Investigating the Impact of Outliers on Dropout Prediction in Higher Education

Daria Novoseltseva [1], Kerstin Wagner[2], Agathe Merceron[2], Petra Sauer[2], Nadine Jessel[3], Florence Sedes[1]

Abstract: Many institutions of higher education seek to reduce the dropout rate through the development of models which can detect students with a high risk of dropping out to provide specific advice for them. Classical models usually ignore the students-outliers with uncommon and inconsistent characteristics although they may show significant information to domain experts and affect the prediction models. The present paper provides an analysis of students' performance and aims to answer the following research questions: What kinds of students-outliers can be detected? Do outliers affect dropout prediction models? To answer the first question, students-outliers have been detected and their characteristics have been analyzed. To address the second question, the dropout prediction models have been compared in terms of different algorithms and the presence of outliers in the data. The results of the work indicate that the performance of prediction models, particularly in terms of recall, can be improved by removing outliers.

Keywords: Outlier Detection, Dropout Prediction, Machine Learning, Models' Comparison.

1 Introduction

Nowadays, one of the main research lines in the educational field is predicting students with a high risk of dropping out from their studies. As mentioned in [Au19], university attrition negatively affects social and economic aspects, such as misuse of public and private resources or social stigmatization. Therefore, domain experts and researchers seek to develop models to predict students with a high risk of dropping out to provide specific advice for them. Classical models tend to perform better for the majority of students with common and consistent characteristics. However, they ignore the students which are not aligned with the majority - outliers. Students-outliers may show significant information to domain experts and affect the prediction models.

The recent studies in dropout prediction [Au19, Be19, BRGS19] achieve good results, however, they do not mention how they deal with students-outliers with deviating characteristics. Thus, to close this gap, we investigate the impact of outliers on dropout prediction in higher education. This study aims at responding to research questions such

[1] IRIT, Paul Sabatier University, 118 Route de Narbonne, F-31062, Toulouse, France, {daria.novoseltseva, florence.sedes}@irit.fr , https://orcid.org/0000-0002-2987-1403

[2] Beuth University of Applied Sciences Berlin, Fachbereich VI, Luxemburger Straße 10, 13353 Berlin, Germany, {kerstin.wagner,merceron, sauer}@beuth-hochschule.de

[3] IRIT, Jean Jaurès University of Toulouse, 5 allées Antonio Machado, 31058 Toulouse, nadine.baptiste@irit.fr

as: RQ1. What kinds of outliers can be detected? Herein, we examine which students' characteristics are outlying; RQ2. Do outliers affect dropout prediction models? In this part, we investigate the impact on dropout prediction in two directions: the impact of outliers in general and the impact of specific classes of outliers.

The paper is organized as follows. Section 2 describes related work. Section 3 describes data collection. Sections 4 and 5 respond to the research questions and present experimental results. Finally, Section 6 concludes the paper and discusses the future work.

2 Related work

Outliers exist in almost every real-world dataset and can arise due to a variety of reasons [HA04]. An outlier is a data point that is very different from the rest of the data based on some measure. This measure strongly depends on the defined similarity/dissimilarity context. Many researchers try to avoid problems coming from outliers by removing them from the data. Meanwhile, another group of researchers believes that outliers are an important part of reality. Therefore, they try to retain such observations in their learning and testing samples [NV19].

In the educational field, outlier detection techniques attend to analyzing abnormal behavior or noisy reactions [BGS16]. According to [AASF19], only 2.25% of investigated papers are focused on outlier detection. For instance, in [Ca19] the authors perform outlier detection for student assessment in distance learning programs (e-learning) based on face-to-face exams. They define outliers as students, which do not use resources from the learning platform but still pass face-to-face exams, and detect them applying the isolation forest technique. In [ATEH12] the authors use unsupervised outlier detection algorithms (k-Nearest Neighbors and Local outlier factor) as part of educational data mining pipeline to improve graduate students' performance and overcome the problem of low grades of students who graduated. Their findings show that the detected outliers are students with excellent results in some degrees. In this current work, outliers have been defined as students who differ from the majority of students according to a certain set of features related to students' performance.

