# The Actionable Explanations for Student Success Prediction Models: A Benchmark Study on the Quality of Counterfactual Methods

Mustafa Cavus
Eskisehir Technical University, Department of Statistics, Eskisehir, Turkiye
mustafacavus@eskisehir.edu.tr

Jakub Kuzilek
Humboldt University of Berlin, Unter den Linden 6, Berlin, Germany
jakub.kuzilek@hu-berlin.de

## ABSTRACT

Digital transformation in higher education resulted in a surge of information technology solutions suited for the needs of academia. The massive use of digital technology in education leads to the production of vast amounts of education and learner-related data, enabling advanced data analysis methods to explore and support the learning processes. When focusing on supporting at-risk students, the dominant research focuses on predicting student success. Enabling prediction models to help at-risk students involves a reliable technical solution and a transparent and explainable solution to build trust among the target learners and educators. Counterfactual explanations (aka *counterfactuals*) from explainable machine learning tools promise to enable trustful explainable models, provided the features are actionable and causal. However, determining the most suitable counterfactual generation method for student success prediction models remains unexplored. This study evaluates standard counterfactual methods —Multi-Objective Counterfactual Explanations, Nearest Instance Counterfactual Explanations, and What-If Counterfactual Explanations. The methods are evaluated using a black-box machine learning model trained on the Open University Learning Analytics dataset, demonstrating their practical usefulness and suggesting concrete steps for model prediction alteration. Our results indicate that the Nearest Instance Counterfactual Explanation method based on the sparsity metric provides the best results regarding several quality criteria. Detailed statistical analysis finds statistically significant differences between all methods except the difference between the Nearest Instance Counterfactual Explanation and the Multi-Objective Counterfactual Explanation method, which suggests that the methods might be interchangeable in the context of the given dataset.

## Keywords

Counterfactual explanations, Explainable artificial intelligence, Contrastive explanations, Learning analytics

## 1. INTRODUCTION

The pace of digital transformation in higher education increased over the decade. With this increase, the data generated by the learners, lecturers, and educational institutions are multiplied. The data growth enabled the use of advanced Data Science methods for the analysis within the field of Learning Analytics [1]. With the extensive use of analytical tools in all areas of human life concerns about security and privacy emerged, resulting in new data protection regulations (e.g., GDPR in EU) [2]. Consequently, trust in advanced analytical tools and Machine Learning methods in higher education has been reduced. To overcome the distrust, a new approach called Trusted Learning Analytics emerged [3]. The TLA approach emphasizes using 'white box' Machine Learning (ML) methods and systems. Within this focus, the Explainable Artificial Intelligence (XAI) methods play a crucial role because they unlock the potential of the 'black box' models for use within the TLA systems [3].

A typical task in Learning Analytics (LA) is the predictive modelling of learner success, which enables identifying the learners needing help with their studies [4]. The ML model is trained with historical data collected within the same educational context. This particular model is then used as a trigger for educational intervention to support learners in need (i. e. [5], [6] or [7].

In the ML modelling process, black box models, known for their high predictive accuracy, are often preferred over interpretable models [8, 9, 10]. The XAI tools are primarily categorized into global and local. At the global level, they reveal which variables are important in the model. In contrast, at the local level, they answer questions about the contributions of variables in generating individual predictions [11, 12, 13]. However, commonly used global and local tools, while sufficient for understanding the prediction made for a particular observation, are not sufficient for generating a counterfactual understanding of an undesirable outcome. Therefore, counterfactual explanations have become popular, defined as the necessary changes in the values of variables to flip the model's prediction into the intended outcome [14]. Although student success prediction models may indicate an unfavorable prediction for a student, they do not generate output for reversing the student's situation. Using counterfactual explanations alongside such models is highly promising for addressing this issue. Students, teachers, and

curriculum designers are guided toward actions or measures to be taken through their generated explanations.

The use of counterfactual explanations in LA has been explored in several studies [15, 16, 17]. Yet, the focus of counterfactual explanations is in the frame of delivering actionable insights to the relevant stakeholders. None of the studies have investigated the quality of the generated counterfactual explanations. Facing numerous counterfactual explanations due to the nature of optimization problems requires selecting those explanations that fulfil specific criteria beneficial for the stakeholder. Because of their background, challenges, and needs differences, each learner requires personalized counterfactual [18]. Thus, several desired quality measures that a counterfactual explanation must satisfy.

