Auditing Group-Level Error in Student Performance Prediction: Reweighting, Uncertainty and a Bharatiya Framework for Trustworthy Educational AI

Vipin Bihari Yadav[1,*], Chandan Singh[1], Shweta Chauhan[2]

[1] *Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Satellite Centre Amethi, Amethi 227413, Uttar Pradesh, India*

[2] *MPCAMS, Mandhana, Kanpur, Uttar Pradesh, India*

* Corresponding author. ORCID: 0009-0006-6672-6789

Email: vipin884060@gmail.com

Abstract

Machine-learning models used in schools and universities are usually judged by overall error, for example mean absolute error (MAE) or R^2. A model can look accurate on average and still make much larger errors for some groups of students, and the average hides it. This paper examines that gap with the UCI Student Performance data (Cortez, 2008), a public benchmark of 395 mathematics students from two Portuguese secondary schools. Linear Regression, Random Forest and XGBoost were trained to predict the final grade (G3) and were evaluated on a held-out set of 79 students and by five-fold cross-validation. Error disparity was measured across sex and home address with group-wise MAE, performance disparity (PD), the standard deviation of group MAEs, and their coefficient of variation. Random Forest was the most accurate model (test MAE = 1.165, R^2 = 0.820), yet it also had the largest sex-based disparity (PD = 0.468) and the smallest address-based disparity (PD = 0.057). Inverse-frequency reweighting over the four sex–address groups lowered sex-based PD for all three models, by 3.5% to 28.7%, without a consistent loss of accuracy. Paired bootstrap intervals on the small test set were wide, however, and every interval for the change in PD included zero, so these reductions are suggestive rather than established. Removing the earlier grades G1 and G2 changed the disparity pattern; for instance, XGBoost's sex-based PD rose from 0.350 to 0.399. A fairness audit therefore holds only for the feature set that was audited. Because the data are Portuguese, no finding is carried over to Indian students. The empirical work is a reproducible demonstration of an audit procedure. On that base the paper proposes the Bharatiya Trustworthy Educational AI (BTEAI) framework, an eight-component cycle that links fairness measurement with teacher oversight, contextual validation and the aims of India's National Education Policy 2020.

Keywords: algorithmic fairness; educational machine learning; student performance prediction; bias mitigation; sample reweighting; trustworthy AI; Bharatiya education; NEP 2020; performance disparity

1. Introduction

Machine-learning models now sit inside many parts of education. They predict grades, flag students who may drop out, recommend study material, score written answers and inform decisions about progression (Romero & Ventura, 2020; Zhai et al., 2021). When they work well, teachers can notice struggling students earlier and spend limited time where it is needed most.

That benefit rests on an assumption that is seldom tested: that a model is about equally accurate for every kind of student. If it is not, the consequences are practical. A student whose grade is over-predicted may miss out on extra help, and one whose grade is under-predicted may be flagged without reason or passed over for an opportunity. In admissions, scholarship allocation and early-warning systems, errors of this kind can do real harm (Baker & Hawn, 2022).

The specific problem taken up here is the gap between overall accuracy and accuracy within groups. Evaluation in educational data mining normally reports one MAE, one RMSE and one R^2 for the whole test set. A low overall MAE is compatible with a much higher MAE for one group, and averaging over everyone conceals it. The results in Section 5 show this happening: the model with the best overall accuracy is not the model with the smallest disparity between groups.

The issue is sharper in India, where the school system is both very large and very uneven. UDISE+ 2024–25 covers roughly 1.47 million schools (Department of School Education and Literacy, 2025b). Digital infrastructure shows the unevenness clearly. Nationally, the share of schools with a computer rose from 57.2% in 2023–24 to 64.7% in 2024–25, and the share with internet access from 53.9% to 63.5% (Department of School Education and Literacy, 2025a). State figures differ widely; in Bihar only 25.2% of schools reported computers in 2024–25 (Ministry of Education, n.d.). A model trained mostly on data from well-connected schools may behave differently elsewhere, and a system that has never been checked for group-level error risks widening the gaps it is meant to help close. That would sit uneasily with the equity and accountability aims of the National Education Policy (NEP) 2020 (Ministry of Education, 2020).