Dropout prediction in higher education is an active research topic that has been recently investigated in many studies [OM18]. The increasing rate of dropout negatively affects social aspects, such as reducing the number of people with higher education, and economic aspects, such as financing students who do not complete their studies as mentioned in [Au19, Be19]. To produce better results, existing works in dropout prediction examine various features (e.g., sociodemographic or performance), machine learning algorithms models (e.g., classification or clustering), and study cases (e.g., online courses or specific courses). Aulck et al. [Au19] predict graduation and re-enrollment of students in a US university. The authors use data of students after the first calendar year and achieve the best results with logistic regression. Furthermore, this study shows that including sociodemographic features in the prediction model hardly improves the results. Berens et al. [Be19] develop an early detection system to predict students' dropout in German universities. AdaBoost is implemented to improve the results of logistic regression, random forest, and neural networks. Baneres et al. [BRGS19] apply naive Bayes, decision tree, k-Nearest Neighbors, and support vector machines to performance features of students from a fully online university in order to identify at-risk students as soon as possible. All these works have good prediction results, however, they do not mention whether they deal with outliers. This work aims to understand which kinds of outliers can be detected and whether outliers impact dropout prediction models.

3 Data collection

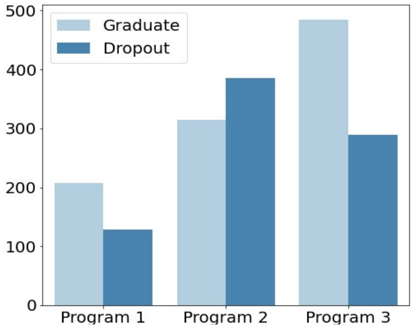
This work uses data of students' performance collected during their study in a six-semester bachelor's degree program at a German university. The data includes students from three educational programs, who started their study from winter 2012 to summer 2019, and have dropped out or graduated. Each degree program has a curriculum that contains a list of planned courses for each semester. Students may follow this curriculum or not. For instance, they can enroll in courses from the 1st semester when they are in the 2nd semester, and vice versa. The data records contain courses taken by each student including the earned grade and the respective semester as well as the semesters of graduation or dropout for each student. The grade scale for passing a course is [1.0; 1.7; 2.0; 2.3; 2.7; 3.0; 3.3; 3.7; 4.0], where the best grade is 1.0 and the worst is 4.0. If a student fails the examination, the grade is 5.0. Students may enroll in courses without taking the exam. To graduate, students must complete all mandatory courses and a program-specific number of electives with a maximum of three exam attempts.

To make the analyses applicable across study programs, we used global features, i. e. features that can be generated independently of study programs, as opposed to program-specific features e.g. achieved grade in the course *B01*. We considered features such as average grades and numbers of courses passed, failed, and enrolled per semester, similar to [Au19, Be19]. In terms of the number of courses passed, we made a further distinction: whether students passed them as planned or earlier, or later than planned. Empty values that have occurred during aggregation have been handled as follows: if no grade was obtained in the respective semester, the value of the grade is set to 6.0; if no course was passed, failed and/or enrolled in, the respective total is set to 0. Since the number of planned courses is program-specific (e.g. 5 planned courses in S2 for program 1, and 6 planned courses for program 2), the features related to courses were converted to proportions by division by the respective number of planned courses.

Outlier analysis and dropout prediction have been conducted for students after their 1st and 2nd semesters of study. Fig.1a depicts the analyzed features and their descriptive statistics. The codes of the features are presented as follows: **S1** and **S2** correspond to the 1st and 2nd semesters, **Av_grade** is the mean semester grade for passed courses, **F_ex** is the proportion of failed exams, **En_ex** is the proportion of courses which students enrolled in but did not take the exam, **P_ex_p**, **P_ex_d, P_ex_a** are the proportions of passed exams according to the plan, with delay, and in advance respectively.

In contrast to the data after the 1st semester, data after the 2nd semester reveals the development of performance over time, e.g. whether more or fewer courses were taken or whether exam grades improved. We removed students who dropped out after the 1st semester due to the absence of data for them in the 2nd semester. The considered dataset includes 1 809 students, among them, 1 007 students are labeled *Graduate*, and 802 students are labeled *Dropout*. The distribution of labels according to the educational program is shown in Fig. 1b.

