Should Machine Learning Have a Place in the Admission Office of Portuguese Higher Education Institutes?

Altea Fogh (alfo@itu.dk), Sara Pissarra Gouveia Vieira (sapi@itu.dk), Mikas Jankeliunas (mikja@itu.dk)

IT University of Copenhagen, May 16, 2025

1 Motivation

Machines have become better at recognizing patterns and making predictions than humans. On the one hand, machines have no feelings or opinion which should make them more impartial than humans [1, ch.0]. On the other hand, the machine learning (ML) models we create are only as impartial as the data we feed them with. As a result, when creating ML models it's important to keep in mind biases in the data and if the model will propagate these biases or not when used to make decisions in real world problems.

Using machine learning techniques to identify students at risk of dropping out could help higher education institutes (HEI) to provide the support needed to assure they finish their degrees in due time. This was the original intention of the data collecting process of the dataset used [2, p.1] on which we base our project. However, admission offices could also take advantage of such models to stop high risk students from ever entering higher education. Which criteria are used when picking who gets admitted to higher education is an ethical question. When Rawls [3, chapter 2] writes about justice and fairness, he argues that positions that lead to inequalities should be open to all. In Portugal, higher education graduates earn more on average [2, p.14]. Therefore, to assure a fair society, it is important to guarantee that higher education institutes are places of equal opportunity, and not responsible for propagating existing social-economic biases.

Dropout rates in Portugal are comparatively high. Particularly older students are more likely to dropout or take longer to complete their degrees [2, p.48]. Furthermore, when looking at gender distribution in Portuguese HEI, we can see that while the split between genders entering them is quite even [4, p.125], there's a bigger gap between genders when we look at people that have obtained their diploma [4, p.139].

In this project we aim at creating a model to predict the risk of a student dropping out based on information at the time of the student's application. Our goal is to achieve a model that does not discriminate between sensible groups. Once the model is created, we will evaluate the fairness of the model and reason whether it would be possible to use such a tool in a place like the admission office of a HEI in a fair way. We aim to answer two question throughout this report. First, can a machine predict such information without being discriminatory? If so, should admission offices use these models when picking who gets admitted?

2 Dataset and EDA

The dataset used for this project is called "Predict Student's Dropout and Academic Success" [5]. It is created from the data of Polytecnic Institute of Portalegre (IPP), Portugal. It relates information about students from different Bachelor degrees, including different types of information for each student, in the span between the academic years 2008/2009 - 2018/2019. It contains information known at the time of enrollment, such as previous academic paths, socio-economical background and demographics, plus information about the academic performance of the students at the end of their first two semesters. For a detailed description of each of the columns, please refer to the data sheet from UC Irvine [6]. Following is a description of the features we used for our project, with the necessary processing.

The dataset includes 4424 datapoints described with 37 columns. Some of the data present in the dataset was generated with data augmentation techniques [5, p.5] such as ADASYN and SMOTE, with the aim of increasing the balance between dropout and graduating. The dependent variable, called "Target", is divided in three: Dropout, Enrolled or Graduate. Enrolled represent taking up to three extra years to graduate, while Dropout only represents more than 3 extra years/never completing the education. We have decided to drop the category Enrolled, as we are interested in analysing only the high-risk category. Taking more than the standard 3 years to finish higher education, but still graduating, is considered a success, so we cannot argue for having the same "risk level" as dropping out. The target variable is thus composed of 2209 Graduating students, and 1421 Dropout.

As part of the pre-processing of our data, we have dropped a number of features. The reasoning is applied on a case by case basis, and it is supported by the need to reduce the complexity of the data due to the use case and scope of the project. As we are focusing on the information at the admission time, we have dropped any category that describes the performance of the student during their degree (*Curricular units, Tuition fees up to date*). We also decided to drop the column *Nationality*, as the distribution is best represented by the binary column *International*. *Application mode* was dropped due to the lack of clear understanding of the values, and their impact on the performance of the student. They represent very detailed ordinances that are difficult to understand without legal domain knowledge [2, p.37] . *Previous qualification (grade)* was ignored due to already having *Admission grade*, and the lack of clear definitions. Both parents occupations were also dropped, as even if there are sources suggesting that the parent's occupation is related to academic performance of the students in math, and high school [7], [8], there is no clear understanding of the strength of the relationship for university performance in particular [2, p.49]. Finally, we decided to drop the columns describing the *GDP*, *Inflation and unemployment rates* in the country at the time of the student being accepted. We have made the hypothesis of the numbers being connected to the student's admission year, but

without the information on the actual year, we are not able to be certain of it.

The dataset comprises binary, categorical, and continuous values. We have kept the binary columns as such, manipulated the categorical to either be binary or one-hot-encoded (OHE), and standardised the continuous ones. Following is a brief description of the final columns and the processing done.

