# Predicting Dropout Rates: A Study of Equity and Bias

Ananya Kulshrestha  Ronak Saluja

## 1 Introduction

Education is one of the most important foundations of societal progress, shaping individuals' future and crafting economic opportunities. Studies from the Bureau of Labor Statistics [5] show that those with higher education degrees have both higher financial earnings and lower unemployment rates, demonstrating the necessity of education for living a comfortable life. Therefore, addressing the challenge of identifying students at risk of underperformance or dropout is essential for fostering equity and opportunity in education.

Machine learning (ML) models have now emerged as powerful tools for addressing such challenges by analyzing large public datasets and uncovering patterns indicative of specific student outcomes. Unfortunately, these models are not immune to bias or errors. There are various critiques the lower quality and biased datasets that are used in predictive algorithms, such as those used in healthcare [6], policing [3], or education, risk perpetuating systemic biases. Similarly, research from the American Educational Research Association (AERA) [1] exemplifies how algorithms used to predict student success may actually disadvantage certain racial groups. Specifically, they find that these models often predict failure more often for Black and Hispanic students, while also overestimating the predicted success for White and Asian students. This demonstrates how existing models often fail to account for historical inequities embedded in data, resulting in biased model outputs further perpetuating societal biases.

As Birhane et al. [2] note, ML research is not value-neutral; the priorities embedded in models often reflect corporate goals rather than broader societal needs. Their analysis, in fact, found that metrics like performance and efficiency are more frequently prioritized, while factors like societal relevance and potential harms are rarely addressed. These findings underscore the importance of inspecting values that are a key part of ML systems, particularly in high-stakes domains like education. For example, when models focus on optimizing overall performance, they risk overlooking disparities in subgroup outcomes, thereby exacerbating existing inequities.

This study hopes to address these challenges by investigating the fairness and ethical implications of ML models used to predict student dropout risk. By examining the sensitivity of predictions to "socially relevant" features like gender and ethnicity, we aim to ensure that machine learning advancements avoid reinforcing existing disparities. This work highlights the importance of ethical considerations in the design and evaluation of machine learning models, striving to create systems that foster equitable educational opportunities for all students.

## 2 Approach

### 2.1 Dataset and Preprocessing

This study utilized the *Students Performance* dataset [4] from Kaggle, which includes data on about 2,500 high school students with ages between 15 and 18 (inclusive). The dataset contains demographic attributes (e.g., gender, ethnicity), parental involvement metrics (e.g., parental support), academic performance indicators (e.g., study time, absences), and extracurricular involvement indicators. The target variable, *GradeClass*, was transformed into a binary classification task to predict students at risk of dropout. A GradeClass score of 4.0 (GPA $< 2.0$) identified students as at risk, while scores

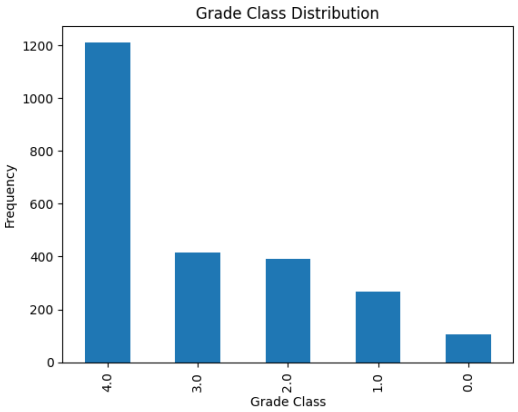


Figure 1: Distribution of the data by GradeClass

of 0–3 were grouped as not at risk. Features such as GPA were removed to avoid direct influence on predictions. Figure 1 shows the GradeClass distribution, with a majority of students classified as at risk. The distribution also shows that our split of the target variable results in balanced class sizes, enabling more reliable training and evaluation of our model.

