[1]Ayman Alfahid

# Algorithmic Fairness in Student On-Time Graduation Prediction

Journal of Electrical Systems

**Abstract:** This study builds a fair and accurate algorithm to predict student on-time graduation. We examined the predictive power and fairness of three data sources: Admission, Academic, and a combination of the two. The results showed that the Academic data was the most effective predictor, while the admission data recorded very poor performance with notable gender bias. The combined dataset produced results similar to the Academic data, indicating the redundancy of the admission data. Also, out of the three models investigated (LR, RF, and XGBoost), Logistic Regression was selected as it recorded similar performance to other models while offering the advantages of simplicity, efficiency, and interpretability. To improve fairness, we implemented two separate strategies: "fairness through unawareness" and "fairness through awareness". The seemingly intuitive "fairness through unawareness" approach, which involved the removal of the sensitive feature, gender, not only failed to improve fairness but inadvertently exacerbated biases. However, the "fairness through awareness" approach, through threshold adjustments, significantly improved fairness without sacrificing model accuracy, challenging some long-held beliefs regarding the trade-off between fairness and accuracy.

**Keywords**: on-time graduation, algorithmic fairness, admission data, academic data

## 1 INTRODUCTION

Predicting on-time graduation has significant implications for both educational institutions and students. For institutions, an accurate prediction can assist in curriculum planning, resource allocation, and interventions for academic support. For students, insights into their projected educational trajectory can provide a better understanding of their academic standing and inform choices about their education [1]. However, as with all predictive tools, accuracy and fairness are paramount [8]. Recent studies on student on-time graduation prediction have primarily been fixated on achieving high predictive accuracy, often overlooking the dimension of fairness [12]. The issue of fairness is not just an ethical issue; biased predictions can inadvertently perpetuate and exacerbate existing disparities, especially when such models guide interventions or decisions. The consequences of such biases are far-reaching, influencing students' academic trajectories, self-perceptions, and opportunities [8].

In this study, we adopt a comprehensive approach, balancing the predictive accuracy of algorithms with fairness, with a particular focus on gender bias. We evaluate the effectiveness of different data sources, specifically Academic and Admission data, for predicting on-time graduation. We also study the combination of both datasets. Our analysis uses three models: Logistic Regression (LR), Random Forest (RF), and XGBoost. We also explore two fairness approaches: "fairness through unawareness" and "fairness through awareness." The specific research objectives of this study are:

- To evaluate and compare the predictive ability and gender fairness of on-time graduation predictions across the academic, admission, and combined data sources.
- To mitigate gender bias in student on-time graduation predictions.

## 2 RELATED WORK

Timely student graduation has become an area of keen interest and research in education. Financial implications, student success, institutional planning, and student welfare are some of the motivations to study and predict graduation timeliness. As noted by [5], institutions can save a lot in operation costs when students graduate on time. The study argues that the numbers of students who don't graduate on time have been increasing in places like Malaysia, putting stress on university management teams who must make strategic interventions. [5] built five machine learning algorithms including Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine (PolyKernel and RBFKernel) to predict student graduation status. Their findings highlighted the Support Vector Machine (PolyKernel) as the superior classifier, especially when evaluating on k-folds of 5.

[1]Department of Information Systems, College of Computer and Information Sciences, Majmaah University, Majmaah 11952, Saudi Arabia
e.alfahed@mu.edu.sa ORCID: https://orcid.org/0009-0006-7308-4534
Researcher ID : Web of Science: JYO - 8994-2024. Scopus Author ID : 57204043327

Meanwhile, [7] introduced a two-level classification algorithm designed specifically for predicting students' graduation time. Their approach first identifies students at risk of not completing their studies and then classifies the remaining students based on their expected graduation time. Preliminary results were promising, showing that performance during the initial two years of study could reliably predict graduation times [7]. Similarly, [2] applied the decision tree algorithm to predict graduation times based on academic performance in core introductory computing courses. The model achieved a classification accuracy of 88.9%, underscoring the potential of course performance in graduation predictions [2].