To explore how the typical ML black box model trained for the predictive modelling of student success within the frame of TLA, we employed the open-access dataset Open University Learning Analytics Dataset (OULAD) [19] to answer the following research questions:

RQ1: *What is the most appropriate method for generating the counterfactual explanations?*

RQ2: *What is the most relevant quality measure of the methods for generating counterfactual explanations?*

This study compares the qualities of different counterfactual generation methods for students whose success prediction model developed on the OULAD anticipates failing. It is essential in two ways: (1) because the missing evaluation of the counterfactual quality can lead to inefficient explanations, and this may compromise their trustworthiness [20], and (2) there is no uniformly better method for each domain [21] and this is the first benchmark in the domain of LA.

The remainder of the paper introduces our approach for analysis and selecting the most appropriate counterfactual generation method followed by the results and their discussion. Finally, the conclusions are presented.

# 2. METHODS

## 2.1 Data

**Dataset.** We employed the OULAD dataset released by the Open University, the largest distance learning institution in the United Kingdom, to analyse counterfactual generating methods. The typical courses at OU take approximately nine months and consist of multiple assignments and a final exam. The most crucial assignments are Tutor Marked Assignments (TMAs), which represent milestones in the course schedule. The dataset contains data about learners' demographics, assessment results, and interaction with Moodle-like Learning Management System (LMS). For the analysis, we selected STEM course `FFF` and its presentation 2013J studied by 2283 students. The course contains five TMAs in weeks 2, 5, 13, 18, and 24. The last TMA was used as a target variable for model training. Learners can achieve scores from 0 to 100; we set a threshold for passing to 40 points. The following groups of students were excluded from the data set: actively withdrawn students (n = 675) and students who did not submit all TMAs (n = 500). The resulting

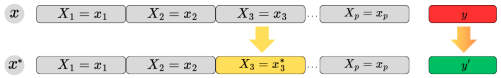


**Figure 1: An illustration of the counterfactual generation**

dataset contains the data of 1108 students. It consists of 14 predictors from which 6 of categorical variables are encoded numerically. The online interactions of learners with the LMS (i.e., 'n_clicks_xy' variables) have been computed for the top five most common activity types in the VLE, and they represent 95% of all student click-stream data. Table 1 presents the details of selected variables.

## 2.2 Counterfactual Explanations

Let $X = [x_1, x_2, ..., x_p]$ be a data matrix of $n$ observations from $p$ variables, and $y$ be the response vector. The goal is to find $f : X \rightarrow y$ that minimizes the expected value of the loss function $L$ in predictive modelling. A counterfactual $x' \in \mathbb{R}^p$ of an observation $x \in \mathbb{R}^p$ is calculated through an optimization problem:

$$argmin_{x' \in \mathbb{R}^p} L[f(x'), y'] + d(x, x') \quad (1)$$

where $\mathbb{R}^p$ denotes the p-dimensional real space, $L$ denotes a loss function that penalizes deviation of the prediction $f(x')$ from the interested outcome $y'$ and $d$, represents a distance function between the observation and its counterfactual. A counterfactual explanation can be briefly defined as the necessary changes in one or more than one variable to flip the model prediction. The distance function $d$ controls the distance between the target observation and the counterfactual. Figure 1 illustrates a counterfactual generation example. The value of the variable $X_3$ must be changed to $x_3'$ to flip the model's prediction $y$ to $y'$. To illustrate this in the context of the OULAD dataset: *An at-risk student can pass the course if the student increases assessment results or the total number of clicks in the discussion forum before the final exam.*

Counterfactuals aim to minimize the distance between the target observation and the counterfactual; however, there are more properties for a counterfactual explanation [22, 23]. **Sparsity** advocates for a minimal number of variable alterations, thereby maintaining its simplicity. **Minimality** focuses on the smallest possible changes in variable values. **Validity** is maintained by minimizing the disparity between the counterfactual instance, denoted as $x'$, and the observation $x$ while ensuring the model output aligns with the desired label $y'$. **Proximity** denotes the necessity of a slight divergence between the factual and counterfactual features. **Plausibility** mandates that counterfactual explanations remain realistic and adhere closely to the underlying data distribution. There are more than known 120 counterfactual generation methods; see [24] for details. However, we considered three commonly used counterfactual methods to make comparing the quality of counterfactuals feasible.