Feature Name	**Feature Type**	**Manipulation**	**Specifics**
Debtor	Binary	/	1-yes, 0-no
Scholarship holder	Binary	/	1-yes, 0-no
Marital Status	Integer between 1 and 6	OHE	Dropped first of 6 columns
Previous Qualifications	17 integers	Split in having/not having higher education/ no secondary education + OHE	Dropped first of 3 columns
International	Binary	/	1-yes, 0-no
Mother's Qualification	29 integers	Binarised: having/not having higher education	1-yes, 0-no
Father's Qualification	34 integers	Binarised: having/not having higher education	1-yes, 0-no
Displaced	Binary	/	1-yes, 0-no
Educational Special Needs	Binary	/	1-yes, 0-no
Gender	Binary	/	1-Men, 0-Women
Age at Enrollment	Range 18-62	Binarised: 23 and under/ over 23	1-Older, 0-Younger
Application Order	Integer between 1 and 9	Binarised: first priority or not	1-First priority, 0-not first priority
Course	17 different course codes	Mapped to course name - merged day/evening - OHE	Dropped first of 14 columns
Daytime/evening attendance	Binary	/	1-Daytime, 0-Evening
Admission Grade	Range 0-200	Standardized	Removing mean and scaling to unit variance

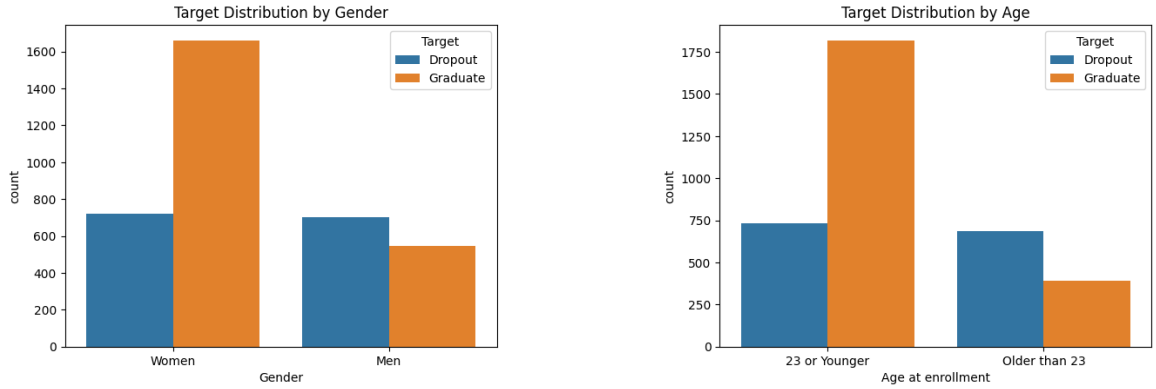
Table 1: Features and data manipulation

The final dataset has 34 columns and 3630 rows. There were no missing values.

2.1 Exploratory Data Analysis

Figure 1 presents the breakdown of the distributions of Gender and Age compared to the Target variable. In 1a we can see the difference both in number of students for each gender, and the disparity the distribution of people graduating and dropping out. Significantly more women graduate than drop out, while the numbers for men are quite similar. Similarly, figure 1b shows that there is a much larger

amount of people 23 and younger that graduate compared to dropping out, while the people above 23 have a similar distribution, more skewed towards dropping out.



(a) Target Distribution with Gender

(b) Target Distribution with Age

Figure 1: Gender and Age distribution

More plots for specific data distributions are present in appendix A (5) and in the code.

3 Methodology

3.1 Model

In models categorised as high risk, such as in our use case [9, p.9], we need to be able to assess the decisions of the model we are working with and how it has generated the outputs. Of the interpretable ML models available, we chose Logistic Regression, as it is a well-established and simple model for binary classification problems such as ours [10, ch.7]. It provides insight into the input feature weights, which in turn allows us to understand how the changes in features affect the probability of a given outcome. In the basic model we can observe that both the feature gender and age or present in the 10 most impactful features of the model as seen in figure 2. This means that the gender/age is directly affecting whether the model will predict a success or a drop-out.

On top of that, having access to the probability of the output label allows us to understand the model better, as there are obvious differences between being 99% sure of the label compared to 55%. On the other hand, we have to make sure to understand the weights of the features correctly (multiplicative odds ratio change), which makes LR a non-trivial model to use. Overall, Logistic regression is a suitable model for fairness analysis due to its transparency and interpretability, though it requires careful considerations.

