features selected by the evolutionary algorithm and the attributes that have the highest global importance, i.e. the highest absolute correlation with the 'Status = Graduated' target variable, together with the results of two model dependent feature importance metrics (using Gradient Boosted Trees and Deep Learning). Note that categorical variables such as Program ID, Freshman or Re-enrolled and Financial Status also turned out to be important variables. Despite the fact that it is a technical university, performance in humanities (e.g. Hungarian Language and Literature, History and Foreign Language) turned out to be a surprisingly important factor on university performance. Furthermore, the students who strategically focused only on the two required subjects (mostly mathematics and a science subject for engineering programs) with the intention of taking advantage of the "doubling" APS calculating method, are more likely to drop-out than those students who have good performance also in humanities i.e. whose points were calculated in the traditional way (see Figure 1). That is why APS calculating method is also an important variable. The fact that this is the most negatively correlated attribute with the target variable also confirms that generalists have better chances than specialists. A thorough explanation of this phenomenon is beyond the scope of this study. Table II also includes the Re-enrolled attribute, that is because re-enrolled students are more likely to drop-out again (see Figure 1). The aforementioned discoveries may also have policy implications (e.g. changing the calculation of admission point score to reflect more the later university performance).

TABLE II
THE MOST IMPORTANT VARIABLES ACCORDING TO DIFFERENT ALGORITHMS

Method	Most important variables
Feature selection (evolutionary)	APS, SP, EP, Hungarian Lang. and Lit. (mult), Maths (mult.), Grades from HS, Program ID, Financial Status
Correlation (Global importance)	APS, Grades from HS, Maths (mult.), Hun.Lang. and Lit. (mult.), History (mult.), MP, SP, Calculating Method, Financial Status
Gradient Bosted Trees.	Program ID, Maths HS, History HS, APS, MP, SP, Maths (mult.), Financial Status, Maths (add) Hung. Lang. and Lit. (mult).
Deep Learning	Program ID, Re-enrolled, APS, Maths HS, Maths (mult.), Grades from HS, Financial Status, Hun.Lang. and Lit., Calculating method.

IV. Modeling and Evaluation

Our problem, formulated in the language of data science, is a binary classification of students, where the class (label) is the final status variable (graduated or dropped-out) and the explanatory variables consist of the regular (not special) attributes from Table I. For this classification problem we used the following machine learning models on the different datasets described in III-A: Decision Tree, Random Forest, Gradient Boosted Trees, Logistic Regression, Generalized Linear Model, Deep Learning.

The Deep Learning model has been tested with multiple hyperparameter settings, e.g. evaluated with different layer structures and activation functions with and without dropouts. Finally, we attained the best performances with the Rectifier activation function and with two hidden layers containing 50-50 nodes. Both Generalized Linear model and Gradient Boosted trees have been executed using H2O's algorithm. In order to achieve the best performance, we fine-tuned the hyperparameters of the models, e.g. Figure 2 shows the result of the grid optimization of the Decision Tree.

In our experiments, we evaluated the models with 10-fold cross-validation on stratified samples (random samples, such that the class distribution in the subsets is the same as in the whole dataset) of the data. The receiver operating characteristic (ROC) curves for the models are presented in Figure 3. Table III shows the prediction accuracy, recall, precision and AUC (area under curve) of the ROC curves of the models trained and evaluated on the imputed dataset. Note that the threshold was chosen to have better recall values at the expense of having slightly worse precision results. This decision depends on the desired application and on the cost of false positive and false negative errors. For example, for a web-application, described in Section V, this setting fits better, since it may be more harmful to discourage fresh students with an unjustified dropped-out prediction than to not identify at-risk students.