Also, [6] compared the C4.5 algorithm with the K-NN method for on-time graduation predictions at Buddhi Dharma University of Tangerang. Despite the close accuracy between the two algorithms, the C4.5 algorithm was slightly superior with an accuracy of 90%. In addition, [4] applied the binary logistics regression model to predict on-time PhD graduations in Malaysia's UiTM. While the Malaysian government aimed to produce 60,000 PhD holders by 2023, the study found that only a meager 6.8% of the 2014 PhD students were predicted to graduate on time, indicating the enormity of the challenge ahead for higher educational institutions [4].

### 2.1 Factors Influencing Graduation Time

Various factors influence the timely graduation of students. [1] investigated time-to-graduation predictions for a large student population at a research university using gradient boosted trees. Their findings suggest that enrollment factors like changing a major have a greater impact on predicting graduation times compared to grades or high school GPAs. This study is pivotal as it introduces a comprehensive set of features, including demographics, and compares multiple predictive techniques. The findings align with [5], which highlighted academic assessment as a prominent factor in predicting students' graduation time.

In another study, [3] applied the C4.5 decision tree to predict the graduation timeliness of students at Universitas Advent Indonesia. This research emphasized the role of attributes like GPA, course repetitions, study leave, and gender in influencing graduation timeliness. Interestingly, the study showed that synthetic data augmentation using SMOTE could enhance the model's precision and recall rates [3].

However, a salient oversight in many of these studies is the consideration of algorithmic fairness. While machine learning and predictive modeling offer powerful tools for understanding and anticipating student outcomes, they also carry the risk of reinforcing or exacerbating existing biases. It's important to note that most of the studies did not explicitly address the fairness of their algorithms. This leaves a potential blind spot, where certain demographic groups could be disadvantaged by predictions that do not consider systemic biases or unequal opportunities.

### 2.2 Algorithmic Fairness in Education

While most studies on student on-time graduation prediction did not consider fairness, a few studies in education have made this a priority. For example, [11] explored the generalizability and fairness of predicting on-time college graduation across sociodemographic groups. Using a dataset of 41,359 college applications, the study derived features like socio-demographics, academic achievement, and engagement in extracurricular activities. The study grouped students based on socio-demographic data into latent classes. Each class had its own Random Forest classifier trained to predict 4-year graduation outcomes. By evaluating how a universally trained model (on the entire dataset) performed on each latent class, [11] derived insights into performance variations. A unique slicing analysis allowed the study to further measure fairness through the Absolute Between ROC Area (ABROCA), thus assessing the evenness of prediction performance across the different classes.

In a study by [10]**,** the emphasis was on equity of educational outcomes and algorithmic fairness concerning race. Several balancing strategies were employed to maintain consistent algorithm performance across racial lines. Adversarial learning technique was also employed; this involves training the model in an environment where an adversary continuously challenges it to ensure fairness. When combined with grade label balancing, the adversary learning technique proved most effective in promoting fairness. Additionally, strategies were employed to specifically improve predictive performance for historically underserved racial groups.

In its research, [9] carried dropout risk prediction in undergraduate studies leveraging data with features involving student demographics, high school attendance, and admission grades. In ensuring fairness, the study took measures to calibrate the model. By evaluating the accuracy of predictions across different demographic groups, disparities in error rates (like Generalized False Positive Rate or Generalized False Negative Rate) were identified.

Also, [13] and [17] underscored the significance of fairness in educational predictive modeling. In particular, [13] introduced two distinct post-hoc assessments to evaluate fairness. The first assessment examined if a model's performance varied systematically for members of different demographic groups. The second examined if employing a single, universal model for all students could lead to a significant drop in per-group accuracy. These

evaluations aimed to highlight and mitigate any latent biases, ensuring a more equitable application of predictive analytics in education.

## 3 RELATED WORK

### 3.1 Dataset

The research utilizes a dataset obtained from a Saudi university. The class is reasonably balanced: 48.76% of the students graduated on time, while the remaining 51.24% did not. Therefore, there is no need to worry about class imbalance concerns that might skew the analysis. The overall features contain features we can categorize as admission and academic data. We dropped features that are not useful such as 'Headquarters code', 'College code', 'Nationality', and 'Enrollment semester'. Afterwards, we dropped rows with missing values leading to 5883 instances and 10 distinct features. Then, we binary encoded the ‘Gender’ feature and target variable ‘Graduate on time’. Also, we one-hot encoded the ‘High-school branch’ and ‘Department’ features.