Intersectional groups were constructed by combining gender and ethnicity to evaluate fairness across diverse demographics. This resulted in the creation of eight intersectional groups, for each unique combination of Race and Ethnicity represented in the dataset. Stratified train-test splits, stratified by these intersectional groups, ensured proportional representation across training and testing subsets, with 75% of each intersectional group used for training, and the remaining 25% used for testing. This preprocessing step was critical for analyzing fairness and subgroup performance across models.

### 2.2 Experiments and Model Selection

We evaluated fairness by analyzing predictions across demographic groups. This analysis was repeated under four feature inclusion scenarios: using all features not already dropped, excluding just gender, excluding just ethnicity, and excluding both ethnicity and gender. The sensitivity of predictions to these demographic features such as gender and ethnicity was used to assess their influence on subgroup disparities.

Three machine learning models were trained and evaluated for the task: Logistic Regression, Random Forest, and a Multilayer Perceptron (MLP). The hyperparameters for each model were optimized to balance general performance and fairness: for *logistic regression* we used $C = 0.1$, class balancing, `lbfgs` solver, and $L_2$ regularization; for *random forest* we used 200 estimators, max depth of 5, minimum leaf size of 4, and no max features; for the *MLP* we used two hidden layers of 10 neurons each, ReLU activation, learning rate of 0.001, and up to 1000 iterations. Each model was evaluated on overall performance and within demographic subgroups under the four feature inclusion scenarios described in Section 2.1 to assess both accuracy and fairness.

### 2.3 The Absences Feature

The *Absences* feature was identified as a strong predictor of dropout risk, demonstrating a high correlation value of **0.73** with the target variable of GradeClass scores. Its prominence highlighted the significance of attendance in predicting academic outcomes. However, it also raised concerns that absences could act as a proxy for other factors, such as parental support or socioeconomic background, potentially introducing biases. This prompted further analysis of its influence on both overall performance and subgroup disparities.

## 3 Evaluation

We will begin by investigating the accuracies with the Absences feature included as a part of training, showing the results in the form of three charts. For each chart, we will have a bar graph for each of our four model configurations and its corresponding overall accuracies and subaccuracies for each demographic for that experiment. We will do the same for the results when Absences is removed as a feature. To further inspect the results when the Absences features are removed, we will look into F1 scores and display those in chart forms as well, so as to analyze how the model deals with correctly identifying instances of dropout rates while also not neglecting any students at danger of dropping out.