Although Deep Learning has the best performance providing that we use every available attribute (AUC = 0.790), Gradient Boosted Trees (GBT) outperform providing that we only use the variables selected by FS algorithm (AUC = 0.776). In case it is important to work with a small number of attributes (e.g. the web application of Section V) GBT has

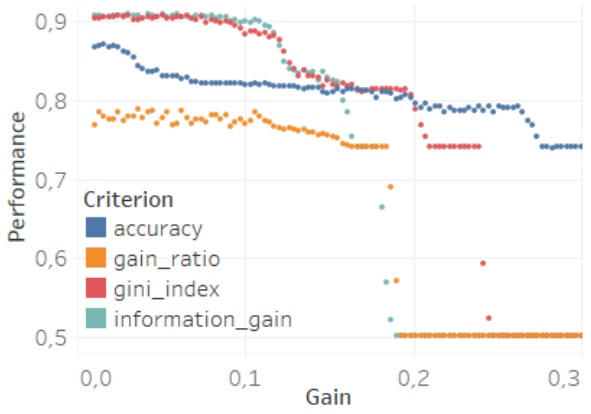


Fig. 2. Visualization of grid optimization of the parameters of the Decision Tree. The different colors indicate different settings of the 'criteria' parameter on which the attributes are selected for splitting. The x-axis corresponds to the 'minimal gain' above which a node splits. The figure was created in Tableau.

TABLE III
PERFORMANCE OF THE MODELS, USING THE IMPUTED DATASET WITH EVERY VARIABLE

Model	Accuracy	Recall	Precision	AUC
Decision Tree	63%	89.7%	59.6%	0.619
Random Forest	65.5%	85.6%	62%	0.736
Generalized Linear Model	67%	72.8%	66.5%	0.734
Naive Bayes	68.3%	80%	66%	0.756
Adaptive Boost	68.8%	70.1%	69.7%	0.746
k-NN	69%	78.3%	67.2%	0.759
Logistic Regression	70.3%	75.2%	69.7%	0.757
Gradient Boosted Trees	70.6%	75%	70.2%	0.769
Deep Learning	73.5%	82.4%	70.8%	0.811

found to be the best performing model. The performances of the models on the other two datasets are summarized in Table IV. The Deep Learning algorithm performs weakly on the filtered dataset, since it contains much fewer rows than the imputed/weighted datasets. To compare our results to related works, the accuracy of models on small and homogeneous data is around 95% [17], [22], however, the performance on a large heterogeneous dataset is 66% of accuracy and 0.729 of AUC [7]. Thus, even though we solved a more difficult problem (we did not used the students' result from their first year), we achieved excellent efficiency.

V. WEB APPLICATION

Once we have found the most influencing factors and the best performing model (that is GBT), we can use our model to predict the newcomer students' final status and identify at-risk students right at the time of enrollment in order to recommend tutoring sessions and offer other assistance. For this reason, with the help of Rapid Miner Server, we deployed a simple

TABLE IV
PERFORMANCES ON THE FILTERED AND WEIGHTED DATASETS

Models	**AUC**		**Accuracy**	
Dataset:	Filtered	Weighted	Filtered	Weighted
Gradient Boosted Trees	0.741	0.806	81.1%	73.3%
Deep Learning	0.740	0.806	73.8%	72.2%
Logistic Regression	0.742	0.763	80.8%	70.1%

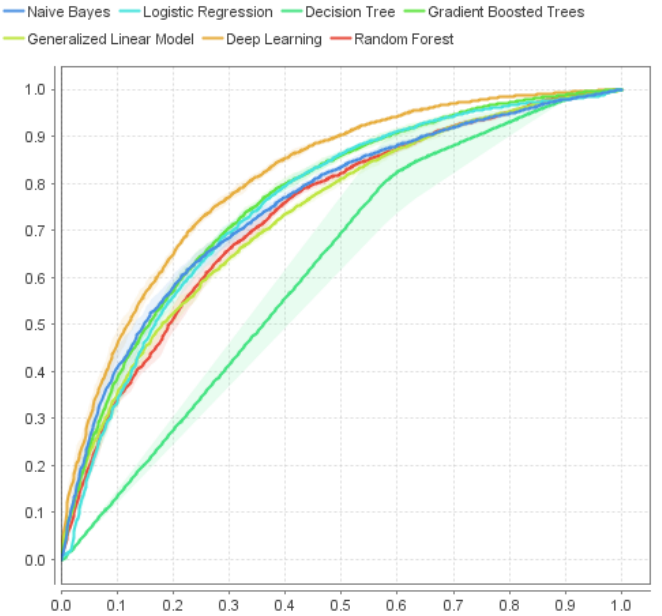
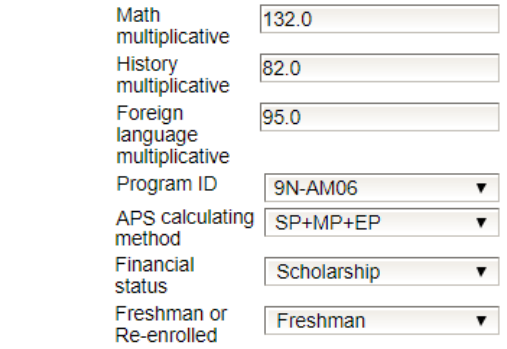


