Balancing Act: Fairness and Accuracy in University Dropout Prediction

Literature Review

Bryan Chi Fai Pang

Student ID: 501210081

TMU: The Chang School of Continuing Education

CIND 820 Big Data Analytics Project

Dr Ceni BABAOGLU

20 October, 2023

Abstract

University dropout has long been a matter of concern for educational institutions and policymakers worldwide. Dropping out not only hinders students' academic and career prospects but high dropout rates also have broader implications for society. With the emergence of machine learning and learning analytics, early intervention is now possible by identifying a student's likelihood to drop out. Nevertheless, the extensive use of personal data, particularly protected features like gender, age, and ethnic origin, has raised questions regarding algorithm fairness. While there is extensive research in university student dropout prediction, there is relatively little research done on the relationship between demographic, socioeconomic data and the performance of prediction model.

This project aims to explore the influence of including specific categories of student data – demographic, socioeconomic, macroeconomic and academic– on the performance of machine learning models in predicting university dropout.

To accomplish this goal, we initiate the process by establishing a baseline model that exclusively relies on macroeconomic and academic data. Following this, we proceed to create additional models that incorporate both demographic and economic data. We then systematically evaluate and compare the outcomes of these models using standard performance and fairness metrics.

We anticipate that the findings of this research will make a valuable contribution to the ongoing discourse surrounding the balance between algorithm fairness and predictive accuracy in the context of student dropout prediction.

Literature Review

I. Landscape of University Student Dropout Prediction in Machine Learning

[1] Lorenz Kemper et al., (2020) ***Predicting student dropout: A machine learning approach***

Kemper's study endeavors to illustrate that high predictive accuracy can be achieved through the utilization of logistic regression and classification. Notably, Kemper's focus lies in the efficacy of the model rather than algorithm fairness. His dataset and models employ just four basic demographic attributes – gender, origin, age, and date of enrollment – while deliberately omitting other students' socioeconomic data. The rest of the dataset is composed of attributes related to student academic career.

Kemper's observations offer valuable insights that are relevant to our study:

1. The four personal demographic features he employs are in alignment with the information routinely gathered by universities.
2. By intentionally reducing the reliance on demographic data, Kemper claims that his models sidestep the potential for pre-existing discrimination based on criteria unrelated to academic achievement.

Kemper's models yield remarkable results, achieving an 88% accuracy rate. Additionally, they accurately predict student dropout in 62% of cases, starting as early as the first semester. When considering all three semesters' marks, the models' accuracy soars to an impressive 92%.

Kemper's research serves as compelling evidence that it is entirely feasible to achieve a high degree of predictive accuracy in student dropout prediction, even when utilizing a minimal set of demographic data as long as extensive academic performance data is used for prediction.

[2] Bernes et al., ***Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods***

Much like the Kemper dataset, Bernes' dataset encompasses only 4 demographic attributes: gender, place of birth, nationality, and immigration background (achieved through name-based imputation), with the remaining dataset features focused on academic performance. An intriguing observation from Bernes' study, relevant to our research inquiry, underscores the predictive capabilities of the models. When exclusively leveraging academic data, these models demonstrate a 71% accuracy rate in forecasting student dropout. However, when incorporating demographic data into the equation, the accuracy climbs modestly by 3%, reaching 74%, a point explicitly stated by Bernes.This marginal increase suggests that demographic information may only contribute minimally to the overall predictive performance and warrants further study.

[3] Francesco et al., ***Deep learning approach for predicting university dropout: a case study at Roma Tre University***

Building upon the insights gleaned from the prior two studies, which underscore the critical role of academic data in predicting student dropout, Francesco's research delves into the possibility of achieving strong model performance using solely administrative (personal and socioeconomic) data. The findings of this investigation clearly indicate that models relying exclusively on administrative attributes yield subpar results, falling short when compared to models trained with a combination of administrative and academic performance attributes.

While the primary focus of these three papers isn't algorithm fairness, their collective findings converge to the same conclusion: demographic data may make only a modest contribution to the predictive capacity of models in the context of student dropout. Conversely, academic data, such as course load and examination results, emerges as the pivotal factor in accurately forecasting student attrition.