3.2 Protected Features

As for the features in our dataset, admission grade would be the most reflecting factor of the student's merit (grade distribution available in the Appendix 5, figure 6), since grades should be a reflection

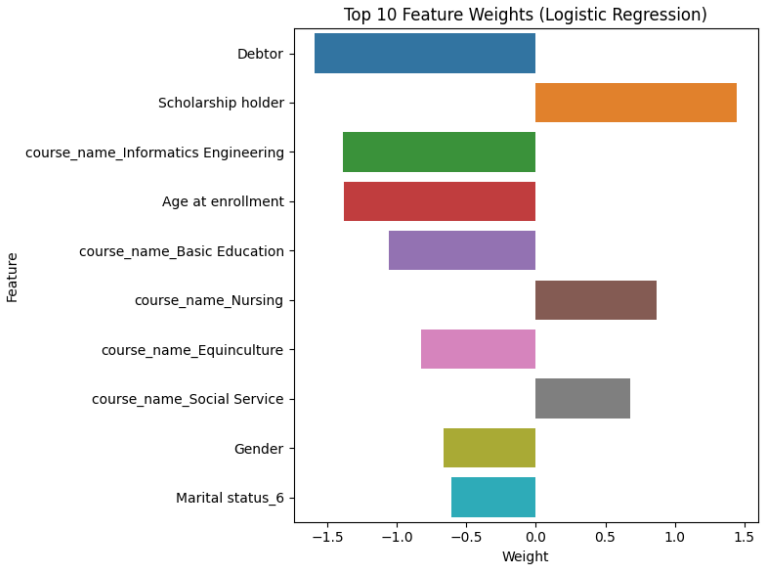


Figure 2: Feature weights for benchmark model

of the student's skills. However, just by looking at a student's grade it would be hard to correctly predict the risk of a student dropping out. Socio-economical and demographic factors, such as having a scholarship or the student's marital status, play a bigger role at determining the risk of a student dropping out as seen in figure 2.

When we consider socio-economical and demographic factors for a decision model, it is very easy to fall in the fallacy of unwanted discrimination. As described in the motivation for our project the dropout rate for men and older students is a problem that has been observed in the Portuguese higher education institutions. Furthermore, according to article 21 of EU Non-discrimination act [11] it is prohibited to discriminate based on sex or age.

As a result, considering the specifics of our data, we decided to focus on age and gender as protected features. There is a clear imbalance in our data between 2868 women and 1555 men. Similarly, after transforming Age into a binary feature, there are 3155 students below 23 and only 1269 above.

3.2.1 Fairness Metrics

Of the group fairness metrics presented in [12, p.12-13] and [13], we are going to focus on Satistical Parity (SP) and Equalised Odds (EO). As group metrics, they aim at measuring if the groups are treated equally, and they allow us to compare a metric that ignores the true label, by only looking at the predicted outcomes, (SP), and one that conditions on them (EO).

Statistical parity requires the probability of the positive outcome (eg. graduating within three years) to be the same across different groups (women/men or under 23/over 23). It does not take into consideration the true outcome (actually graduating within three years), but only the rate of positive predictions. This implies that as long as there is the same proportion of women and men

(or under/over 23) predicted as graduating, statistical parity will be satisfied, independent on the true label/correctness of the prediction.

Equalised odds, on the other hand, takes the actual predictions into account, by requiring the two groups to have the same true positive rate (correctly predicting graduation when the student does graduate) and false positive rate (incorrectly predicting graduation when the student doesn't graduate). This means that among the individuals that have the same true label (actually graduated), the model should make positive predictions, at the same rate for both groups. Equalised odds ensures also the model error to be equally distributed between women and men (and under/over 23), accounting for the actual target label. In practice, Equalized Odds helps ensure that neither group is disproportionately over- or under-predicted. It is important to understand the risk of incorrectly labeling a student as dropout when this student would have in fact succeeded (false negative) and the other way around (false positive).

The two metrics are mutually exclusive [14]: we can only adjust for one metric (which will be Equalised odds TPR).

3.3 Performance metrics

To evaluate the performance of our models, we are going to use Accuracy, Balanced Accuracy and F1 score [15]. This will allow us to assess initial performance and how the model changes after fairness considerations are included.

Accuracy measures the overall performance of the model on the two classes, by returning the number of correctly classified samples. It does not consider the nuances in the data. We use this metric as a benchmark for overall model performance.

Both the target and the protected groups have high disparities. Thus we need to also use metrics that reflect the performance within each group. We decided to use F1 score and Balanced Accuracy. Both metrics allow us to assess model performance for unbalanced classes, by making sure that neither class is at a disadvantage.

Balanced accuracy allows us to evaluate the performance for the groups while considering the imbalance in the protected classes, allowing us to assess the disparity within them [5]. It highlights the disparity across each group even if the overall performance is quite high.