Indian educational thought adds a moral vocabulary of its own: the all-round development of the learner (*sarvāṅgīṇa vikāsa*), the central place of the teacher–student relationship (*guru–śiṣya paramparā*), and the responsibility of the person who teaches. These ideas do not define fairness in a computational sense. They do offer an ethical frame for deciding how prediction tools should be used, questioned and, when necessary, overruled.

Empirical work on group-level fairness in student-performance prediction is still sparse, and we are not aware of a framework that ties fairness metrics to the institutional and policy conditions of Indian schooling. This paper makes four contributions:

- an audit of group-level prediction-error disparity for three regression models on the UCI Student Performance data, used as a reproducible benchmark;
- a comparison of baseline and inverse-frequency-reweighted training on the same test set, with paired bootstrap intervals that show how much of any change could be sampling noise;
- a sensitivity experiment showing that disparity depends on which predictors are used; and
- the Bharatiya Trustworthy Educational AI (BTEAI) framework, which links computational fairness checks to Indian educational values, NEP 2020 and responsible-AI principles.

Section 2 reviews related work and Section 3 states the research gap, objectives and questions. Section 4 describes the data, models, metrics, mitigation method and uncertainty analysis. Section 5 reports the results. Section 6 presents the BTEAI framework, Section 7 discusses the findings, Section 8 sets out implications, Section 9 covers limitations and future work, and Section 10 concludes.

2. Related Work

2.1 AI in education

Machine learning has been applied to educational data for a long time. Cortez and Silva (2008) used decision trees, random forests, neural networks and support vector machines to predict the grades of Portuguese secondary-school students, and found that the earlier period grades were among the most informative predictors. Romero and Ventura (2020) survey educational data mining and learning analytics, from early-alert and dropout systems to personalised sequencing of learning material, and Zhai et al. (2021) review AI research in education over 2010–2020.

On structured tabular data, tree ensembles such as Random Forest and gradient-boosted trees (Chen & Guestrin, 2016) are usually competitive, which the baseline comparison in this paper again shows. Sequence models such as Deep Knowledge Tracing (Piech et al., 2015) address a different setting, in which a student's interactions with a learning system are modelled over time. Across this literature the dominant yardstick is aggregate accuracy. Zawacki-Richter et al. (2019), reviewing 146 articles on AI in higher education, pointed to limited critical reflection on the ethical risks of these systems.

2.2 Algorithmic fairness and bias

Algorithmic fairness is now an established research area. Mehrabi et al. (2021) survey sources of bias and a range of fairness definitions, and Barocas et al. (2019) distinguish individual fairness, which asks that similar people be treated similarly, from group fairness, which asks for comparable outcomes across groups. Group fairness is easier to measure and is the approach used here. Bias can enter at several stages: through historical patterns in the training data, through unrepresentative sampling, through features that are measured better for some groups than others, and through fitting one model to a mixed population (Mehrabi et al., 2021). Barocas and Selbst (2016) analysed how data mining can produce disparate impact under United States anti-discrimination law even when nobody intends to discriminate.

Mitigation methods are usually grouped into pre-processing, in-processing and post-processing. Pre-processing changes the training data or the weight of each example. Kamiran and Calders (2012) proposed reweighing for classification, Feldman et al. (2015) repaired data to remove disparate impact, and Calmon et al. (2017) optimised the data transformation itself. The Fairlearn toolkit (Bird et al., 2020) packages several assessment and mitigation methods. The method used in this paper is close to reweighing but simpler, because the weights depend only on group membership and not on the outcome.

Whether fairness costs accuracy is an empirical question. Menon and Williamson (2018) characterise the accuracy cost of fairness constraints in binary classification and show that it depends on the problem. Mitchell et al. (2021) catalogue the choices and assumptions behind fairness definitions and stress that several definitions cannot all hold at once, so the choice has to be made explicitly. Blanzeisky and Cunningham (2022) show that a model with reasonable overall accuracy can still be biased against an under-represented group, and that multi-objective optimisation can reduce that bias while keeping accuracy high.