II. Algorithm Fairness and Discrimination-aware Data Mining Practice in General

[4] Žliobaitė, I. ***Measuring discrimination in algorithmic decision making***

This paper serves as a foundational introduction to the potential for machine learning-based decisions to inadvertently perpetuate discrimination against specific demographic groups. It elucidates the various metrics and methodologies available to gauge the fairness in machine learning. Notably, despite government regulations aimed at safeguarding individuals from differential treatment on the basis of gender and race etc, machine learning models can paradoxically utilize these "protected features" to formulate decision criteria. This leads to two individuals, sharing otherwise identical profiles and characteristics in all other aspects, can receive disparate predictions solely due to difference in a protective feature, such as gender.

Furthermore, the paper delves into the critical concept of the trade-off between fairness and accuracy and the metrics that measure them. While the omission of sensitive attributes can render machine models more equitable, this might come at the expense of reduced predictive accuracy.

[5] Kelley et al. ***Removing Demographic Data Can Make AI Discrimination Worse***

Kelley contends that refraining from collecting and utilizing sensitive features may appear to be a viable means to attain algorithm fairness. However, it is precisely the inclusion of these features that enables the assessment of bias and fairness. According to Kelly et al, a more nuanced and constructive approach would entail proactively establishing guidelines pertaining to the acquisition and utilization of sensitive features, as well as actively monitoring and assessing algorithmic outcomes for any signs of discrimination.

These two papers provide basic understandings surrounding algorithm fairness and the fairness-accuracy trade-off. Familiarity with these issues is essential for the two next two papers as they deal directly with university dropout prediction and algorithm fairness.

III. Research on Algorithm Fairness in Student Dropout Prediction

A limited number of studies are dedicated exclusively to the intersection of algorithm fairness and student dropout prediction. Intriguingly, two consulted studies on the topic employ very similar methodology and share a common finding: the incorporation of protected features has a negligible impact on the predictive performance of models in forecasting student dropout in univeristy.

[6] Deho, O. et al. (2023). Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why. ***IEEE Transactions on Learning Technologies***

[7] Yu, Renzhe, et al. (2021). Should College Dropout Prediction Models Include Protected Attributes?

These two studies, although different in terms of dataset size and the specific protected attributes involved, employ a shared methodology: utilizing their respective source datasets, they create multiple subsets: one set including protected attributes and another without them. Deho's study includes protected features such as gender, age, disability, and home language, while Yu's study covers gender, first-generation college status, minority membership, and high financial need. Subsequently, models are trained using these subsets, and their performance is assessed using standard performance metrics.

Both studies' findings indicate that the inclusion of protected attributes does not significantly impact the models' performance in predicting student dropout. Instead, the most predictive power in student dropout prediction stems from academic performance data.

While these results are surprising, it's essential to note that both studies construct and tailor data features to address their specific research questions. Moreover, the comparison centers on the isolation of individual protected features, such as gender, rather than considering the entirety of demographic or socioeconomic features as a whole.

The summary of literature will follow, as it is incorporated with the research questions.

Research Questions

Our study aims to contribute to the understanding of algorithm fairness and university dropout prediction in machine learning by asking these research questions:

Attribute Types and Prediction Performance in University Dropout

How do different types of attributes impact the performance of student dropout prediction models?

What are the consequences of including or excluding specific classes of features on the accuracy of student dropout predictions?

The existing body of literature suggests that demographic features have only a marginal impact on model performance in predicting student dropout. However, this conclusion is often drawn from experiments involving the removal of a single feature or a small subset of demographic attributes (up to four), especially when they only make up a very small portion of the total features. Given the increasing awareness and sensitivity surrounding the collection and utilization of individual demographic and socioeconomic data, further research in this area holds significant potential. This question is pivotal because if accurate dropout prediction models can be developed without relying on such data, it can provide reassurance that model fairness is not compromised for the sake of accuracy.

—---

[This 2nd part, I need to do more research to decide feasibility / appropriateness and the fairness metrics .]

Attribute Types and Model Fairness [Further consultation needed for this question]

To what extent does the exclusion of demographic and socioeconomic attributes affect the fairness of a predictive model?

How can the measurement of fairness in models inform decisions regarding the use of sensitive data in the context of student dropout prediction?

By evaluating the fairness of predictive models, we aim to quantify the trade-off between fairness and accuracy when predicting student dropout. This quantitative insight can guide decision-making processes regarding the inclusion or exclusion of sensitive data, offering valuable information to institutions and stakeholders.