**What-if counterfactual explanations.** What-if method (WhatIf) finds the observations closest to the observation $x$ from the other observations in terms of Gower distance,

**Table 1: The details of the variables used to train our student success prediction model**

| Variable | Description | Class | Values |
|---|---|---|---|
| `gender` | student's gender | categorical | $\{0, 1\}$ |
| `region` | the geographic region, where the student lived while taking the module presentation | categorical | $\{1, 2, ..., 13\}$ |
| `education` | the highest student education level on entry to the module presentation | categorical | $\{1, 2, ..., 5\}$ |
| `imd_band` | the IMD band of the place where the student lived during the module presentation | categorical | $\{1, 2, ..., 10\}$ |
| `age_band` | a band of student's age | categorical | $\{1, 2, 3\}$ |
| `num_of_prev_attempts` | the number of how many times the student has attempted this module | numeric | $\{0, 1, ..., 4\}$ |
| `credits` | the total number of credits for the modules the student is currently studying | numeric | $[60, 360]$ |
| `disability` | indicates whether the student has declared a disability | categorical | $\{0, 1\}$ |
| `assessment_results` | the weighted sum of all previous assessments $\sum_{i=1}^{4} w_i a_n$ where $w_i = (0.125, 0.125, 0.250, 0.250)$ is the vector of weights T = (0.125,0.125,0.25,0.25) is the vector of corresponding weights | numeric | $[24.25, 72.75]$ |
| `n_clicks_disc_forum` | the number of clicks on discussion forum | numeric | $[0, 7670]$ |
| `n_clicks_disc_hpage` | the number of clicks on discussion homepage | numeric | $[4, 3150]$ |
| `n_clicks_assignments` | the number of clicks on assignments | numeric | $[0, 7193]$ |
| `n_clicks_quizzes` | the number of clicks on quizzes | numeric | $[0, 4857]$ |
| `n_clicks_course_page` | the number of clicks on course page | numeric | $[0, 1196]$ |

solving the following optimization problem [25]:

$$x' \in argmin_{x \in X} d(x, x'). \quad (2)$$

**Multi-objective counterfactual explanations.** The multi-objective counterfactual explanations method (MOC) objects to find counterfactuals corresponding to the validity, proximity, sparsity, and plausibility of solving a multi-objective optimization problem [26]:

$$x' \in min_x[o_v(\hat{f}(x), y'), o_p(x, x'), o_s(x, x'), o_{pl}(x, X)] \quad (3)$$

where the objectives correspond to the desired properties, *validity, proximity, sparsity, plausibility*, respectively. Thus, it generates valid, proximal, sparse, and plausible counterfactuals.

**Nearest instance counterfactual explanations.** The nearest instance counterfactual explanations method (NICE) finds the observations most similar to the observation in terms of the heterogenous Euclidean overlap method [27]. Because of the NICE method, there are two options in the object function based on the properties *proximity*, and *sparsity*, it can be used in these two ways.

The WhatIf method generates valid, proximal, and plausible counterfactuals. It is shown that the MOC method generates more counterfactuals than other counterfactual methods that are closer to the training data and require fewer feature changes [26]. Moreover, NICE generates the proximity counterfactuals. However, there is no uniformly better method in the datasets from different domains [21]. Thus, evaluating the quality of the generated counterfactual is necessary, and we conduct the experiments in the following section.

## 2.3 Experiment design

This study focuses on which method provides the highest quality counterfactual explanations for the student success prediction model trained using the OULAD dataset. Thus, our approach is (1) selecting the most appropriate ML model, (2) generating the counterfactuals, and (3) producing the evaluation criteria.

**Modeling.** We used `forester` [28] for model selection and hyperparameter optimization. It is an AutoML tool that adjusts the hyperparameters of tree-based models using Bayesian optimization. The reason for using this tool instead of manual modelling is its ability to make Bayesian optimization highly practical with its relevant parameters. Additionally, the fact that tree-based models exhibit lower prediction performance than alternative complex models in classifying tabular datasets [29] supports the idea that using this tool does not limit model selection. The number of optimization rounds `bayes_iter` is taken as 5, and the number of trained models `random_evals` is taken as 10 in the AutoML tool, respectively. forester returns 28 models, including decision trees, random forests, XGBoost, LightGBM, and their fine-tuned versions with Bayesian optimization and random search in Table 2. Because the best-performing one is a fine-tuned random forest model with random search —accuracy 0.900, AUC 0.771, and F1 0.946— the counterfactuals are generated on it.