	MIN	MEAN	STD	MAX
Av_grade S1	1	2.8	1.3	6
Av_grade S2	1	3.2	1.5	6
F_ex S1	0	0.067	0.13	0.8
F_ex S2	0	0.075	0.13	0.83
En_ex S1	0	0.17	0.28	2.8
En_ex S2	0	0.22	0.32	1.6
P_ex_p S1	0	0.7	0.38	1
P_ex_p S2	0	0.5	0.4	1
P_ex_d S1	0	0.041	0.13	1
P_ex_a S1	0	0.078	0.45	5.4
P_ex_a S2	0	0.074	0.16	3.5



a) b)

Fig. 1: Data description: a) descriptive statistics of analyzed features; b) distribution of labels Dropout/Graduate for three educational programs

4 RQ1: What kinds of outliers can be detected?

4.1 Methodology

The definition of outliers strongly depends on the context of the analysis. However, outliers always correspond to the two characteristics: (i) they are different from the norm with respect to their features; (ii) they are rare in a dataset compared to normal instances [GU16]. Therefore, such aspects as norm and rareness should be determined.

To specify the norm, the similarity metrics between objects have to be chosen. In the current work, we defined outliers as students which differ from the majority of students according to the set of numerical features described above. This is the task of finding global outliers - objects which are outlying for the entirety of the dataset in which it is contained. Furthermore, the ground truth indicating which students are outliers is absent and non-evident. Hence, only unsupervised algorithms can be used. Thus, we considered one of the most commonly used unsupervised algorithms to detect global outliers: distance-based k-Nearest Neighbors (kNN). The kNN outlier detection algorithm is described in [AP02] and is based on the idea that outliers are data points are distant from their neighbors and have sparse neighborhoods. The algorithm requires one predefined parameter – the number of nearest neighbors k, which should be in the range $10 \leq k \leq 50$ [GU16]. In our case, the outliers obtained with various values of k significantly intersected, therefore we focused on $k = 50$. Concerning the rareness aspect, we investigated two assumptions: (i) 5% of students are outliers; (ii) 10% of students are outliers. The outliers scores obtained by the kNN algorithm were ranked, and 5% or 10% of the students with top ranks were considered as outliers.

To understand the nature and the kind of the detected outliers, k-means clustering has been run, and the number of clusters has been chosen according to the elbow curve. The mean values of features for each cluster have been examined and compared, which allowed emphasizing the outlying characteristics.

4.2 Results

The distribution of labels (Dropout/Graduate) of detected outliers is as follows: 47/53% for the 5% assumption, 62/38% for the 10% assumption. Hence, the number of outliers with the dropout label is increasing when increasing the threshold of the assumption.

Running the k-means clustering algorithm, eight main clusters were obtained for both assumptions. The clusters for the two assumptions overlap significantly, and clusters 1 to 3 for the 5% assumption case are subsets of clusters 1 to 3 for 10% assumption. Therefore, we focus on the clusters obtained with the 10% assumption in this study. These clusters are presented in Tab.1. The table includes information about the number of outliers in each cluster (N), their labels (Dropout/Graduate), the number of outliers that are also detected as outliers in 5% assumption (N 5%), and mean cluster values for each feature. Although the features for outlier detection and clustering were standardized, Tab. 1 includes average values for non-standardized features of the clusters to help the interpretation:

- **Cluster 1 - Intense S1 and missed S2.** Students with an extremely high number of passed exams in S1, but without any passed exams in S2: the proportion of passed exams in S1 according to the plan is 0.64, while the proportion of passed exams ahead of the plan is 2.39. Meanwhile, the average grade in S2 is 6, which points out the fact that students from this cluster have not passed any exams in S2.
- **Cluster 2 - Intense S1.** Students with an extremely high number of passed exams in S1: the proportion of passed exams according to the plan is 0.99 and passed exams ahead of the plan is 2.99. Unlike cluster 1, this cluster contains students who have performance in S2. The majority of students from this class have the label graduate.
- **Cluster 3 - Intense S2.** Students with an extremely high number of passed exams in S2. The proportion of passed exams according to the plan is 1, the proportion of passed exams ahead of the plan is 2.75, the proportion of passed exams behind the plan is 0.83. In this small cluster all students have the label graduate.

- **Cluster 4 - Procrastination in S1.** Students with a bad performance in S1, which try to reestablish in S2: they pass exams from S1 in S2 with average marks. The labels of students in this cluster are mixed.
- **Cluster 5 - Average performance and intention to pass exams in advance.** Students with average grades during all semesters, who prefer to pass exams ahead of the plan in each semester instead of exams that correspond to the plan. The labels of students in this class are mixed.
- **Clusters 6-8 - Various types of bad performance.** These students belong to different clusters, and have various characteristics: some of them have low grades in all semesters, some have a high number of failed exams, others have a high number of enrollments without attending the exam. The majority of them have a dropout label.