Dataset, Methodology and Codes Repository

The dataset used in this project is introduced by Valentim Realinho in a data descriptor paper named *Predicting Student Dropout and Academic Success*, with 4424 records and 35 attributes featuring 4424 students of an anonymized university, created by Polytechnic Institute of Portalegre. The dataset is also available in the UC Irvine Machine Learning Repository.

Github repository of this project: https://github.com/bryantoca/capstone_project

- Original Dataset: You can access the original dataset here.
- Reverse Encoded Version: If you need the dataset with nominal/categorical features reversed from numeric encoding, you can find it here.
- Capstone_module_02.ipynb Notebook: This notebook contains basic exploratory data analysis (EDA) and summary statistics. You can access it by clicking here.
- NBViewer: To view the notebook in NBViewer, please click here.

Information and descriptive statistics of the datasets, including heat map showing correction, can be viewed by clicking the link below. (It is a rendered version of jupyter notebook for module 2 (Capstone_module_02.ipynb) and can be viewed directly online.

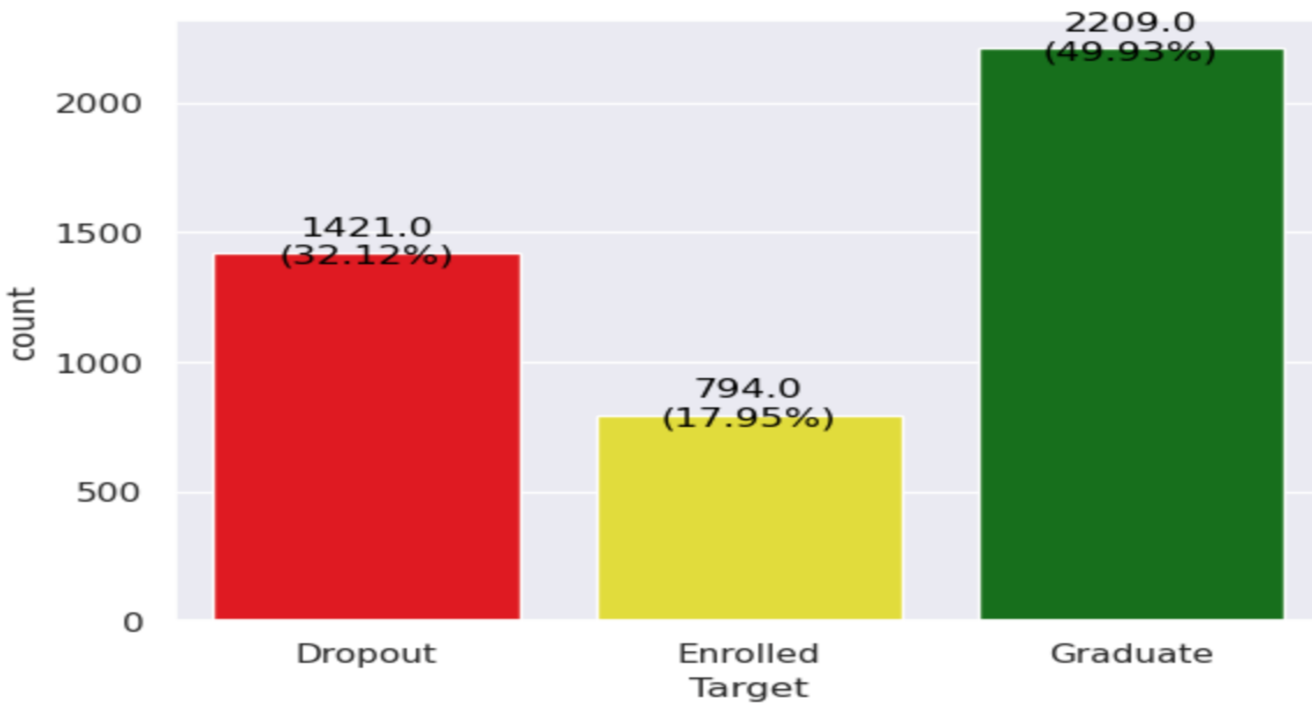
https://nbviewer.org/github/bryantoca/capstone_project/blob/59ae45b4d4dfe117d56f179768cb4a20a5cb9d6c/Capstone_module_02.ipynb

This research adopts Deho's methodology and calculation metrics for performance and fairness; modifications are made in order to serve our research questions.