## 3.1 Results with Absences

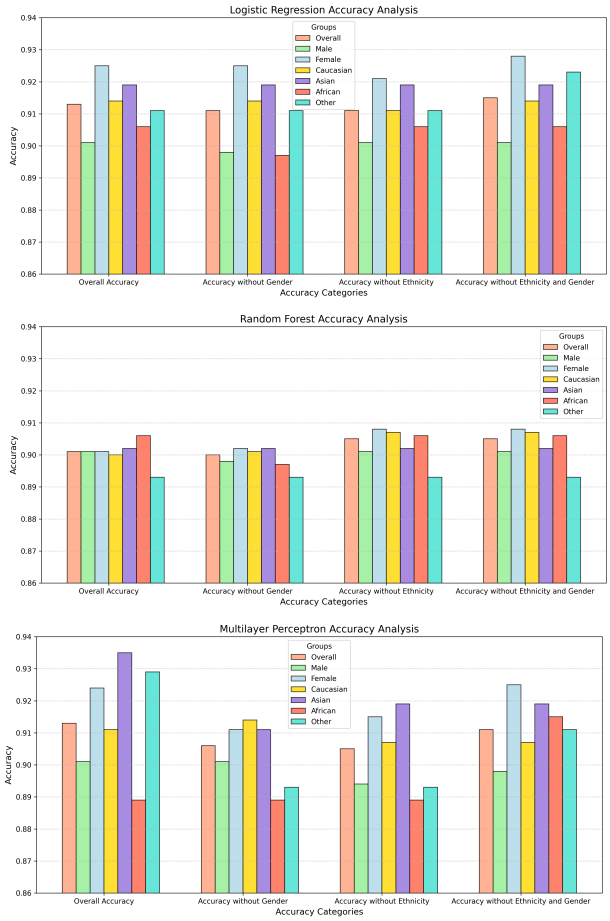


Figure 2: Model Performances With Absence

## 3.2 Results without Absences

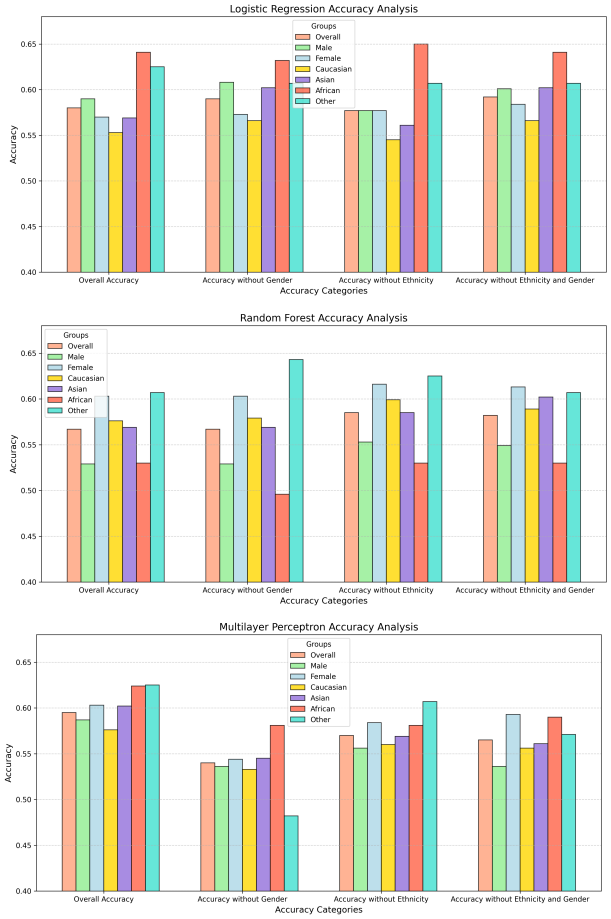


Figure 3: Model Performances Without Absence

| Group | Overall | No Gender | No Ethnicity | No Both |
|---|---|---|---|---|
| Overall | 0.918 | 0.916 | 0.916 | 0.919 |
| Male | 0.904 | 0.901 | 0.904 | 0.904 |
| Female | 0.930 | 0.930 | 0.927 | 0.933 |
| Caucasian | 0.921 | 0.921 | 0.918 | 0.921 |
| Asian | 0.926 | 0.926 | 0.926 | 0.926 |
| African | 0.906 | 0.897 | 0.906 | 0.906 |
| Other | 0.898 | 0.898 | 0.898 | 0.920 |
| Overall | 0.909 | 0.907 | 0.912 | 0.912 |
| Male | 0.906 | 0.903 | 0.906 | 0.906 |
| Female | 0.911 | 0.911 | 0.917 | 0.917 |
| Caucasian | 0.911 | 0.911 | 0.917 | 0.917 |
| Asian | 0.912 | 0.912 | 0.912 | 0.912 |
| African | 0.911 | 0.903 | 0.911 | 0.911 |
| Other | 0.880 | 0.880 | 0.908 | 0.880 |
| Overall | 0.918 | 0.912 | 0.911 | 0.911 |
| Male | 0.904 | 0.907 | 0.898 | 0.898 |
| Female | 0.931 | 0.915 | 0.922 | 0.925 |
| Caucasian | 0.918 | 0.918 | 0.916 | 0.907 |
| Asian | 0.941 | 0.908 | 0.926 | 0.919 |
| African | 0.889 | 0.917 | 0.889 | 0.915 |
| Other | 0.920 | 0.880 | 0.880 | 0.911 |