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
N	14	22	2	33	35	29	25	21
Graduate	3	22	2	17	17	2	3	3
Dropout	11	0	0	16	18	27	22	18
N 5%	8	22	2	21	17	5	8	8
Av_grade_S1	2.31	2.28	4.58	5.85	2.28	4.08	4.01	4.69
Av_grade_S2	6.00	2.04	2.50	2.61	2.09	4.65	3.13	3.45
F_ex_S1	0.06	0.01	0.00	0.04	0.03	0.14	0.49	0.13
F_ex_S2	0.00	0.02	0.00	0.07	0.02	0.51	0.09	0.14
En_ex_S1	0.24	0.08	0.00	0.17	0.24	0.15	0.14	0.87
En_ex_S2	0.20	0.08	0.00	0.04	0.20	0.37	0.61	0.36
P_ex_p_S1	0.64	0.99	0.00	0.00	0.31	0.34	0.34	0.05
P_ex_p_S2	0.00	0.08	1.00	0.04	0.22	0.07	0.19	0.05
P_ex_d_S2	0.00	0.01	0.83	0.71	0.11	0.04	0.22	0.33
P_ex_a_S1	2.39	2.99	0.20	0.00	0.55	0.10	0.03	0.02
P_ex_a_S2	0.00	0.66	2.75	0.02	0.41	0.05	0.11	0.06

Tab. 1: Characteristics of clusters of outliers detected by kNN algorithm with 10% assumption

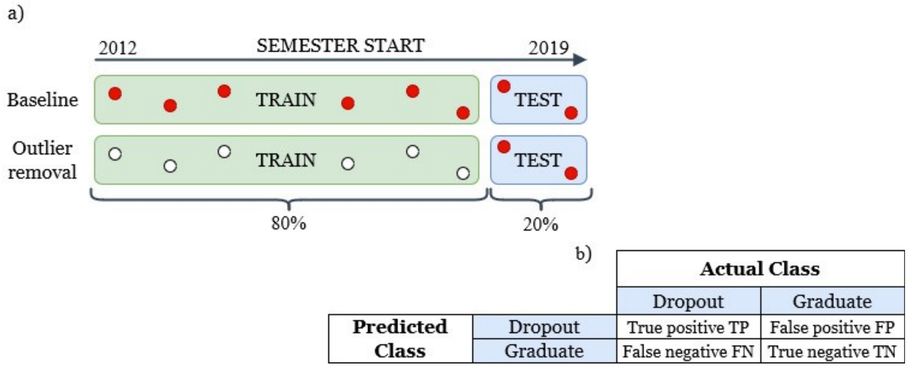
Students-outliers from clusters 1-3 have a high number of passed exams in advance. This can be explained by the recognition of courses from a former study when students changed their educational program. Thus, we denoted these clusters as specific classes and investigated their impact on dropout prediction as a separate case.

5 RQ2: Do outliers affect dropout prediction models?

5.1 Methodology

Using the data from the three undergraduate programs, we built 15 cross-program models, which predict dropout, where independent variables are global features (Fig.1a) and the dependent variable is the feature with *Dropout/Graduate* labels. The models based on different approaches to handle the outliers from the previous step (a. keep all outliers, b. remove all outliers from training data, c. remove only specific outlier clusters from training data) with different types of algorithms: decision tree (DT) and logistic regression (LR) to train interpretable models and random forest (RF) as an ensemble method that usually performs well. The analysis has been implemented using the Python scikit-learn library.

The dataset that keeps the outliers in the training data (approach a) served as the baseline to evaluate the outlier handling approaches. For approach b, regarding each assumption (5%, 10%) all outliers have been removed (datasets 2, 3), and for approach c, the outliers from specific clusters have been removed (datasets 4, 5). The detailed description of the analyzed datasets with the size of train and test sets is shown in Tab. 2.



		Actual Class	
		Dropout	Graduate
Predicted Class	Dropout	True positive TP	False positive FP
	Graduate	False negative FN	True negative TN

Fig. 2: Dropout prediction outlines: a) time-aware train/test splits b) the confusion matrix

As argued in [KMA19], time-aware nested cross-validation has been used to build the models to predict drop out: firstly, we sorted the students' data by their program start semester, then split without shuffling into training and test disjoint sets (80/20%), so that the test set contains the most recent students. Finally, we used the 10-fold cross-validation for the training set and took the best model in each case to predict dropout. The test set remained the same for each model and contained both inliers and outliers since outliers cannot be excluded from the prediction in a real-life scenario (362 students). The evaluation of the models included the following measures: recall, accuracy, precision, and area under the ROC curve. The detailed outlines of dropout prediction are depicted in Fig.2. In the context of our work, the most important metric is recall, since it returns a ratio of correctly predicted dropouts and the actual dropouts TP/(TP+FN). The higher the recall score, the more dropouts are correctly predicted.