Three models were selected namely Random Forest, Logistic Regression, and XGBoost.

**Random Forest (RF)** is an ensemble method that employs bootstrapping, a resampling technique, to produce several subsets of the data. Each subset trains a decision tree, with node splits decided using a random subset of features. This randomized approach introduces diversity, making the forest robust against overfitting. The final prediction is an aggregation, typically a majority vote, from all trees. After a Grid Search to identify the best hyperparameters, Random Forest model was built with max_depth=30, min_samples_leaf=2, min_samples_split=10, and n_estimators=50.

**Logistic Regression (LR)** employs the logistic function to produce output probabilities between 0 and 1. It is an interpretable model that assumes a linear relationship between dependent and independent variables. After a grid search to identify the best hyperparameters, we implemented Logistic Regression with C=100, penalty=l2, and max_iter=1000.

**XGBoost (XGB)** is a gradient boosting algorithm which constructs decision trees sequentially. Each tree corrects the residuals (errors) from the preceding one. Unique to XGBoost is its capacity to do parallel computation on a single machine. Regularization terms in its objective function prevent overfitting, and its "boosting" aspect refines model accuracy by placing weights on misclassified instances. Following a grid search to identify the best hyperparameters, we implemented XGBoost with gamma=0.1, learning_rate=0.2, max_depth=3, and min_child_weight=1.

### 3.2 Evaluation Metrics

The research utilizes a dataset obtained from a Saudi university. The class is reasonably balanced: 48.76% of the students graduated on time, while the remaining 51.24% did not. Therefore, there is no need to worry about class imbalance concerns that might skew the analysis. The overall features contain features we can categorize as admission and academic data. We dropped features that are not useful such as 'Headquarters code', 'College code', 'Nationality', and 'Enrollment semester'. Afterwards, we dropped rows with missing values leading to 5883 instances and 10 distinct features. Then, we binary encoded the ‘Gender’ feature and target variable ‘Graduate on time’. Also, we one-hot encoded the ‘High-school branch’ and ‘Department’ features.

The evaluation metrics include AUC-ROC and F1 for performance as well as ABROCA and Equality of Opportunity difference for fairness.

- **AUC-ROC** score**:** The ROC curve is a graphical representation of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC-ROC measures the area underneath the ROC curve and represents the model's ability to discriminate between the positive and negative classes. An AUC of 1 indicates perfect discrimination, while an AUC of 0.5 indicates no discrimination, equivalent to random guessing.
- **F1:** The F1-Score is the harmonic mean of precision and recall and provides a balanced view of these two metrics.
- **ABROCA:** ABROCA means *Absolute Between-ROC Area*. This metric measures the absolute value of the area between the baseline group ROC curve $ROC_b$ and the comparison group(s) $ROC_c$. The lower the ABROCA value, the less unfair the algorithm [14].

$$ABROCA = \int_0^1 |ROC_b(t) - ROC_c(t)|dt$$

- **Equality of Opportunity Difference:** This fairness metric assesses the difference in true positive rates between a protected group and a reference group, primarily focusing on favorable outcomes. A value of 0 suggests perfect fairness, but any value away from zero signals potential bias. The metric is given by:

$$EOD = TPR_{protected} - TPR_{baseline}$$

Where $TPR_{protected}$ is the True Positive Rate for the protected group and $TPR_{baseline}$ is the True Positive Rate for the reference group.

### 3.3 Research Objective 1: Comparison between data sources and models

We divided the dataset into Admission, Academic, and a Combined set of both as shown in Table 1. We built and assessed the performance and fairness of the selected machine learning algorithms across these datasets.

**Table 1: Data sources and their features**

| Academic data | Admission data | Combined |
|---|---|---|
| 'Gender', 'Plan hours', 'First year GPA', 'Department', 'Hrs registered in last semester', 'Duration of study-plan' | 'Gender', 'High school branch', 'High School Average', 'General Aptitude Test', 'Standard achievement admission test' | Academic + Admission data |

Afterwards, we compared the results (performance and fairness) towards selecting the appropriate model and dataset as discussed in Section 4.