Python, as well as Pandas, Seaborn, Matplotlib, Numpy libraries, are used in this project. The codes are saved in jupyter notebook format.

The dataset has a number of unique characteristics:

1. Target has three classes: Graduate, Enrolled, Dropout. (Status at the end of expected degree completion time frame)



2. Rich in demographic, socioeconomic features.
3. Inclusion of macroeconomic data at the time of student enrollment.
4. The dataset is prepared and shared for general machine learning.
5. Nominal features, such as job, qualification etc, are encoded as numerics.
6. No missing or null values.

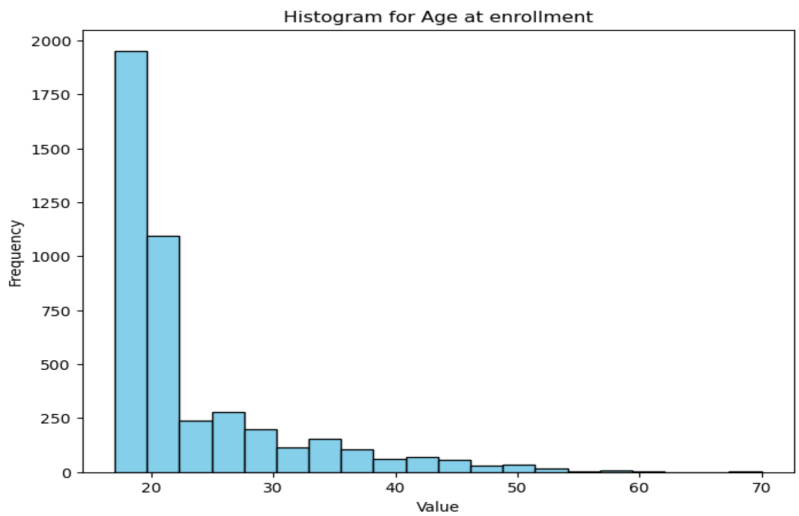
I will first highlight some interesting features related to attributes; the full attribute list follows.

At this stage, 75 records have been identified as anomalies and are identified by using

```
anomalies = data[(data['Target'] == 'Graduate') & (df.iloc[:, 21:33].eq(0).all(axis=1))]
```

These 75 students graduated with zero as value in number of courses taken and grades; they should be either removed or classified as dropout or enrolled.

Attributes "Age at enrollment" and "Grades" have unique features:

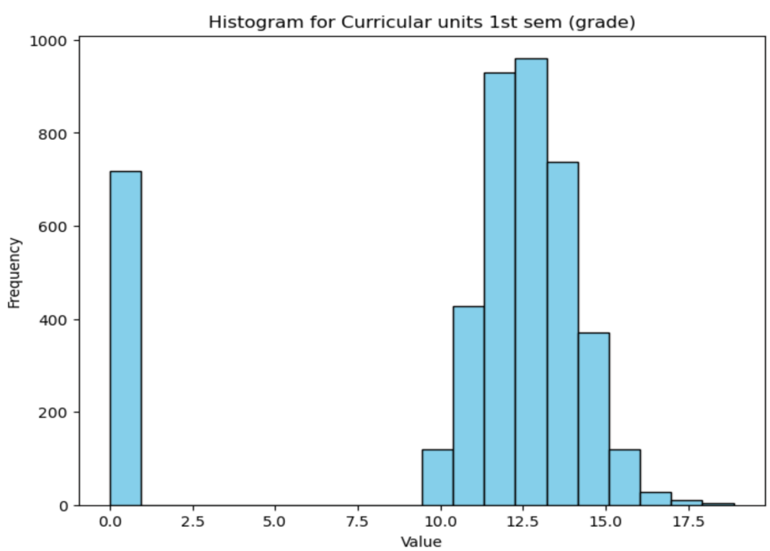


Mean age: 23.26

SD: 7.59

Range: 17-70

Age is not normally distributed; most of the students are in their 20's.



Mean grade : 10.64

SD: 4.84

Range: 0 - 18.88

Grades below 10 are considered as fail and are reported as 0.