5.2 Results

The results of the prediction models are presented in Tab. 2. The rows show the results for each considered dataset. The columns correspond to the prediction metrics and set size. The numbers in bold show best values per metric among considered datasets for each algorithm (DT, LR, RF). The colored numbers in italic correspond to the best values per metric across all models.

Models		Prediction metrics				Set size	
Alg	Dataset	Rec	Acc	Prec	Auc	Train	Test
DT	1 All data	82.07%	**83.15%**	**92.79%**	**83.83%**	1447	362
	2 w/o 5%	**83.27%**	81.77%	89.70%	80.82%	1384	362
	3 w/o 10%	80.88%	81.22%	91.03%	81.43%	1312	362
	4 w/o 5% clusters 1-3	80.88%	81.49%	91.44%	81.88%	1432	362
	5 w/o 10% clusters 1-3	80.88%	81.77%	91.86%	82.33%	1428	362
LR	1 All data	79.28%	84.53%	98.03%	87.84%	1447	362
	2 w/o 5%	**80.48%**	**85.36%**	***98.06%***	***88.44%***	1384	362
	3 w/o 10%	80.08%	84.81%	97.57%	87.79%	1312	362
	4 w/o 5% clusters 1-3	79.28%	84.25%	97.55%	87.39%	1432	362
	5 w/o 10% clusters 1-3	79.68%	84.53%	97.56%	87.59%	1428	362
RF	1 All data	81.27%	85.08%	**96.68%**	**87.48%**	1447	362
	2 w/o 5%	82.87%	83.70%	92.86%	84.23%	1384	362
	3 w/o 10%	***84.86%***	***86.46%***	95.09%	**87.48%**	1312	362
	4 w/o 5% clusters 1-3	81.67%	84.53%	95.35%	86.33%	1432	362
	5 w/o 10% clusters 1-3	82.87%	85.36%	95.41%	86.93%	1428	362

Tab. 2: Results of dropout prediction models for datasets with outliers (All data), without outliers (w/o), and for algorithms (decision tree (DT), logistic regression (LR), random forest (RF))

Training a model without outliers improves the performance of the model on some metrics. Among 12 columns (each metric for each algorithm), for 8 the best result is somewhere else in the column, not in the first line with dataset 1 (all data). Comparing with the model trained with all data (dataset 1), recall has improved: 1.2% for decision tree and for logistic regression, and 3.59% for random forest. The improvement of the other metrics is more modest. The best results are never obtained by removing the outliers from specific classes (datasets 4, 5). These outliers are characterized by a large number of courses passed ahead of the plan in semester 1 or 2. Most probably, these courses have been passed before enrolling formally in the degree and recognized after formal enrollment. Concerning the other kinds of outliers, there is no dataset that works better for all metrics. In terms of algorithms, one winner is random forest which has the best recall (84.86%) and accuracy (86.46%), another winner is logistic regression that shows the best values for precision (98.06%) and AUC ROC (88.44%).

6 Conclusion and future work

The present study investigates the impact of outliers on dropout prediction models in higher education. Analyzing the students' performance after the two first semesters, we aimed to answer the following research questions: RQ1. What kinds of outliers can be detected? RQ2. Do outliers affect dropout prediction models?

To answer the first research question, the students-outliers have been detected by an unsupervised outlier detection algorithm (kNN) with two assumptions (5% or 10% of students considered as outliers) and clustered (k-means algorithm). The first three clusters are characterized by high values of passed exams in the first two semesters. This can be explained by the recognition of courses that have been passed before enrolling formally (e.g. students who changed their educational program). Clusters 4 and 5 show other examples of rare performance: students who failed the first semester and then tried to reestablish as well as students who prefer courses from semesters going beyond the current curriculum. These clusters are interesting since they contain students who seem to have the same probability to graduate or to drop out. Some targeted intervention could be designed for such students. Students from clusters 6 to 8 have low marks, a high number of enrollments without attending exams, or a high number of failed exams. They tend to drop out.

