On the Fairness of Ensemble Learning Methods in Student Dropout Prediction

Abdelghafour Aboukacem(✉), Loubna Mekouar, El Houcine Bergou, Youssef Iraqi, and Ismail Berrada

College of Computing, Mohammed VI Polytechnic University, Ben Guerir, Morocco
{abdelghafour.aboukacem,loubna.mekouar,elhoucine.bergou,youssef.iraqi,ismail.berrada}@um6p.ma

Abstract. Student Dropout Prediction (SDP) is a crucial task in Educational Data Mining (EDM). It aims to help institutions intervene early to improve retention. Despite notable advancements in the predictive performance of the developed models, the topic of fairness is usually overlooked. Algorithmic bias can lead to disparate impacts across demographic groups. In this study, we explore the use of ensemble learning, specifically bagging, boosting, voting, and stacking techniques to improve both the predictive performance as well as the fairness aspect. Using a real-world dataset that spans the whole K-12 system of Morocco with 14 sub-datasets, we evaluate various Machine Learning (ML) models trained using the previously mentioned ensemble learning techniques. Our study examines model fairness for several protected attributes, including gender, handicap, financial aid, and boarding school availability.

Keywords: Student dropout prediction · Fairness · Ensemble learning

1 Introduction

With the recent advancements in learning technologies [24], the education field has been undergoing substantial improvements in learning quality. Leveraging Educational Data Mining (EDM), data mining techniques are being applied to extract meaningful insight to further improve key objectives[17]. These advancements span from personalized and adaptive learning [25], to curriculum optimization and policy decision improvement [7]. However, a growing interest is being dedicated toward Student Dropout Prediction (SDP) [1,20], driven by the severity of the issue [5,6]. To this extent, Dropout Early Warning Systems are being developed to predict the possibility of students dropping out. These systems are based on Machine Learning (ML) Deep Learning (DL) models and are based on datasets that span the academic, socio-economic, and demographic aspects of the students. A variety of models, such as eXtreme Gradient Boosting (XGB), Light Generalized Linear Model (LGBM), are being used to predict students at risk [8]. Studies also explored even more complex models such as Large Language Models (LLMs) to predict and explain student dropouts [2].

A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 364–371, 2025.
https://doi.org/10.1007/978-3-031-98462-4_46

However, despite their predictive performance, the topic of fairness, transparency, and ethical accountability in SDP remains underexplored. Biased dropout prediction models risk reinforcing systemic inequalities, particularly for marginalized and under-resourced students. The fairness of AI in education is not just a technical issue but an ethical imperative, influencing student agency, teacher interventions, and institutional decision-making.

On the other hand, ensemble learning [16] has been known to improve the performance of simple machine learning models. This method aggregates several weaker models to create a more performant model. The aggregation can be done is a variety of ways. The most common ones are bagging [4], boosting [9], stacking [23], and voting [22]. However, the question of the fairness of the ensembled classification models is still unexplored, especially in the case of SDP.

In our work, we aim to explore the use of ensemble learning techniques, specifically bagging, boosting, stacking, and voting, in improving both the predictive performance as well as minimizing the disparities for several protected groups. Our experiment is based on a real life dataset that spans the whole K-12 system of Morocco. Currently over three hundred thousand students dropout each year in the Moroccan K-12 system[1], this calls for accurate and fair predictive methods to mitigate the issue. To summarize, the main contributions of this paper are as follows:

- **First exploration of ensemble learning for fairness**: To the best of our knowledge, this is the first study that investigates the effects of ensemble learning on model fairness for student dropout.
- **Multiple protected groups**: We consider four protected groups in our study based on gender, mental and physical handicaps, availability of boarding school, and financial aid. This diversity in protected groups is rare among prior research.
- **Evaluations on real-life dataset**: We compare the predictive performance as well as the fairness of ensemble learning methods to those of ML models. We provide a comprehensive comparison through real-life dataset of the Moroccan educational K-12 system.

The rest of this paper is structured as follows, Sect. 2 showcases previous studies on fairness for educational settings, Sect. 3 provides details on our experiment detailing important information about the dataset used and performance metrics. Section 4 presents our findings comparing the ensemble learning methods to the base ML models in terms of predictive performance and fairness. Finally, Sect. 5 provides a comprehensive conclusion and future work opportunities.

2 Related Work

Prior research has explored the topic of fairness in educational machine learning models through a variety of metrics and bias mitigation strategies. Several works

[1] https://www.moroccoworldnews.com/2023/06/32538/over-300-000-students-drop-out-of-school-each-year-in-morocco/.

studies the fairness of these models in broader areas than dropout prediction such as in language assessment [13], success prediction [3,12,27], and form post classification [19] based on gender.