**Counterfactual generation.** We used `counterfactuals` package [21] to generate the counterfactual explanations for the at-risk students using the counterfactual generation methods WhatIf, proximity-based NICE (NICE_pr), sparsity-based NICE (NICE_sp), and MOC. The non-actionable variables that are impossible to change are kept constant, such as `gender, disability, region, age_band, education, imd_band, num_of_prev_attempts, cummulative_assessment_results`. The MOC, NICE_pr, NICE_sp, and WhatIf methods generate 191, 39, 19, and 120 counterfactuals for the 12

**Table 2: The best score test table of `forester`**

| No | Name | Engine | Tuning | Accuracy | AUC | F1 |
|---|---|---|---|---|---|---|
| 1 | ranger_RS_3 | ranger | random_search | 0.900 | 0.771 | 0.946 |
| 2 | xgboost_RS_3 | xgboost | random_search | 0.900 | 0.801 | 0.946 |
| 3 | lightgbm_RS_1 | lightgbm | random_search | 0.900 | 0.787 | 0.946 |
| 4 | xgboost_bayes | xgboost | bayes_opt | 0.900 | 0.753 | 0.946 |
| 5 | decision_tree_bayes | decision_tree | bayes_opt | 0.900 | 0.809 | 0.945 |
| 6 | lightgbm_bayes | lightgbm | bayes_opt | 0.900 | 0.745 | 0.945 |
| 7 | ranger_model | ranger | basic | 0.892 | 0.726 | 0.942 |
| ... | ... | ... | ... | ... | ... | |
| 28 | xgboost_RS_4 | xgboost | random_search | 0.092 | 0.190 | 0.086 |

failed students predicted by the student success prediction model. It is essential to compare the counterfactual generation methods in terms of the number of generated counterfactuals because it shows the diversity of alternative ways to flip the model decision. The higher number of counterfactuals is better. The materials for reproducing the experiments performed and the dataset are accessible in the following anonymized repository: `https://github.com/mcavs/HEXED2024_paper`.

## 3. RESULTS AND DISCUSSION

The quality metrics *minimality, plausibility, proximity, sparsity, validity* are calculated to evaluate the generated counterfactuals by the methods WhatIf, NICE_pr, NICE_sp, and MOC. It should be highlighted that the lower values are better for each metric. Some user studies have shown that the users prefer to use the counterfactuals, which perform well on the criteria in [30, 31]. Thus, we compared their qualities in two steps. First, we used the average values and the standard deviations of these metrics given in Table 3, and second, we compared the distribution of the results in Figure 2.

It is seen that the quality of counterfactuals is quite good in terms of proximity, plausibility, and validity. However, the results are not promising for WhatIf in minimality and sparsity. It is expected because it is known the WhatIf method generates valid, proximal, and plausible counterfactuals. Therefore, we do not recommend using this method in this domain. On the other hand, counterfactuals generated by the NICE method that optimizes based on sparsity showed better results in sparsity and other quality metrics than the one that optimizes based on proximity. There are differences between the NICE_pr and NICE_sp in terms of minimality and sparsity. NICE_sp shows better performance because it optimizes based on sparsity and the metrics sparsity and minimality are quite related metrics. Sparsity refers to the changes in the number of variables while minimality refers the the smallest possible changes in the variable values. Therefore, using the NICE_sp method may be preferred to obtain better-quality explanations in this domain. Although the MOC method shows results competing with NICE_sp, it is poor on average.

Figure 2 shows the distribution of the quality metrics of the counterfactuals, providing deeper insights. The WhatIf method appears to produce explanations that are not minimal compared to the others. Although the NICE_pr was better than the WhatIf method in this regard, it performed worse than the other methods. When the methods are compared in terms of plausibility, it is seen that the WhatIf is better than the others, but the difference is low. While the WhatIf method produced fewer proximity explanations, other methods produced proximity explanations at a similar level. A similar pattern against the WhatIf has also been observed for sparsity. As expected, the NICE_sp method shows the best performance in terms of sparsity. Surprisingly, no method other than the MOC produced non-validity explanations. This is the most problematic quality feature for the MOC. The intriguing observation is the quality of counterfactuals generated by the MOC is better than the NICE_pr in terms of proximity, even though the NICE_pr method aims to create the proximity counterfactuals.