F1 score keeps a balance of precision and recall, taking both false positives and false negatives into consideration. If one class was consistently under-predicted, F1 score would help us see that. For example, as men are much less present than women in the dataset, F1 score can help us identify if the model under-predicts for them, despite overall.

3.4 Method for Improving Fairness

As we have full access to the data, we decided to start by training our model on a dataset without the protected features to see if we could obtain an equally accurate model without discriminating for gender and age. This is the concept of fairness through unawareness presented in [13, p.6]. The new model improved slightly both in accuracy (77%) and f1-score (83%) compared to our initial model, which had access to this features when training. The resulting model improved significantly in fairness between genders (88% vs 81%). However, it was still showing a significant difference between the over/under 23 groups (90% vs 70%). Originally, age shares a significant correlation with marital status (0.43) and daytime/evening attendance (0.46), while gender shares a significant correlation with age (0.19) and scholarship holder(0.19), as seen in figure 3. There are more features with a higher correlation with age than with gender which might explain why the improvement in fairness is bigger for gender than age.

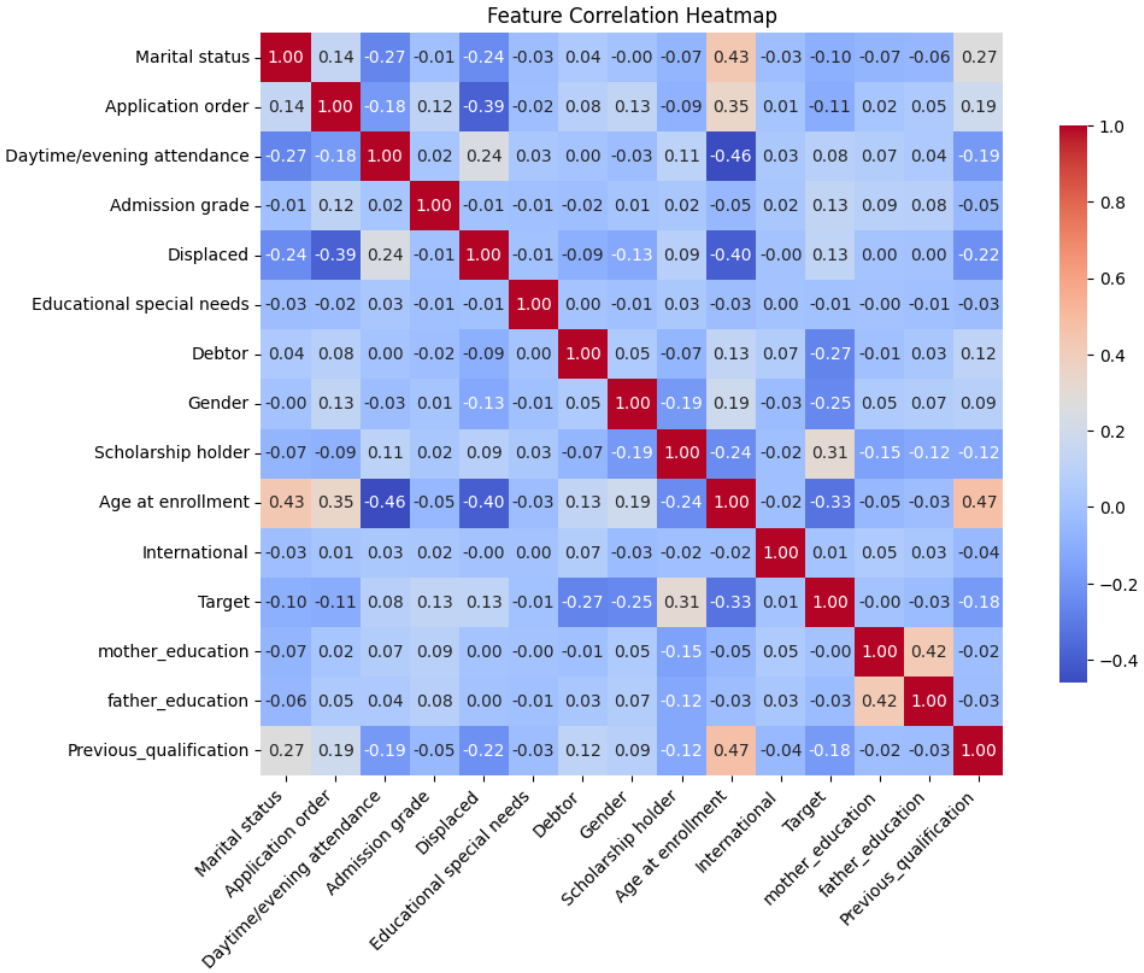


Figure 3: Correlation heat map for the features

Considering our results, we decided to use Fair PCA [16] to reassure that our model would not focus on proxy features. Once again, the model trained on the dataset transformed through Fair PCA obtained a very similar accuracy (75%) and f1-score (80%) to the initial model. The new model did

not increase significantly the fairness metrics neither for age (89% vs 59%) nor for gender (92% vs 68%).