Table 1: F1 Scores With Absences

| Group | Overall | No Gender | No Ethnicity | No Both |
|---|---|---|---|---|
| Overall | 0.601 | 0.610 | 0.595 | 0.613 |
| Male | 0.605 | 0.639 | 0.587 | 0.629 |
| Female | 0.597 | 0.581 | 0.603 | 0.597 |
| Caucasian | 0.599 | 0.609 | 0.572 | 0.594 |
| Asian | 0.607 | 0.647 | 0.603 | 0.647 |
| African | 0.618 | 0.606 | 0.655 | 0.644 |
| Other | 0.553 | 0.522 | 0.577 | 0.560 |
| Overall | 0.586 | 0.584 | 0.609 | 0.601 |
| Male | 0.552 | 0.552 | 0.581 | 0.580 |
| Female | 0.618 | 0.616 | 0.636 | 0.622 |
| Caucasian | 0.621 | 0.623 | 0.637 | 0.622 |
| Asian | 0.576 | 0.576 | 0.617 | 0.623 |
| African | 0.509 | 0.478 | 0.522 | 0.522 |
| Other | 0.542 | 0.583 | 0.604 | 0.577 |
| Overall | 0.603 | 0.604 | 0.583 | 0.586 |
| Male | 0.593 | 0.606 | 0.558 | 0.564 |
| Female | 0.613 | 0.602 | 0.607 | 0.608 |
| Caucasian | 0.600 | 0.613 | 0.575 | 0.581 |
| Asian | 0.637 | 0.619 | 0.602 | 0.620 |
| African | 0.593 | 0.619 | 0.588 | 0.586 |
| Other | 0.553 | 0.419 | 0.577 | 0.520 |

Table 2: F1 Scores Without Absences

### 3.3 Key Findings

Analyzing Figure 2 and Table 1, we can clearly see consistently strong performance throughout all of our model choices and feature selection combinations. Between all accuracy subgroups, we notice that subaccuracy is bounded by ∼87 and ∼94%, indicative of high performance without significant deviations. We also observe a similar level of consistency when examining the F1 scores, with all F1 scores being bounded between ∼88 and ∼93%, indicating that the model is fairly balanced in its decision making without significant biases based on demographics or feature selection. However, as noted in Section 2.3, this behavior is not surprising due to heavy correlation between our target variable and the Absence feature in this specific dataset. As a result, we are interested in observing whether this high performance will persist if we remove this feature. These results can reveal whether the model is still able to determine how to properly predict whether a student will drop out without having access to such vital information.

However, when looking into the results of models which didn't include the Absence feature (Figure 3 and Table 2), we see a significant decrease in both overall and demographic-specific performance. The range of subaccuracies expands to now range from ∼48 to ∼64%, indicative of heavy variation between different demographics and feature selection choices. This behavior demonstrates that the Absence feature was truly vital for the model to perform well, demonstrating that the other features in their raw form are not complex enough to properly model this situation. Qualitatively observing our graph display subaccuracies, while the data mostly displays consistency across demographics and different models (although to a smaller degree than the models including absences), we still notice two key distinctions.

In the Multilayer Perceptron model, we observe unique behavior for the "Other" demographic, which is also the least represented model in the dataset. We observe for this demographic the largest discrepancy with respect to performance for a certain demographic group between different models. Specifically, there is a significant 14.3% reduction from the model trained with all features (62.5%) when compared to the model trained without Gender as a feature (48.2%). However, when adding back Gender as a feature while removing Ethnicity or removing both Ethnicity and Gender, the accuracy is in the low 50/high 60s, much closer to the performance of the model with all features. When observing the F1 scores, we notice a similar trend where the F1 score for Other when the model is trained without the gender feature, which is equivalent to 0.419. Notably, not only is the difference between the F1 score for this model and the other models for this demographic between ∼10-16%, but the other demographics trained without the Gender feature all have a F1 score greater than .6, indicating significantly much poorer performance than all other demographics. This result is indicative of intersectional bias, where not including gender might affect underrepresented groups to a greater extent than groups with greater representation in the dataset.