Grades other zero proximate normal distribution.

Most of the attributes are categorical in nature and encoded numerically.

Class of Attribute	Attribute	Type
Demographic data	Marital status	Categorical
	Nationality	Categorical
	Displaced	Binary
	Gender	Categorical
	Age at enrollment	Numeric / discrete
	International	Binary
Socioeconomic data	Mother’s qualification	Categorical
	Father’s qualification	Categorical
	Mother’s occupation	Categorical
	Father’s occupation	Categorical
	Educational special needs	Categorical
	Debtor	Binary
	Tuition fees up to date	Binary
	Scholarship holder	Binary
Macroeconomic data	Unemployment rate	Numeric, Continuous
	Inflation rate	Numeric, Continuous
	GDP	Numeric, Continuous
Academic data at enrollment	Application mode	Categorical
	Application order	Ordinal
	Course	Categorical
	Daytime / evening attendance	Categorical
	Previous qualification	Categorical
Academic data at the end of 1st semester	CU 1st sem (credited)	Numeric, Discrete
	CU 1st sem (enrolled)	Numeric, Discrete
	CU 1st sem (evaluations)	Numeric, Discrete
	CU 1st sem (approved)	Numeric, Discrete
	CU 1st sem (grade)	Numeric, Continuous
	CU 1st sem (without evaluations)	Numeric, Discrete
Academic data at the end of 2nd semester	CU 2nd sem (credited)	Numeric, Discrete
	CU 2nd sem (enrolled)	Numeric, Discrete
	CU 2nd sem (evaluations)	Numeric, Discrete
	CU 2nd sem (approved)	Numeric, Discrete
	CU 2nd sem (grade)	Numeric, Continuous
	CU 2nd sem (without evaluations)	Numeric, Discrete
Target	Target	Categorical

Methodology and Measurement Metric

We begin our methodology by developing a baseline predictive model using Random Forest, fitting it solely with macroeconomic and academic attributes (Set 1). This process is then iterated, gradually incorporating additional attributes, starting with demographic attributes in Set 2 and ultimately including all types of data in Set 4. Set 5 exclusively comprises demographic and socioeconomic attributes.

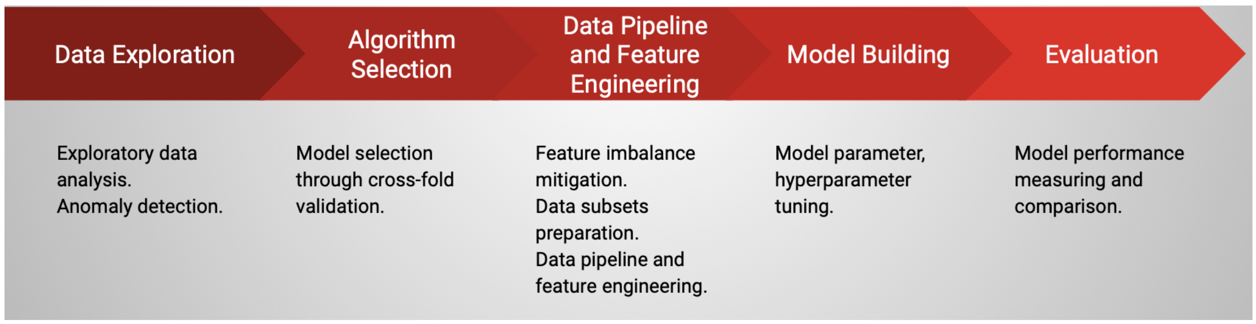
Furthermore, the same process is reiterated with Logistic Regression.

		Set 1	***Set 2***	***Set 3***	***Set 4***	***Set 5***
Class of Attribute						
Demographic			x		x	x
Socioeconomic				x	x	x
Macroeconomic		x	x	x	x	
Academic		x	x	x	x	

The performances of these models are recorded and compared in the following manners. (to be finalized)

		Set 1		***Set 2***		***Set 3***		***Set 4***		***Set 5***	
Metric		*Model 1*	*Model 2*	*Model 1*	*Model 2*	*Model 1*	*Model 2*	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Accuracy											
Precision											
Recall											
F1 score											
TPR											
FNR											