For SDP, studies have measured fairness between groups based on features such as gender and race [10,11]. A variety of metrics are being used to measure the fairness of these models which include Absolute Between-ROC Area (ABROCA) first introduced in [10]. This metric computes the absolute difference between the Area Under the ROC Curve (AUC) of each group for a protected attribute. Moreover, the authors in [11] base their evaluations on discrimination, which measures the difference in model performance across demographic groups, and consistency, which evaluates how similar predictions are for similar individuals across groups. The work in [12] utilizes demographic parity, which ensures that predictions are independent of sensitive attributes, and equality of opportunity which in turn ensures equal true positive rates across demographic groups. Other studies rely on classification performance metrics such as accuracy, recall, True Negative Rate (TNR) [26], and AUC [19].

To mitigate these disparities, studies used data resampling strategies [18,19] which aim to train predictive models on a balanced set based on the studied sensitive feature. Adjusting the training set's sample weights is another idea previously explored in [26].

In educational contexts, ensemble machine learning was explored to overcome a variety of issues. For example, ensembled K-Nearest Neighbors (KNN), Neural Network (NN), and other models were used in [14] for curriculum recommendation for higher education. Ensemble learning approaches are also used to perform cheat detection [28]. Moreover, combining models such as XGB, LGBM, Random Forest (RF), and more to predict student failure along with suggesting learning paths is explored in [21].

3 Experiment Setting

3.1 Dataset

The dataset is an anonymized collection from Morocco's Ministry of National Education, Preschool, and Sports[2], it contains academic, demographic, and socio-economic features. Notably, the dataset pertains specifically to public-sector institutions within the Fes-Meknes region. Furthermore, it covers student records from the 2015/2016 academic year to the 2020/2021 academic year of 14 different educational levels and study options. We split our dataset into train and test sets by considering records from the 2015/2016 to 2019/2020 school years as the training set, and the records from the 2020/2021 school year as the test set.

In our fairness analysis we consider four different sensitive features which we find in our dataset: gender, handicaps, boarding school availability, and financial aid. It is worth noting that in some educational levels, we do not find records

[2] https://www.men.gov.ma/.

of particular groups. This in turn, renders the usage of resampling techniques useless.

3.2 Experiment

To evaluate the fairness of ensembled models, we consider cases of the most common ensemble learning techniques. For Bagging, we evaluate bagging of Decision Trees (DT) classifier, considering lower max depth trees (depth of 2) and a high number of estimators (100 estimators) as suggested in the literature [15]. For Boosting, we evaluate the effectiveness of boosted DT, employing Adaptive Boosting (AB) as a booster model with 100 estimators. For Stacking, we evaluate stacking DT,Logistic Regression (LR), and Naïve Bayes (NB) models, using LR as the meta-learner. Finally, for Voting, we consider both hard and soft voting using the three base models: DT, LR, and NB. We compare the four methods to the base models by considering AUC, class-specific precision, and recall for model performance. Additionally, we measure the fairness of the models using the ABROCA metric, defined as the absolute difference between the AUCs of each group, as well as the Demographic Parity (DP) between each group pair.

4 Results and Discussions

Table 1 showcases the results of the predictive performance of the tested approaches. Table 2 provides details on model fairness measurements using the ABROCA metric.

The results from Table 1 indicate that stacking the predictive models as well as the base LR model performed the best, achieving the highest AUC of 0.80. These models were followed closely by the boosted and bagged trees with an AUC score of 0.79. These models provide the best balanced performance in identifying at risk students. In terms of dropout precision, denoted as D precision, boosted trees outperformed the other models with a precision 0.35, indicating that when the model predicts a student as a dropout, it is likely correct. On the other hand, hard voting achieved the highest score on non-dropout precision, denoted as ND precision, making it more reliable for students who will persist. As far as recall goes, soft voting and NB both achieved the highest score in detecting actual dropout cases with 0.81 and 0.83 dropout recalls respectively. However, boosted trees achieved the highest score, of 0.79, for non-dropout recall, denoted ND recall, making it effective at correctly identifying those who are likely to persist.

Overall, stacking and LR emerged as the best-performing models based on AUC, while boosted trees excelled in dropout precision and non-dropout recall. For dropout identification, soft voting provided the best rounded performance while providing an edge in dropout detection with the higher dropout recall value.

After examining the disparities for all four sensitive features using the ABROCA metric (lower values are better). From Table 2, the results can be summarized in the following:

Table 1. Performance comparison of different methods.