Considering that pre-processing techniques were not enough to produce a fair model, our next step was to move to in-processing techniques. We decided to constraint our logistic regression model to penalize it whenever it would treat the groups differently [17]. We used a group constraint, which means that if we make bad predictions for one group we can compensate by making bad predictions for the other group. We achieve the same accuracy (78%) and f1-score (83%) on both the model constraining for age and gender. These were the highest performance values we were able to obtain throughout the different approaches. However, there was no improvement when it came to fairness metrics compared to the original model (age: 92% vs 51%, gender: 92% vs 67%). For the specific comparison, refer to table 2.

Model	Accuracy	F1-Score	EO (TPR) Women	EO (TPR) Man	EO (TPR) Under 23	EO (TPR) Over 23
Inital Model	76	81	93	66	92	53
No Protected Features	77	83	88	81	90	70
Fair PCA	76	82	92	60	89	59
Constrained Model (Age)	78	83	-	-	92	51
Constrained Model (Gender)	78	83	92	67	-	-
Post-Processing (Age)	75	83	-	-	93	99
Post-Processing (Gender)	60	56	41	38	-	-

Table 2: Perfomance and Fairness metrics comparison

As a result, we decided to see if we could improve the fairness of our model in a post-processing phase [18]. We recognize that the different groups have different behaviors so it would make sense to apply different threshold to each of the groups. By using ROC curves, we found the point where true positive rate (TPR) and false positive rate (FPR) are closest to each other. This was the technique that most improved our fairness metrics for age. The new threshold for age kept the accuracy (75%) and f1-score (83%) of the original model. The same was not verified for gender, where accuracy (60%) and f1-score (56%) dropped a lot (figure 4).

4 Discussion

We have attempted different ways to debias our model: pre-processing of the dataset by trying to remove the protected features, in-processing by constraining the model and post-processing by defining

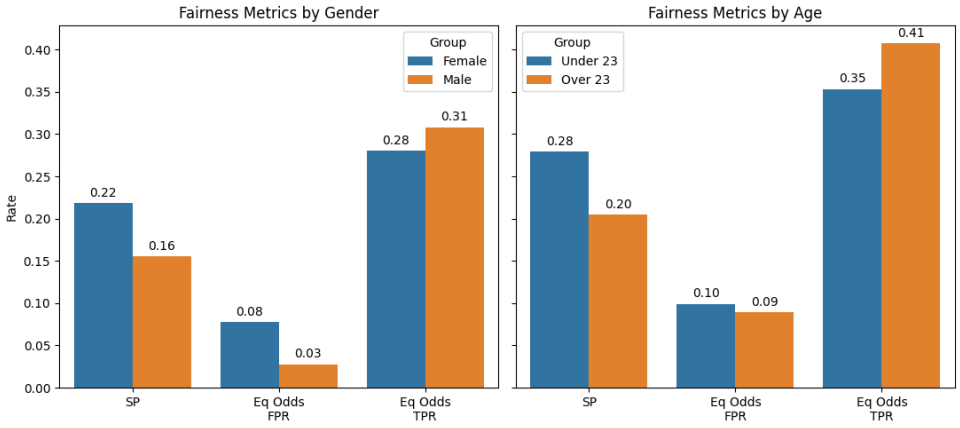
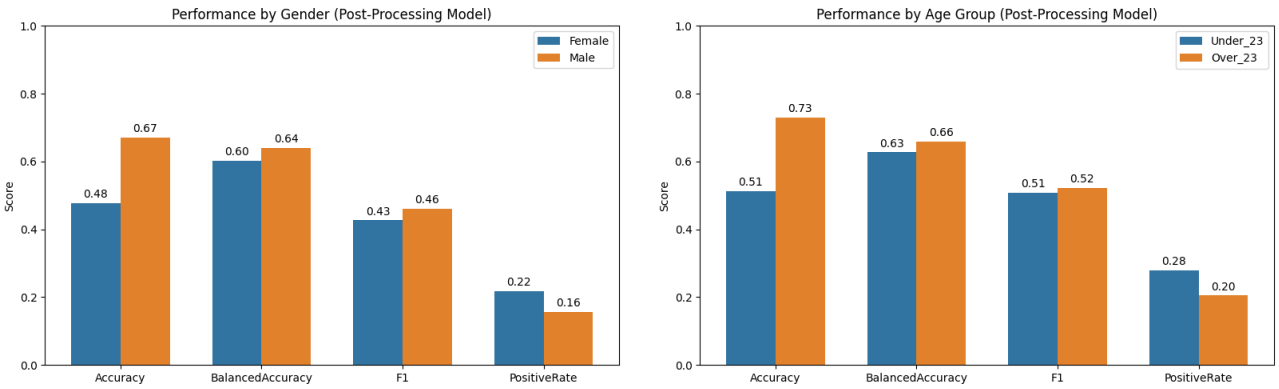