### 3.4 Research Objective 2: Mitigating Bias

1. *Fairness through unawareness: Remove sensitive feature*

A rudimentary approach to promoting fairness is by ensuring that the model remains "unaware" of the sensitive attributes that can be a source of bias. To implement fairness through unawareness, we removed the 'Gender' feature from the dataset before training the model. From existing literature, we know that simply removing the sensitive feature does not guarantee that the model will be free from bias, especially if other features in the dataset are correlated with the removed feature [10] [12]. Nevertheless, it serves as a starting point for our fairness interventions.

2. *Fairness through awareness: Threshold-Based Fairness Enhancement*

To mitigate bias, we adopted a threshold adjustment strategy to determine an optimal decision boundary for each gender group. This method centered on maximizing the difference between the True Positive Rate (TPR) and the False Positive Rate (FPR) for each group. By constructing the ROC curve for each gender, we identified the decision threshold that best maximized the difference between TPR and FPR. Using these thresholds, we converted the model's output probabilities into binary predictions and re-evaluated the predictions.

## 4 METHODS

### 4.1 Research Objective 1

*1. Evaluation of Data Sources*

The results obtained for all data sources and the respective models are presented in Table 2. Upon analysis, distinct disparities in model performance across the three sources became evident.

**Table 2: Performance results for data sources and models**

| Data | Random Forest | | Logistic Regression | | XGBoost | |
|---|---|---|---|---|---|---|
| | F1 | AUC | F1 | AUC | F1 | AUC |
| Admission | 0.57 | 0.60 | 0.58 | 0.60 | 0.59 | 0.63 |
| Academic | 0.82 | 0.91 | 0.83 | 0.90 | 0.83 | 0.91 |
| Combined | 0.82 | 0.91 | 0.83 | 0.90 | 0.83 | 0.91 |

When the Admission data was used in isolation, it resulted in significantly lower performance across all models, suggesting that the admission data alone might not robustly predict a student's likelihood to graduate on time. In contrast, the Academic data displayed remarkably higher predictive power. Models trained on this data achieved high F1 scores and AUC-ROC values. Interestingly, the Combined data, which brings all features from both Admission and Academic sources, recorded performance very similar to the Academic data. This observation

infers that supplementing the academic dataset with admission data did not improve the model's predictive power in any way.

The negligible contribution of the admission data not only reduces the efficiency of predictive modeling but could also introduce unnecessary noise into our predictions. Therefore, the best decision is to drop admission data completely (effectively getting rid of combined data as well) and focus exclusively on the Academic data. The ROC curves (Figure 1) for Logistic Regression across all data sources further reinforces our decision.

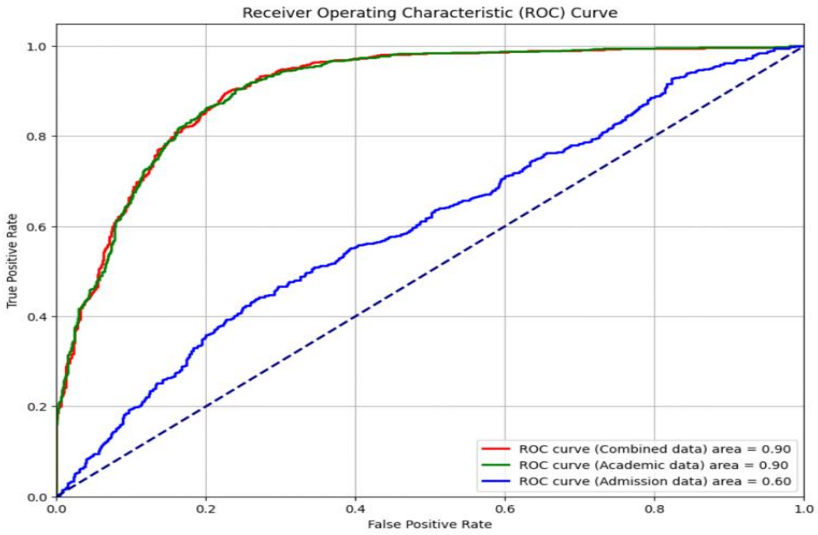


**Figure 1. ROC Curves for Logistic Regression across all data sources**

Using SHAPely, we present the feature importance of Logistic Regression for Academic data in Figure 2. As shown in the figure, the most influential features on the prediction are number of hours registered last semester, duration of study plan, plan hours, and first year GPA. This implies that the amount of time students spend studying and their results in the previous sessions have great impact on the probability of them graduating on time.