To address the second research question, we build 15 cross-program dropout prediction models based on different approaches: a. keep all outliers (dataset 1), b. remove all outliers from training data (datasets 2, 3), c. remove only specific outlier clusters from training data (datasets 4, 5), using three different types of algorithms (decision tree, logistic regression, random forest). The models trained without all outliers have improved the performance with respect to some metrics, particularly for the recall, where the highest improvement is 3.59%. The best recall value is always achieved by models from approach b - removing all outliers from the training dataset.

Our results suggest that removing outliers could improve dropout prediction. Several limitations are noteworthy regarding this study, and they point out directions for future work. Firstly, the outlier detection has been performed using one unsupervised algorithm and with two threshold assumptions. The choice of the threshold is the most challenging task of outlier detection when the ground truth is unknown. Therefore, in future work, other thresholds as well as other outlier detection algorithms should be considered. Another limitation concerns the specific classes of outliers that have been removed from the training sets. The results of this study show that removing outliers from the clusters 1 to 3 does not improve dropout prediction. Thus, we intend to investigate how other kinds of outliers may impact dropout prediction. Regarding the dropout prediction, only three classification algorithms have been trained (decision tree, logistic regression, and random forest). However, in future work, more advanced algorithms such as AdaBoost or artificial neural networks can be applied. Moreover, we have investigated the impact of outliers on dropout prediction taking the scikit-learn standard values for the hyper-parameters of the models without feature preselection. The encouraging results of this work motivate us to improve the results through optimizing the hyper-parameters and selecting relevant features. Furthermore, similar analyses should be undertaken for data after the 1st, 3rd, and 4th semesters. These findings may enhance existing practices of dropout prediction in higher education. Furthermore, the outlier detection might be a part of the pipeline in the early warning system of detecting students with high risk of dropout.

Bibliography

[AASF19] Aldowah, Hanan; Al-Samarraie, Hosam; Fauzy, Wan Mohamad: Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 37:13–49, 4 2019.

[AP02] Angiulli, Fabrizio; Pizzuti, Clara: Fast outlier detection in high dimensional spaces. In: European conference on principles of data mining and knowledge discovery. Springer, pp.15–27, 2002.

[ATEH12] Abu Tair, Mohammed M; El-Halees, Alaa M: Mining educational data to improve students' performance: a case study. International Journal of Information, 2(2), 2012.

[Au19] Aulck, Lovenoor; Nambi, Dev; Velagapudi, Nishant; Blumenstock, Joshua; West, Jevin: Mining University Registrar Records to Predict First-Year Undergraduate Attrition International Educational Data Mining Society, 2019.

[Be19] Berens, Johannes; Schneider, Kerstin; Gortz, Simon; Oster, Simon; Burghoff, Julian et al.: Early Detection of Students at Risk-Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. JEDM| Journal of Educational Data Mining, 11(3):1–41, 2019.

[BGS16] Bansal, Rashi; Gaur, Nishant; Singh, Shailendra Narayan: Outlier Detection: Applications and techniques in Data Mining. 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), pp. 373–377, 2016.

[BRGS19] Baneres, David; Rodríguez-Gonzalez, M Elena; Serra, Montse: An early feedback prediction system for learners at-risk within a first-year higher education course. IEEE Transactions on Learning Technologies, 12(2):249–263, 2019.

[Ca19] Carneiro, Rubens E; Drapal, Patrícia; Fagundes, Roberta AA; Maciel, Alexandre MA; Rodrigues, Rodrigo L: Anomaly Detection on Student Assessment in E-Learning Environments. In: 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT). volume 2161. IEEE, pp. 168–169, 2019.

[GU16] Goldstein, Markus; Uchida, Seiichi: A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one, 11(4):e0152173, 2016.

[HA14] Hodge, Victoria J.; Austin, Jim: A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22(2):85–126, Oct 2004.

[KMA19] Krauss, Christopher; Merceron, Agathe; Arbanowski, Stefan: The timeliness deviation: A novel approach to evaluate educational recommender systems for closed-courses. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. pp.195–204, 2019.

[NV19] Nyitrai, Tamás; Virág, Miklós: The effects of handling outliers on the performance of bankruptcy prediction models. Socio-Economic Planning Sciences, 67:34–42, 2019.

[OM18] Ochoa, Xavier; Merceron, Agathe: Quantitative and Qualitative Analysis of the Learning Analytics and Knowledge Conference 2018. Journal of Learning Analytics, 5(3):154–166,2018.