Figure 4: Fairness metrics for the post-processing model

different thresholds to each protected group. We noticed that removing the protected features was not enough since the behavior of these features is reflected by proxy ones. Applying constraints to the logistic regression when training the model also did not obtain satisfactory results. It is possible that a model which is able to capture non-linear relations would perform better in terms of fairness metrics after being constrained, at the detriment of explainability. We achieved the best fairness results when we opted for post-processing methods. Achieving fairness in this case came at a cost of performance. The model actually becomes more accurate for the underrepresented groups (male and over 23), as seen in figure 5.



(a) Performance metrics for gender in post-processing

(b) Performance metrics for age in post-processing

Figure 5: Perfromance metrics for the post-processing model

By applying different thresholds to people under/over 23 we were able to improve fairness without reducing the model's performance. The same was not possible with gender. When we used different threshold to obtain similar TPR and FPR for gender we dropped accuracy and f1-score significantly. The best trade-off we found for gender was to simply remove it from the original dataset before training the model. When reasoning about our results for this feature, it is important to keep in mind that the gender imbalance in our dataset is not representative of the distribution in higher education institutions in Portugal[4, p.125]. Women are more represented in our dataset that in the real world. In

figure 7 in the appendix 5, it is possible to see the distribution of students by study line, which presents a high number of courses that might be considered as women-dominated, which could explain the discrepancy.

As a result, we need to thoroughly question the purpose of our model. Figuring out whether the likelihood of a student dropping out of a higher education degree before even giving them a chance to be enrolled is unfair. If we look at Rawls justice and fairness theory [3], he argues that positions which might lead to inequalities should be open to all and having a bachelors in Portugal is associated with better salaries. Therefore, getting a bachelors should be an option for everyone.

Furthermore, predictive dropout models rely on past data that reflects structural inequalities. Rawls argues that inequalities should only be allowed if they work in advantage of every part. This is not the case of our analysis since we have not been able to create a model that does not discriminate between gender and age. The existence of a systematical bias in Portuguese higher education institutes is reflected in our dataset. We can not represent a bias system through an unbiased dataset, the same way, we can not capture a bias system through an unbiased model.

In conclusion, we do not believe the prediction of the risk of a student dropping out should be calculated by a machine learning model, especially a Logistic Regression. Turning the value of the risk into a number dehumanizes the nature of the reasons that might lead a student to dropout. The evaluation of this risk should be done by competent people in the aim of providing support, not of depriving a student of a chance.

5 Conclusion

The report aimed at answering two questions. The first being if machine learning can predict if a student will drop out at the time of admission, and the second is if the admission offices should use the model to begin with.

First, we have calculated the fairness metrics for the basic model, to use as a benchmark to assess improvement. Subsequently, we have used pre-processing techniques, which have shown a positive impact on gender, but not for age. In-processing techniques improved overall accuracy but did not impact fairness significantly. We then decided to assess if post-processing techniques could have positive results. Age benefited from the manipulation, while gender did not.

Lastly, we have argued for the benefit of using this model in the specific setting. We have reached the conclusion that, based on the use case, ethical considerations and the results of our analysis, it is not advisable to predict the risk of students dropping out based on admission information.

The code for the project can be found here: GitHub Repository [19]