Fig. 3. Comparison of the ROC curves, using the imputed dataset with all the attributes

web-application, which is an online interface, where students can type their results from high-school and the application, based on the training dataset using Gradient Boosted Trees, predicts whether the student will graduate or not together with the confidence of the prediction that is a "continuous rating of positiveness". This application is aimed at both students and educational directorate. However, we emphasize that such application has to be handled cautiously, since predictive analytics applied to the behavior of human beings can cause ethical dilemmas [36]. Moreover, it can lead to negative self-fulfilling prophecies, i.e. the outcome of the application may affect the students' behavior e.g. a dropped-out prediction can make a student feel unfit for university and finally cause the student's drop-out. This problem can be avoided using recommendation systems instead of concrete predictions i.e. recommending personalized course and tutoring sessions from the relevant subjects to the at-risk students.

VI. CONCLUSION

In this work, we used advanced machine learning techniques including data imputation, feature selection and several ML models, in order to identify the factors influencing students' university performance and identify at-risk students. In contrast to other works in this field, we only used data from the students' achievements from high-school. We found that the current composition of Application Score Points has notable predictive power on college graduation. Furthermore, high school performance in humanities have surprisingly significant impact even on the university performance of engineering students. We also presented a data-driven decision support platform for education directorate and stakeholders, and suggested a solution to the possible ethical issues of such systems.



(a) Some input parameters

confidence(Graduated)	prediction(Status)
0.7056270445331028	Graduated

(b) The prediction with its confidence

Fig. 4. Screenshots from the Web Application

As a next step in research, we aim to improve the performance of the models by semi-supervised learning and using new attributes. In addition, we intend to carry out a recommendation system that suggests tutoring sessions and remedial courses for students based on their high school performance.

Acknowledgement

The research reported in this paper was supported by the BME - Artificial Intelligence FIKP grant of EMMI (BME FIKP-MI/SC) and by the EFOP-3.4.4-16 grant. The research of R. Molontay was partially supported by NKFIH K123782 research grant. The authors are grateful to the Central Academic Office for delivering the data this work is based on. We thank Mihály Szabó, Bálint Csabay and István Bognár for their assistance with data collection and data understanding.

References

[1] G. S. Abu-Oda and A. M. El-Halees, "Data mining in higher education: university student dropout case study," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 1, p. 15, 2015.

[2] D. S. Fike and R. Fike, "Predictors of first-year student retention in the community college," *Community college review*, vol. 36, no. 2, pp. 68–88, 2008.

[3] M. Yorke, *Leaving early: Undergraduate non-completion in higher education*. Routledge, 2004.

[4] J. Lin, P. Imbrie, and K. J. Reid, "Student retention modelling: An evaluation of different methods and their impact on prediction results," *Research in Engineering Education Sysmposium*, pp. 1–6, 2009.

[5] Z. Kovacic, "Early prediction of student success: Mining students' enrolment data." *Proceedings of Informing Science & IT Education Conference*, 2010.