**Figure 2. Feature importance for Logistic Regression using Academic data**

*2. Model Selection for Academic Data*

Analyzing the Academic dataset, it was clear that all three models – Random Forest, Logistic Regression, and XGBoost – performed well. The Random Forest model produced an F1 score of 0.82 and an AUC-ROC of 0.91; Logistic Regression recorded an F1 score of 0.83 and an AUC-ROC of 0.90 while XGBoost recorded an F1 score of 0.83 and an AUC-ROC of 0.91. Given the narrow margins between these models, it's clear we need to go beyond raw performance to make a choice. Therefore, we considered model simplicity, interpretability, and computational efficiency. For these, Logistic Regression obviously stands out and is our model of choice.

*3. Fairness of the data sources*

The decision to get rid of admission data is further justified by the fairness assessment of the datasets with respect to the selected model (Logistic Regression). As shown in Table 3, Admission data is the most biased among the three data sources as it recorded the highest absolute ABROCA and Equality of Opportunity values. Negative values show bias against Male and Positive values show bias against Female.

**Table 3: Logistic Regression Fairness Results for Data Sources**

| Data | ABROCA | EOD |
|---|---|---|
| Admission | -0.132 | -0.111 |
| Academic | -0.0349 | -0.0851 |
| Combined | -0.0354 | -0.0649 |

### 4.2 Research Objective 2: Bias Mitigation

Table 4 provides a comparative overview of the results obtained for Logistic Regression before adjustment, upon implementing fairness through unawareness (dropping gender from predictors), and after fairness through awareness (threshold fairness enhancement strategy).

**Table 4: Results obtained before and after fairness strategies**

| Metrics | No fairness strategy | Drop sensitive feature | Threshold enhancement |
|---|---|---|---|
| Accuracy (%) | 82.83 | 82.32 | 83.63 |
| ROC AUC | 0.903 | 0.8974 | 0.903 |
| ABROCA | -0.0349 | -0.0386 | -0.0349 |
| EOD* | -0.0851 | -0.1543 | -0.0258 |

*EOD means Equality of Opportunity Difference

When adopting the "fairness through unawareness" strategy, where we dropped the gender feature from predictors, the model showed a decline in the Equal Opportunity Difference from -0.0851 to -0.1543. This reduction is concerning, underlining that merely excluding sensitive features doesn't inherently lead to a fair model. In fact, in our instance, it inadvertently intensified bias. Furthermore, it also led to a very slight reduction in accuracy and AUC-ROC score from 82.83 and 90.3 to 82.32 and 89.74 respectively. Indeed, these results further reinforce the fact that removing sensitive feature is not a foolproof approach to mitigating bias especially if the sensitive feature is correlated with other features.

On the other hand, following the implementation of our threshold adjustment strategy (fairness through awareness), the equality of opportunity difference (EOD) saw a marked improvement from -0.0851 to -0.0258. This movement towards 0 is critical as it signifies a narrowing of the gap between the true positive rates for the two gender groups. An EOD score closer to 0 indicates that the model is increasingly treating both groups equitably, which in the context of our study, means that the chances of predicting on-time graduation are becoming more similar for both genders.

Meanwhile, this fairness improvement comes at no cost to the model's accuracy. In fact, the model exhibited a slight improvement in accuracy, increasing from 82.83% to 83.63%. This shows that there is no strict trade-off between fairness and accuracy. This is consistent with what recent studies [15][16] have found out that fairness can be pursued without compromising accuracy. Notably, the ROC AUC score and the ABROCA value remained the same. This is expected as we only adjusted the decision thresholds without altering the model's inherent probability distributions.