Project Progression



1. Data Exploration
    a. EDA to gain insights into the dataset and calculate general summary statistics.
    b. Review EDA findings which can be accessed at the following link https://nbviewer.org/github/bryantoca/capstone_project/blob/59ae45b4d4dfe117d56f179768cb4a20a5cb9d6c/Capstone_module_02.ipynb
    c. Anomaly detection and correction.
    d. Consideration of the target variable - three class (graduate, enrolled, dropout) or binary (graduate / not graduate)
    e. Addressing class imbalance through techniques such as SMOTE or other methods (on going research)

2. Algorithm Selection
    a. Random Forest and Logistic Regression (initial choice).
    b. K-fold cross-validation will be used for further model selection consideration.
    c. The two best performing models will be used.

3. Data Pipeline and Feature Engineering and Model Building

    a. From the original dataset, five sets will be created according to the table below.

		Set 1	***Set 2***	***Set 3***	***Set 4***	***Set 5***
Class of Attribute						
Demographic			x		x	x
Socioeconomic				x	x	x
Macroeconomic		x	x	x	x	
Academic		x	x	x	x	

b. The two chosen prediction models will be fitted using ten-fold cross validation approach, repeated 10 times, resulting in 100 train-test sets. The results (average) will be used for performance and fairness analysis.

Throughout the model building process these steps will be performed iteratively:

Feature engineering and transformation: changing ordinal categorical data to one-hot encoding, discretization of continuous numeric features such as age and grade (if needed for better performance)

Feature selection and dimensionality reduction will be performed, using a combination of correlation statistics, low variance and trial and error.

c. Results will be recorded for the two models in the following manner.

		Set 1		***Set 2***		***Set 3***		***Set 4***		***Set 5***	
Metric		*Model 1*	*Model 2*	*Model 1*	*Model 2*	*Model 1*	*Model 2*	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Accuracy											
Precision											
TPR											
TNP											
NPV											

4. Evaluation and Results
    a. The findings will be presented with appropriate graphics and tables.
    b. Writing of final paper research paper and slide for presentation.

As of the current stage, I've identified several key challenges that need to be addressed for the project's success:

Multiclass Classification Complexity: Handling multiclass classification introduces complexities, particularly in the interpretation of the confusion matrix and ROC curve. I will explore effective strategies to navigate these challenges and ensure a clear understanding of the model's performance.

Data Leakage Mitigation: The risk of data leakage, especially during the data preparation and model building stages.

Test Harness for Various Models and Sub-Datasets: Evaluating different models and various sub-datasets demands a structured test harness. I will develop a systematic approach to compare model performance across different data subsets and ensure the results are reliable. (at the research stage)

Algorithm Fairness: further studies and consultation needed to ensure methodology can in deed measure fairness (Research Question 2)

References

Agrusti, F., Mezzini, M., & Bonavolontà, G. (2020). Deep learning approach for predicting university dropout: a case study at Roma Tre University . *Journal of E-Learning and Knowledge Society*, *16*(1), 44-54.https://doi.org/10.20368/1971-8829/1135192

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, *11*(3), 1–41. https://doi.org/10.5281/zenodo.3594771

Deho, O. B., Joksimovic, S., Li, J., & Zhan, C. (2023). Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why. *IEEE Transactions on Learning Technologies, 16*(4), 560-572. 10.1109/TLT.2022.3226474

Kelley, St et al. (2023, March 6).Removing Demographic Data Can Make AI Discrimination Worse. Harvard Business Review https://hbr.org/2023/03/removing-demographic-data-can-make-ai-discrimination-worse#:~:text=and%20machine%20learning-,Removing%20Demographic%20Data%20Can%20Make%20AI%20Discrimination%20Worse,race%20can%20produce%20fairer%20outcomes.

Kemper, L. et al.(2020) Predicting student dropout: A machine learning approach, European Journal of Higher Education, 10:1, 28-47, DOI: 10.1080/21568235.2020.1718520

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, *7*(11), 146. https://doi.org/10.3390/data7110146

Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict students' dropout and academic success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Yu, R., Lee, H., Kizilcec, R. F. (2021, June 8).Should College Dropout Prediction Models Include Protected Attributes? *L@S '21: Proceedings of the Eighth ACM Conference on Learning @ Scale (*2021)

Žliobaitė, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 1060–1089 (2017). https://doi.org/10.1007/s10618-017-0506-1