In summary, the quality of the explanations produced by the methods compete with each other in terms of both average and distribution properties, and it is not possible to say that the NICE_sp method produces the best quality explanations based on visual outputs alone. Therefore, using the Kruskal-Wallis test and the pairwise Wilcoxon test, we statistically test whether the explanations made by the methods differ. A Kruskal-Wallis test was performed on the quality metric values of the four methods (MOC, NICE_pr, NICE_sp, and WhatIf). The differences between the rank totals of the methods were significant, $\chi^2_{(4)} = 48.823$, $p < .001$. Post hoc comparisons were conducted using Wilcoxon Tests with a Benjamini-Hochberg adjusted alpha level of .016. The difference between the MOC and NICE_pr was no statistically significant ($p = .115$). The other comparisons were significant. The results of the statistical tests support the previous results.

**Table 3: The averages and standard deviations of the quality metrics for the methods**

| Metric | MOC | NICE_pr | NICE_sp | WhatIf |
|---|---|---|---|---|
| minimality | 0.07 ± 0.36 | 0.71 ± 0.94 | 0 | 7.83 ± 1.26 |
| plausibility | 0.06 ± 0.03 | 0.04 ± 0.02 | 0.04 ± 0.02 | 0 |
| proximity | 0.02 ± 0.03 | 0.02 ± 0.01 | 0.02 ± 0.01 | 0.10 ± 0.03 |
| sparsity | 1.62 ± 0.83 | 1.95 ± 1.10 | 1 | 8.69 ± 1.25 |
| validity | 0.07 ± 0.05 | 0 | 0 | 0 |

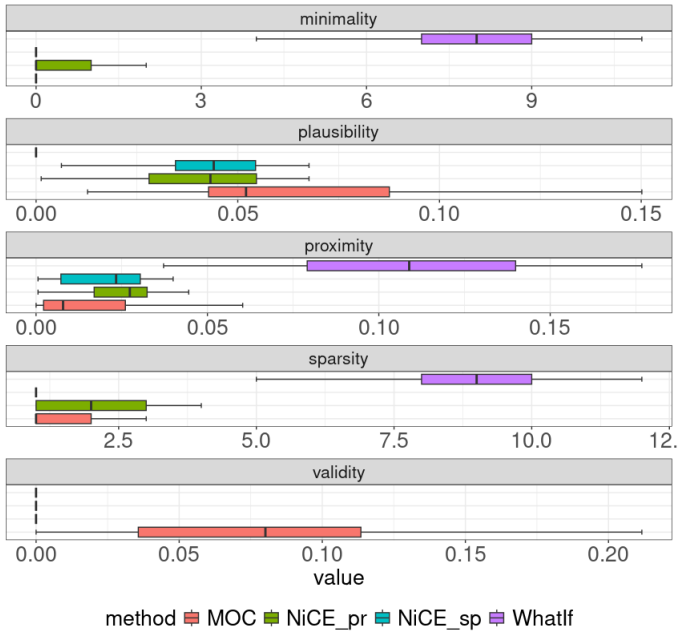


**Figure 2: The distributions of the quality metrics for the methods**

## 4. CONCLUSIONS

In this study, we explored the possibilities of using XAI tools in the frame of the TLA research. Our research focused on deploying the counterfactual explanation methods on the OULAD dataset containing the demographics, results and learner interactions with LMS to answer the following research questions: 1) *What is the most appropriate method for generating the counterfactual explanations?* Selection of the most suitable method depends on the stakeholder requirements and the educational context. However, selecting the most appropriate methods is generally guided by evaluating standard counterfactual properties: Sparsity, Validity, Proximity, and Plausibility. The evaluation of our approach on the OULAD dataset resulted in the finding that explanations generated using the NICE method based on sparsity are of higher quality in terms of all considered metrics than explanations generated through other methods (Table 3). 2) *What is the most relevant quality measure of the methods for generating counterfactual explanations?* As mentioned before, selecting a method depends highly on the educational setting. Yet, it might be defined by the relevant stakeholder as the most essential criteria chosen from those used as a standard evaluation measure. In addition, the statistical hypothesis testing results indicate no statistically significant difference between the Nearest Instance Counterfactual Explanation and the Multi-Objective Counterfactual Explanations method, which indicates the requirement for the deep validation of generated counterfactual explanations for the at-risk students to avoid misconceptions. This suggests that the human-in-the-loop is needed even when selecting the most optimal method in technical validation. In addition, the counterfactuals provide a simple way to understand and uncover the issues about learner learning and open the path to recommendations for possible educational interventions. Finally, the study has some limitations. Due to the focus of the study, data drift was not considered, and only the most common counterfactual explanation methods were used. Furthermore, we believe that conducting qualitative studies and evaluating the explanations solely based on quality metrics would provide further validation for the findings.