Method	AUC	D Precision	ND Precision	D Recall	ND Recall
Soft Voting	0.77	0.23	0.95	**0.81**	0.56
Hard Voting	0.73	0.27	**0.95**	0.79	0.66
Stacking	**0.80**	0.29	0.94	0.72	0.73
Boosting	0.79	**0.35**	0.91	0.61	**0.79**
Bagged Trees	0.79	0.32	0.94	0.76	0.69
DT	0.76	0.33	**0.93**	0.74	0.68
LR	**0.80**	**0.33**	0.93	0.71	**0.73**
NB	0.54	0.18	0.88	**0.83**	0.25

- **Boarding School**: Boosted trees and LR achieved the best results of an ABROCA average of 0.0190 and 0.0208 respectively, exhibiting the lowest bias and indicating more equitable predictions for students from different schooling backgrounds.
- **Gender**: NB was the fairest model in terms of gender with an averaged ABROCA of 0.0182 and a DP score of 0.0232, suggesting minimal prediction disparities across male and female students.
- **Handicap**: Bagged trees along with LR achieved the lowest bias with averaged ABROCA scores of 0.1019 and 0.1024 respectively. Bagged trees however, exhibit slightly better performance in DP. This indicates their fairer predictive ability for students with disability.
- **Financial Aid**: NB exhibited the lowest bias with an ABROCA score of 0.0119, closely followed by boosted trees with an ABROCA score of 0.0283, suggesting that these models are the least sensitive to students' financial aid status.

Table 2. Fairness comparison across different methods using ABROCA and DP.

Method	Boarding School		Gender		Handicap		Financial Aid	
	ABROCA	DP	ABROCA	DP	ABROCA	DP	ABROCA	DP
Soft Voting	0.0221	0.0350	0.0390	0.1806	0.1141	0.3261	0.0241	0.0266
Hard Voting	0.0230	0.0259	0.0354	0.2169	0.1302	0.1851	0.0305	0.0477
Stacking	0.0193	0.0255	0.0354	0.2077	0.1153	0.2148	0.0344	0.0611
Boosting	**0.0190**	0.0524	0.0323	**0.1078**	0.1284	0.1652	**0.0283**	**0.0105**
Bagged Trees	0.0282	**0.0248**	**0.0319**	0.2155	**0.1019**	**0.1265**	0.0308	0.0747
DT	0.0238	**0.0219**	0.0325	0.2114	0.1084	**0.1324**	0.0335	0.0850
LR	**0.0208**	0.03266	0.0358	0.2103	**0.1024**	0.1761	0.0378	**0.0176**
NB	0.0321	0.0358	**0.0182**	**0.0232**	0.1166	0.4962	**0.0119**	0.0282

4.1 Ensemble Learning and Fairness

The findings suggest that ensemble learning methods can enhance fairness. However, the evidence is not strong enough to conclude that ensemble methods inherently provide better fairness overall. Although methods like bagging and boosting for trees demonstrated lower bias in certain sensitive attributes, fairness outcomes varied across ensemble methods and were not consistently superior to certain models like LR. We can summarize the findings of our experiment in the following points:

1. Boosted and Bagged trees showed enhanced fairness for certain attributes, e.g. disability and financial aid, however, other ensemble methods such as voting and stacking did not consistently outperform simpler models in fairness. Moreover, LR, a non-ensemble learning model performed comparably to ensembles in terms of fairness across multiple sensitive attributes. This challenges the assumption that ensembles are always fairer.
2. Some ensembles such as hard voting exhibited the highest bias for students with handicaps, indicating that simply aggregating model decisions may reinforce biases from individual models. On the other hand, despite the best AUC of 0.80, stacking did not significantly improve the fairness. This indicates that even an advanced ensemble method does not automatically mitigate disparities.
3. Averaging the fairness of each model over all the protected features reveals that boosted trees, the best ensembled model in terms of fairness, scored an average ABROCA of 0.052. Comparing it to LR, the best performing model, which scored an average ABROCA of 0.049, enforces that simpler models can achieve better fairness with similar predictive performance.

5 Conclusion and Future Work

In this study, we investigated the impact of ensemble learning techniques such as bagging, boosting, voting, and stacking on both the predictive performance as well as the fairness in student dropout prediction. Our analysis was conducted on a real-world dataset of the Moroccan K-12 education system. We compared the ensemble learning methods to the base line machine learning models.

Our findings suggest that while ensemble learning improve the predictive performance, its effect on the fairness is inconsistent. Specifically seeing that methods such as the results varied from one protected group to another. For example we noticed that bagged trees demonstrated the lowest bias for students with disabilities, while boosted trees exhibited the least bias for boarding school availability. Moreover, not all ensemble learning methods improved fairness, techniques such as stacking and voting reinforced biases for students with disabilities. This non uniformity in results across attributes raises critical concerns regarding the ethical deployment of ensemble SDP models. Should institutions prioritize ensemble learning techniques if they do not consistently improve fairness? Our

results suggest that simply increasing model complexity does not guarantee equitable predictions.