To a smaller extent, we notice similar results in the Random Forest model for the African-American demographic. While the performance of the different models on the demographic itself is fairly consistent across the four models for both accuracy and F1 score, we notice that the performance consistently lags behind that of the other demographics. This is especially prevalent when we remove the Gender demographic, where the F1 score for the African-American demographic and the second lowest F1 score has a nearly 10% difference. As above, this could be indicative of intersectional bias, but this time for another relatively underrepresented demographic in the dataset.

In summary, while we mostly observe consistent and stable behavior even without absences, there are a handful of instances where there are clear disadvantages between demographic groups when Absences isn't included as a feature.

### 3.4 Importance of Demographic Features

For our models which utilize Absence as a feature, we notice that Absence is by far the most dominant fea-

ture. We can quantitatively extract this value in our RandomForest models, as the scikit-learn's RandomForestClassifier has a built-in method for extracting the proportion of importance for each feature with respect to predicting the final output. In this case, we notice that across our four models, the average importance given to the Absences feature is 86.43%. This high value is also consistent with the high performance of the model, demonstrating the usefulness of the feature in predicting dropout risk. In order to get a better idea of how ethnicity and gender influence model predictions when this feature isn't involved (representing the remaining ∼13%), we also calculated feature importance for the models not using the Absence feature. With these results, we find that "socially relevant" features play an overwhelmingly minor role in influencing results, indicated by ethnicity having ∼6% feature importance, while gender has 2.27% feature importance in all scenarios where it is used as a feature. Additionally, we notice that for our Logistic Regression and Random Forest models, accuracy mostly stays the same or improves overall across all demographics when both Gender and Ethnicity are excluded, indicative of the lack of necessity regarding this feature. However, accuracy goes down by ∼1-5% when these features are excluded for the MLP model. Nonetheless, as our MLP model did exhibit the most significant evidence of inequitable outcomes and biases, this outcome demonstrates the importance of weighing the tradeoff between fairness and accuracy. As a result, our results suggest that there is no need to include such demographic features in models used in these contexts. Not only is their contribution relatively small, they can still influence potential disparities between demographic groups.

# 4 Conclusion

## 4.1 Takeaways

From our experiments and results, we have three major takeaways.

1. **A model is only as good as its data**: At the beginning of the paper, we mention how models often reflect inequities in real-world data for algorithms in this domain. For our specific experiments, we see this trend continue. Specifically, as we didn't do detailed work towards dealing with class imbalance besides careful stratification between our training and testing data, we also notice class imbalance in especially our MLP model. Additionally, as our dataset had a heavy correlation between one feature and the target, the model heavily reflected this in its preliminary results. As a result, our findings further emphasize the importance of collecting reputable and unbiased data for systems which impact real lives. Similarly, while we find in our specific study that experimenting with different models and feature selection doesn't strongly affect model biases for the most part, that doesn't imply that this is a universal finding. This was a result specific to our data, and other models trained with different datasets will have different results.

2. **Performance and fairness must coexist**: In our models trained without using Absence as a feature, we note that the best overall performance is in the MLP model. However, this model also exhibits noticeable bias, as discussed with respect to the "Other" ethnic demographic. On the other hand, the Logistic Regression model has an accuracy difference of only 0.2%, without displaying a similar degree of difference. As a result, for just a small tradeoff in accuracy, we have a much fairer model across all demographics. This demonstrates the importance of not just striving for the highest possible accuracy, but also being meticulous in ensuring fair AI systems are designed.

3. **Every decision counts**: While our results are mostly consistent, we note that even slight fluctuations in feature selection and model architecture can cause large deviations in performance. Concretely, even though we notice that Gender has a very small feature importance, removing it from our MLP model still caused a signifi-

cant decrease in accuracy for the Other demographic group. We notice a similar effect in the F1 score in our Random Forest model for the African American demographic when removing the Gender feature. This demonstrates that even small decisions can have a huge impact on model performance and fairness, indicating the importance of being methodical and ethical at each step when designing real-world AI systems.