## Acknowledgments

The work in this paper is supported by the German Federal Ministry of Education and Research (BMBF), grant no. 16DHBKI045.

## 5. REFERENCES

[1] George Siemens and Ryan SJ d Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254, 2012.

[2] Tore Hoel, Dai Griffiths, and Weiqin Chen. The influence of data protection and privacy frameworks on the design of learning analytics systems. In *Proceedings of the seventh international learning analytics & knowledge conference*, pages 243–252, 2017.

[3] Hendrik Drachsler. *Trusted learning analytics*. Universität Hamburg, 2018.

[4] Zacharoula Papamitsiou and Anastasios A Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4):49–64, 2014.

[5] Kimberly E Arnold and Matthew D Pistilli. Course signals at purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 267–270, 2012.

[6] Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Salem Alelyani, and Raheel Nawaz. Predicting the academic performance of students from vle big data using deep learning models. *Computers in Human behavior*, 104:106189, 2020.

[7] Muhammad Adnan, Asad Habib, Jawad Ashraf, Shafaq Mussadiq, Arsalan Ali Raza, Muhammad Abid, Maryam Bashir, and Sana Ullah Khan. Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access*, 9:7519–7539, 2021.

[8] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[9] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory model analysis: explore, explain, and examine predictive models*. Chapman and Hall/CRC, 2021.

[10] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable ai methods-a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers*, pages 13–38. Springer, 2022.

[11] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[12] Aditya Bhattacharya. *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd, 2022.

[13] Mustafa Cavus, Adrian Stando, and Przemyslaw Biecek. Glocal explanations of expected goal models in soccer. *arXiv preprint arXiv:2308.15559*, 2023.

[14] André Artelt and Barbara Hammer. On the computation of counterfactual explanations–a survey. *arXiv preprint arXiv:1911.07749*, 2019.

[15] Maria Tsiakmaki and Omiros Ragos. A case study of interpretable counterfactual explanations for the task of predicting student academic performance. In *2021 25th International Conference on Circuits, Systems, Communications, and Computers (CSCC)*, pages 120–125, 2021.

[16] Huijie Zhang, Jialu Dong, Cheng Lv, Yiming Lin, and Jinghan Bai. Visual analytics of potential dropout behavior patterns in online learning based on counterfactual explanation. *Journal of Visualization*, 26(3):723–741, 2023.

[17] Farzana Afrin, Margaret Hamilton, and Charles Thevathyan. Exploring counterfactual explanations for predicting student success. In *International Conference on Computational Science*, pages 413–420. Springer, 2023.

[18] Bevan I Smith, Charles Chimedza, and Jacoba H Bührmann. Individualized help for at-risk students using model-agnostic and counterfactual explanations. *Education and Information Technologies*, pages 1–20, 2022.

[19] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific data*, 4(1):1–8, 2017.

[20] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating robustness of counterfactual explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–09. IEEE, 2021.

[21] Susanne Dandl, Andreas Hofheinz, Martin Binder, and Giuseppe Casalicchio. *counterfactuals: An R Package for Counterfactual Explanation Methods*, 2023. R package version 0.1.2.

[22] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[23] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International conference on artificial intelligence and statistics*, pages 895–905. PMLR, 2020.

[24] Greta Warren, Mark T Keane, Christophe Gueret, and Eoin Delaney. Explaining groups of instances counterfactually for xai: a use case, algorithm and user study for group-counterfactuals. *arXiv preprint arXiv:2303.09297*, 2023.

[25] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2019.

[26] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*, pages 448–469. Springer, 2020.

[27] Dieter Brughmans, Pieter Leyman, and David Martens. Nice: an algorithm for nearest instance counterfactual explanations. *Data mining and knowledge discovery*, pages 1–39, 2023.

[28] Anna Kozak and Hubert Ruczyński. forester: A novel approach to accessible and interpretable automl for tree-based modeling. In *AutoML Conference 2023 (ABCD Track)*, 2023.

[29] Léo Grinsztajn, Edouard Oyallon, and Gaël

Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

[30] Nina Spreitzer, Hinda Haned, and Ilse van der Linden. Evaluating the practicality of counterfactual explanations. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[31] Maximilian Förster, Philipp Hühn, Mathias Klier, and Kilian Kluge. Capturing users' reality: A novel approach to generate coherent counterfactual explanations. 2021.