As suggestions for future work, investigations on fairness-aware ensemble learning that incorporate fairness constraints into ensemble techniques through data pre-processing for example. A broader model selection for such studies should be taken into consideration. In addition, incorporating more fairness metrics such as demographic parity, equalized odds, or counterfactual fairness could provide a more comprehensive analysis. Finally, integrating these models into real-world interventions and assessing their impact on retention strategies could provide deeper insights into their practical utility.

References

1. Abdul Bujang, S.D., Selamat, A., Krejcar, O., Mohamed, F., Cheng, L.K., Chiu, P.C., Fujita, H.: Imbalanced classification methods for student grade prediction: a systematic literature review. IEEE Access **11**, 1970–1989 (2023)
2. Aboukacem, A., Berrada, I., Bergou, E.H., Iraqi, Y., Mekouar, L.: Investigating the predictive potential of large language models in student dropout prediction. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 381–388. Springer, Cham (2024)
3. Anderson, H.J., Boodhwani, A., Baker, R.: Assessing the fairness of graduation predictions. In: Educational Data Mining (2019)
4. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
5. Campbell, C.: The socioeconomic consequences of dropping out of high school: evidence from an analysis of siblings. Soc. Sci. Res. **51**, 108–118 (2015)
6. De Witte, K., Cabus, S., Thyssen, G., Groot, W., van den Brink, H.M.: A critical review of the literature on school dropout. Educ. Res. Rev. **10**, 13–28 (2013)
7. Dutt, A., Ismail, M.A., Herawan, T.: A systematic review on educational data mining. IEEE Access **5**, 15991–16005 (2017)
8. Elbouknify, I., et al.: Student at-risk identification and classification through multitask learning: a case study on the Moroccan education system. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 372–380. Springer, Cham (2024)
9. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
10. Gardner, J., Brooks, C., Baker, R.: Evaluating the fairness of predictive student models through slicing analysis. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, LAK19, pp. 225–234. Association for Computing Machinery, New York (2019)
11. Hu, Q., Rangwala, H.: Towards fair educational data mining: a case study on detecting at-risk students. In: Educational Data Mining (2020)
12. Lee, H., Kizilcec, R.F.: Evaluation of fairness trade-offs in predicting student success (2020)
13. Loukina, A., Madnani, N., Zechner, K.: The many dimensions of algorithmic fairness in educational applications. In: Yannakoudakis, H., Kochmar, E., Leacock, C., Madnani, N., Pilán, I., Zesch, T. (eds.) Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–10. Association for Computational Linguistics, Florence (2019)

14. Nuankaew, W.S., Bussaman, S., Nuankaew, P.: Evolutionary feature weighting optimization and majority voting ensemble learning for curriculum recommendation in the higher education. In: Surinta, O., Kam Fung Yuen, K. (eds.) Multidisciplinary Trends in Artificial Intelligence. pp. 14–25. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20992-5_2
15. Plaia, A., Buscemi, S., Fürnkranz, J., Mencía, E.L.: Comparing boosting and bagging for decision trees of rankings. J. Classif. **39**(1), 78–99 (2022)
16. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. **33**(1–2), 1–39 (2010)
17. Romero, C., Ventura, S.: Educational data mining and learning analytics: an updated survey. WIREs Data Min. Knowl. Disc. **10**(3) (2020)
18. Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V.M., Gasevic, D., Chen, G.: Assessing algorithmic fairness in automatic classifiers of educational forum posts. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) Artificial Intelligence in Education, pp. 381–394. Springer, Cham (2021)
19. Sha, L., Raković, M., Das, A., Gašević, D., Chen, G.: Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. IEEE Trans. Learn. Technol. **15**(4), 481–492 (2022)
20. Shafiq, D.A., Marjani, M., Habeeb, R., Asirvatham, D.: Student retention using educational data mining and predictive analytics: a systematic literature review. IEEE Access **10**, 72480–72503 (2022)
21. Smirani, L.K., Yamani, H.A., Menzli, L.J., Boulahia, J.A.: Using ensemble learning algorithms to predict student failure and enabling customized educational paths. Sci. Program. **2022**, 1–15 (2022)
22. Tumer, K., Ghosh, J.: Error correlation and error reduction in ensemble classifiers. Connect. Sci. **8**(3–4), 385–404 (1996)
23. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992)
24. Xiong, Z., et al.: A review of data mining in personalized education: current trends and future prospects. Front. Digit. Educ. **1**(1), 26–50 (2024)
25. Xiong, Z., et al.: A review of data mining in personalized education: current trends and future prospects. Front. Dig. Educ. **1**(1), 26–50 (2024)
26. Yu, R., Lee, H., Kizilcec, R.F.: Should college dropout prediction models include protected attributes? (2021)
27. Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: evaluating different sources of student data. In: Educational Data Mining (2020)
28. An ensemble learning approach based on tabnet and machine learning models for cheating detection in educational tests. Educ. Psychol. Measur. **84**(4), 708–809 (2024)