### 4.2 Future Work

Future work would prioritize addressing class imbalance using various methods such as undersampling, oversampling, or synthetic data generation (using GANs) to ensure fair representation across demographics. Additionally, investigating the effects of diverse datasets and incorporating other potential fairness-aware machine learning techniques can offer deeper insights into achieving a balance between performance and equity in practical AI applications.

## 5 Ethical and Societal Risks

One of the key goals of our project is examining how demographic bias can affect model fairness and performance. While our results find the existence of bias in just a handful of cases, it is important for audiences of our paper to note, as we mention in the conclusion, that this is not a fact for all datasets and machine learning models. As a result, it is important to understand that dealing with data containing such sensitive information should always be treated with caution and care. It can be extremely detrimental to deploy models or advocate for data which perpetuates social biases, and we want to strongly stress that readers of this paper should always consider the dataset they are working with before utilizing potentially sensitive features. In our specific case, it is also important to note that the dataset we utilized is synthetic, as it is intended for machine learning research purposes. Therefore, making broad generalizations about the importance of demographic features for models in this domain and context is strongly not recommended. To have a more realistic project, we would consider finding real-world datasets. However, it would be important in this scenario to be respectful of data privacy. Working with real-world data requires specific additional caution towards ensuring data cannot be linked to a person's real identity, which requires evaluators to be very vigilant towards not revealing compromising information in raw data or subsequent analysis. In conclusion, we strongly urge future studies building upon our work to not blindly follow our results, and also ensure that data privacy guidelines are followed thoroughly.

## Contributions from Members

Group: Ananya Kulshrestha and Ronak Saluja (both *Recitation 3 F11*)

We both contributed nearly equally to the project, doing most of the work together in synchronous meetings. Specifically, we determined the project idea, did research into related work, determined our experimental design, and evaluated our results together synchronously. We also worked on the presentation and slides together, making sure to split work evenly. As for independent work, Ronak worked on training the RandomForest and LogisticRegression models on his own time, while Ananya worked on training the MLP model on her own time. In this paper, the first half was primarily written by Ananya, and the second half was primarily written by Ronak, with both of us proofreading each other's sections. In all, we both contributed very equally to all deliverables and implementations.

## Acknowledgements

We would like to express our heartfelt gratitude to Professor Hadfield-Menell, Professor Wilson, our TA Stephen Casper, and all other course staff of 6.3950 for their invaluable guidance, support, and constructive feedback throughout the duration of this project. Their expertise and dedication have been instrumental in shaping our understanding of how artificial intelligence affects society and fairness.

## References


[1] American Educational Research Association. "Study: Algorithms Used by Universities to Predict Student Success May Be Racially Biased". In: (2024). URL: https://www.aera.net/Newsroom/Study-Algorithms-Used-by-Universities-to-Predict-Student-Success-May-Be-Racially-Biased.

[2] Abeba Birhane et al. "The Values Encoded in Machine Learning Research". In: *arXiv preprint arXiv:2106.15590* (2022). URL: https://arxiv.org/pdf/2106.15590.pdf.

[3] Will Douglas Heavena. "Predictive policing algorithms are racist. They need to be dismantled." In: *MIT Technology Review* (2020). URL: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/.

[4] Rabie El Kharoua. *Students Performance Dataset*. 2024. DOI: 10.34740/KAGGLE/DS/5195702. URL: https://www.kaggle.com/ds/5195702.

[5] Bureau of Labor Statistics. "Chart: Earnings and Unemployment Rates by Educational Attainment, 2023". In: (2023). URL: https://www.bls.gov/emp/chart-unemployment-earnings-education.htm.

[6] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453. URL: https://www.science.org/doi/10.1126/science.aax2342.