A Appendix

Additional Figures

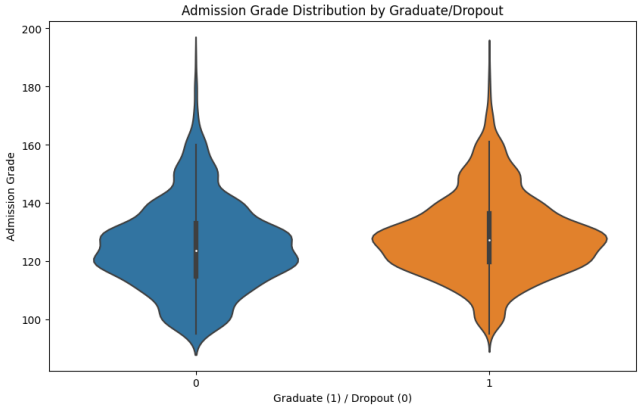


Figure 6: Violin Plot of admission grade distribution

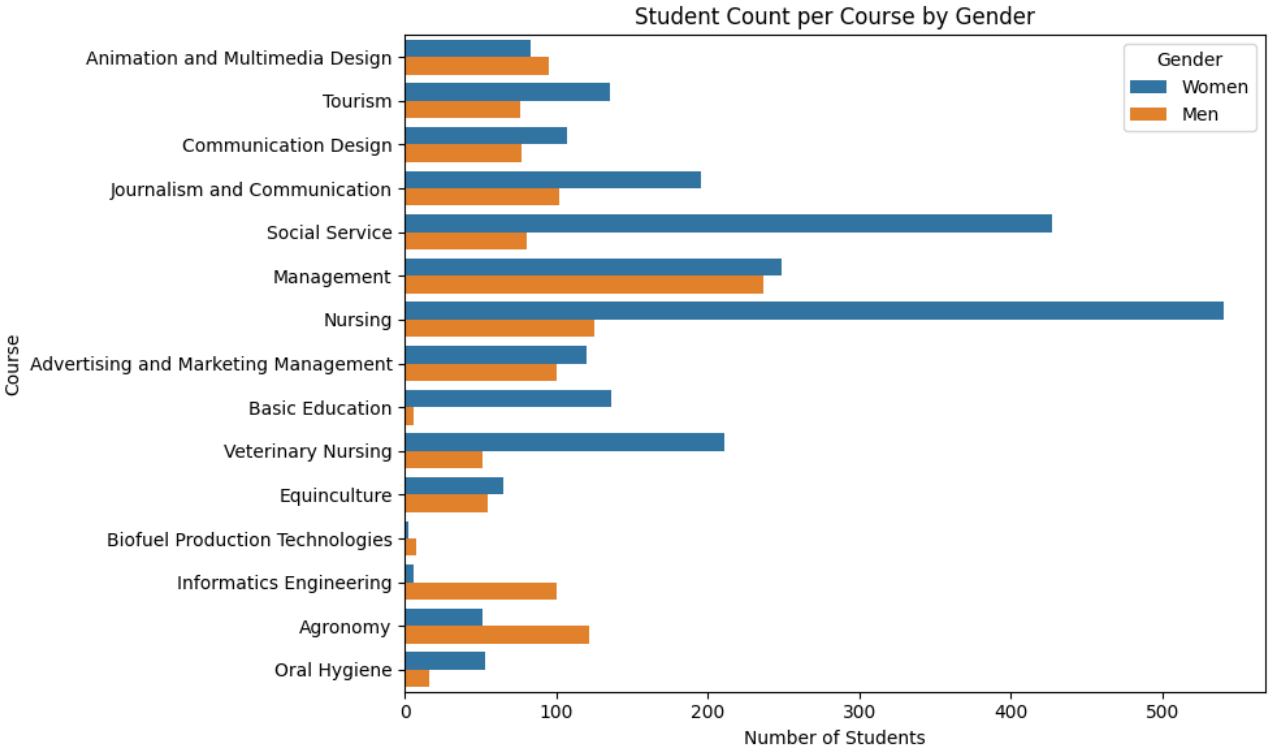


Figure 7: Student count and division by gender for each study line

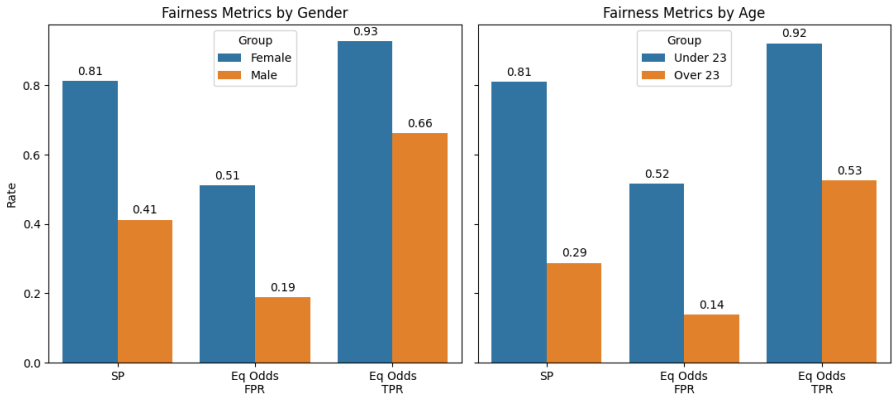


Figure 8: Fairness metrics for benchmark model

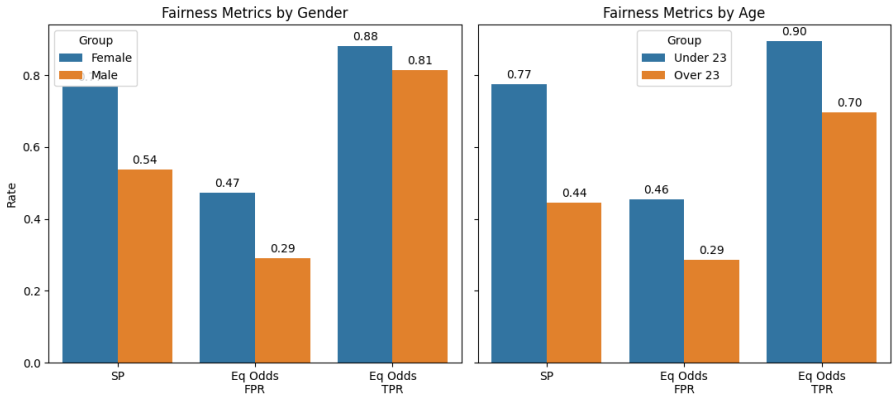


Figure 9: Fairness metrics for model without protected features

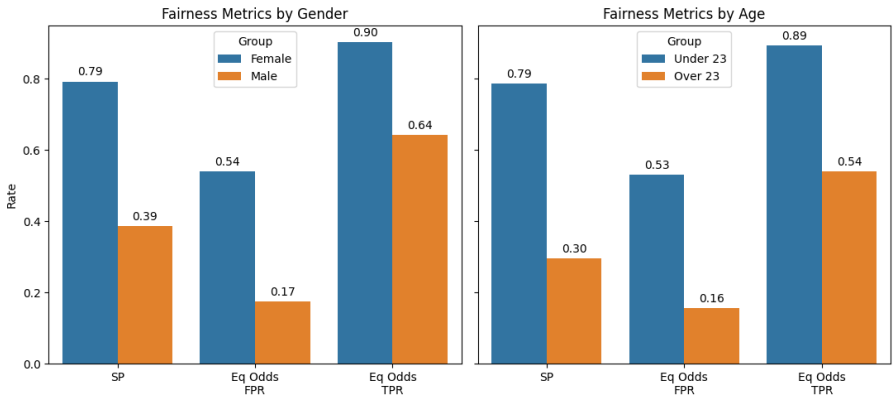


Figure 10: Fairness metrics for in-processing model

References

[1] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, Inc., USA, 2019. ISBN 0190948205.

[2] OECD. Resourcing higher education in portugal. 2022. URL https://doi.org/10.1787/a91a175e-en.

[3] Gordon Graham. *Theories of Ethics*.

[4] Divisão de Estatísticas do Ensino Básico e Secundário (DEEBS) Divisão de Estatísticas do Ensino Superior (DEES Direção-Geral de Estatísticas da Educação e Ciência (DGEEC), Direção de Serviços de Estatísticas da Educação (DSEE). Perfil do aluno 2017/2018. Technical report, Ministério da Educação, Portugal, 2019. URL https://www.dgeec.medu.pt/api/ficheiros/65798b9648cf33c04d6c35f9.

[5] Machado J Baptista L. Realinho V, Vieira Martins M. Predict students' dropout and academic success [dataset]. *UCI Machine Learning Repository*, 2021. URL https://doi.org/10.24432/C5MC89.

[6] UC Irvine. Predict students' dropout and academic success, 2021. URL https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success.

[7] OECD. Do parents' occupations have an impact on student performance? *PISA in Focus, No. 36, OECD Publishing*, 2014. URL https://doi.org/10.1787/5jz8mr7kp026-en.