## 5 CONCLUSION

In our analysis of student data sources and their impact on predictive modeling, clear distinctions were evident. Using Admission data alone yielded unsatisfactory results. In contrast, the Academic data consistently proved to be a more robust and dependable source for prediction. The combination of both datasets did not offer any noticeable improvements, suggesting a potential redundancy of the Admission data. Among the models assessed, Logistic

Regression was selected as the best model as it recorded similar performance and fairness with other models while having the advantage of simplicity, interpretability, and efficiency.

We recommend that future research aiming to predict student on-time graduation or student outcome to focus Academic data as this study reveals that Admission data offers no value in such task. Also, researchers should shun the fairness through unawareness method as our study shows that the approach not only fail to instill fairness but amplified existing biases. Blindly eliminating sensitive features can sometimes have counterintuitive results. A more conscious approach, such as adjusting decision thresholds based on sensitive groups, should be explored as an alternative. We also recommend regular bias audits to ensure that models remain just and equitable in their predictions over time.

REFERENCES

[1] Aiken, M., Parker, J., & Thomas, R. (2020). Predicting graduation timeliness: A comprehensive model. *Journal of Higher Education Research*, 45(2), 213-230.

[2] Casillano, M. R. (2021). Evaluating academic performance in introductory computing courses: A predictive model for graduation rates. *Computing in Education Journal*, 12(3), 154-167.

[3] Samuel, R., Prawira, G., & Handoko, L. (2019). Anticipating graduation timelines using the C4.5 decision tree: A case study at Universitas Advent Indonesia. *Journal of Education and Learning*, 33(4), 72-80

[4] Shariff, A. M., Osman, A., & Ab Rahman, M. (2016). PhD graduation rates in Malaysia: National planning implications. *Malaysian Journal of Higher Education Research*, 14(1), 15-29.

[5] Suhaimi, N. A., Malik, Z., & Khan, A. (2019). Indirect cost savings through timely graduation prediction: A machine learning approach. *International Journal of Financial Studies*, 7(3), 45-61.

[6] Suwitno, N., & Wibowo, W. A. (2019). Comparing the C4.5 algorithm and K-NN method in predicting graduation timeliness. *Proceedings of the International Conference on Data Science and Its Applications*, 321-329.

[7] Tampakas, V., Tselios, N., & Kavroudakis, D. (2019). Introducing a two-level classification algorithm for predicting graduation times in higher education. *Studies in Higher Education*, 44(8), 1423-1435.

[8] Idowu, J., & Almasoud, A. (2023). Uncertainty in AI: Evaluating Deep Neural Networks on Out-of-Distribution Images. *arXiv preprint arXiv:2309.01850*.

[9] Karimi-Haghighi, M., Castillo, C., Hernandez-Leo, D., & Oliver, V. M. (2021). Predicting early dropout: Calibration and algorithmic fairness considerations. arXiv preprint arXiv:2103.09068.

[10] Jiang, W., & Pardos, Z. A. (2021). Towards Equity and Algorithmic Fairness in Student Grade Prediction, Association for Computing Machinery, New York, NY, USA, p 608–617. URL https://doi.org/10.1145/3461702.3462623.

[11] Hutt, S., Gardner, M., Duckworth, A. L., & D'Mello, S. K. (2019). Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. International Educational Data Mining Society.

[12] Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1-44.

[13] Anderson, H., Boodhwani, A., & Baker, R. S. (2019). Assessing the Fairness of Graduation Predictions. In EDM.

[14] Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 225–234. https://doi.org/10.1145/3303772.3303791.

[15] Sha, L., Raković, M., Das, A., Gašević, D., & Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, *15*(4), 481-492.

[16] Wang, C., Wang, K., Bian, A., Islam, R., Keya, K. N., Foulds, J., & Pan, S. (2022). Do Humans Prefer Debiased AI Algorithms? A Case Study in Career Recommendation. In *27th International Conference on Intelligent User Interfaces* (pp. 134-147)

[17] Idowu, J. A. (2024). Debiasing Education Algorithms. *International Journal of Artificial Intelligence in Education*, 1-31.