[6] J. J. Vossensteyn, A. Kottmann, B. W. Jongbloed, F. Kaiser, L. Cremonini, B. Stensaker, E. Hovdhaugen, and S. Wollscheid, "Dropout and completion in higher education in Europe: Main report," 2015.

[7] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *arXiv preprint arXiv:1606.06364*, 2016.

[8] P. von Hippel and A. Quezada-Hofflinger, "The data revolution comes to higher education: Identifying students at risk of dropout in chile," *Social Science Research Network*, 2017.

[9] U. Heublein, "Student drop-out from German higher education institutions," *European Journal of Education*, vol. 49, no. 4, pp. 497–513, 2014.

[10] N. Raisman, "The cost of college attrition at four-year colleges & universities. policy perspectives." *Educational policy institute*, 2013.

[11] A. Latif, A. Choudhary, and A. Hammayun, "Economic effects of student dropouts: A comparative study," *Journal of Global Economics*, 2015.

[12] O. for Economic Co-operation and D. Staff, *Education at a glance: OECD indicators 2013*. OECD, 2013.

[13] A. Brezavšček, M. P. Bach, and A. Baggia, "Markov analysis of students performance and academic progress in higher education," *Organizacija*, vol. 50, no. 2, pp. 83–95, 2017.

[14] W. G. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis," *Interchange*, vol. 1, no. 1, pp. 64–85, 1970.

[15] J. P. Bean, "Dropouts and turnover: The synthesis and test of a causal model of student attrition," *Research in higher education*, vol. 12, no. 2, pp. 155–187, 1980.

[16] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Systems with Applications*, vol. 41, no. 2, pp. 321–330, 2014.

[17] C. M. Vera, C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," *IEEE Journal of Latin-American Learning Technologie*, vol. 8, no. 1, pp. 7–14, 2013.

[18] E. Yukselturk, S. Ozekes, and Y. K. Türel, "Predicting dropout student: an application of data mining methods in an online education program," *European Journal of Open, Distance and E-learning*, vol. 17, no. 1, pp. 118–133, 2014.

[19] EDM. (2011) Educational data mining society. [Online]. Available: http://educationaldatamining.org/

[20] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15 991–16 005, 2017.

[21] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning analytics*. Springer, 2014, pp. 61–75.

[22] A. Pradeep, S. Das, and J. J. Kizhekkethottam, "Students dropout factor prediction using EDM techniques," in *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*. IEEE, 2015, pp. 1–7.

[23] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students performance: A tutoring action plan to prevent academic dropout," *Computers & Electrical Engineering*, 2017.

[24] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study." *International Working Group on Educational Data Mining*, 2009.

[25] K. Shaleena and S. Paul, "Data mining techniques for predicting student performance," in *Engineering and Technology (ICETECH), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–3.

[26] M. Kumar, A. Singh, and D. Handa, "Literature survey on educational dropout prediction," *IJ Education and Management Engineering*, vol. 2, pp. 8–19, 2017.

[27] C. Sammut and G. I. Webb, *Encyclopedia of machine learning and data mining*. Springer, 2017.

[28] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[29] P. Biró, "Student admissions in hungary as gale and shapley envisaged," *University of Glasgow Technical Report TR-2008-291*, 2008.

[30] C. Márquez-Vera, A. Cano, C. Romero, and S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data," *Applied intelligence*, vol. 38, no. 3, pp. 315–330, 2013.

[31] M. Hofmann and R. Klinkenberg, *RapidMiner: Data mining use cases and business analytics applications*. CRC Press, 2013.

[32] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014, vol. 333.

[33] L. H. Willett, "Continuing education student flow analysis," *Research in Higher Education*, vol. 17, no. 2, pp. 155– 164, 1982.

[34] K. S. Sahedani and B. S. Reddy, "Forecast engineering students failure by using data mining techniques," *International Journal of Advance Engineer ing and Research Development (IJAERD)*, vol. 1, no. 5, 2014.

[35] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[36] S. Slade and P. Prinsloo, "Learning analytics: Ethical issues and dilemmas," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1510–1529, 2013.