[8] Ricardo Colaço Luís Catela Nunes, Pedro Freitas. Trends in student achievement in portugal: What does pisa tell us? 2024. URL https://www.edulog.pt/storage/app/uploads/public/666/875/9da/6668759daa0d6352676543.pdf.

[9] Tambiama Madiega. Artificial intelligence act. *European Parliamentary Research Service*, 2024.

[10] Christoph Molnar. *Interpretable Machine Learning*. 3 edition, 2025. ISBN 978-3-911578-03-5. URL https://christophm.github.io/interpretable-ml-book.

[11] European Union. Charter of fundamental rights of the european union. *Official Journal of the European Union*, 2012.

[12] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 54(6), 2021. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

[13] Rubin J. Verma S. Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness - FairWare '18*, 2018. URL https://fairware.cs.umass.edu/papers/Verma.pdf.

[14] Ziyuan Zhong. A tutorial on fairness in machine learning, 2022. URL https://medium.com/data-science/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

[15] Neptune.ai. Balanced accuracy: When should you use it?, 2025. URL https://neptune.ai/blog/balanced-accuracy.

[16] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair pca for fair representation learning. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5250–5270. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/kleindessner23a.html.

[17] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression, 2017. URL https://arxiv.org/abs/1706.02409.

[18] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[19] Group Yay. Afae dropout github repository. URL https://github.com/AlteaF/AFAE_dropout.