2.3 Fairness in educational AI

Baker and Hawn (2022) reviewed the empirical evidence on algorithmic bias in education, covering which groups are affected and at which stages of development and deployment bias enters. Kizilcec and Lee (2022) give an introduction to algorithmic fairness in education that contrasts statistical, similarity-based and causal notions of fairness and offers recommendations to policy makers and developers. Memarian and Doleck (2023) reviewed how fairness, accountability, transparency and ethics are treated in research on AI in higher education. The UCI Student Performance data are public, well documented and include sex and home address (Cortez & Silva, 2008), which makes them convenient for a reproducible demonstration of a group-level audit, though not for conclusions about any other country.

2.4 The Indian and Bharatiya context

NEP 2020 calls for technology-enabled reform aimed at access, equity, quality and accountability (Ministry of Education, 2020). Platforms such as DIKSHA, SWAYAM and NISHTHA have widened access to content and teacher training. AI is now entering the classroom directly: from the 2026–27 session the Central Board of Secondary Education has introduced a curriculum on computational thinking and artificial intelligence for Classes III to VIII (Central Board of Secondary Education

[CBSE], 2026). As AI becomes part of school life, the question of whom it serves well cannot be put off.

The Indian tradition has long treated knowledge as a means of liberation. A well-known verse in the *Viṣṇu Purāṇa* says that true knowledge is the kind that frees (*sā vidyā yā vimuktaye*; Viṣṇu Purāṇa 1.19.41). Together with the *gurukula* ideal of teaching that is tied to the learner's circumstances, this points to the holistic development of the learner, the primacy of the teacher–student bond, and the ethical accountability of the one who holds knowledge. These values give no formal definition of algorithmic fairness, but they are consistent with the aims of trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019; UNESCO, 2021) and can guide how ML tools are used in Indian schools.

3. Research Gap and Objectives

The literature reviewed above points to a common gap. Educational ML models are mostly judged by aggregate metrics, so group-level differences in error go unexamined. Where fairness is discussed, empirical mitigation with explicit disparity measures is less common, and there is no established framework that connects computational fairness checks with the linguistic, socioeconomic and policy conditions of Indian education. The objectives in Table 1 respond to this gap.

Table 1. Research objectives.

Objective	Description
O1	Evaluate selected ML regression models for student academic-performance prediction using standard performance metrics (MAE, RMSE, R^2).
O2	Measure group-level differences in predictive performance across student groups defined by sex and home address.
O3	Quantify group-level prediction-error disparity using PD, SD_{MAE}, and CV.
O4	Examine whether inverse-frequency sample reweighting reduces group-level disparity while maintaining acceptable overall performance.
O5	Develop a Bharatiya-contextual framework for trustworthy, fairness-aware educational AI appropriate to India's diverse educational environment.

The study is organised around four research questions:

- **RQ1.** How accurately can the selected models predict student performance on the UCI Student Performance data?
- **RQ2.** Do prediction errors differ systematically between student groups defined by sex and by home address?
- **RQ3.** Can inverse-frequency reweighting reduce group-level disparity while keeping overall accuracy acceptable, and how certain is that conclusion on a small test set?
- **RQ4.** Which principles should guide the responsible use of fairness-aware educational AI in India's diverse educational setting?

4. Materials and Methods

4.1 Dataset

The study uses the UCI Student Performance dataset (Cortez, 2008; dataset ID 320; https://doi.org/10.24432/C5TG7T), released under a CC BY 4.0 licence. It describes achievement in secondary education at two Portuguese schools, Gabriel Pereira and Mousinho da Silveira, and was gathered from school reports and questionnaires in 2005–2006. Only the mathematics file is used. It

holds 395 students and 33 attributes (4 nominal, 13 binary and 16 numerical), with no missing values or duplicate rows; both were checked in the copy used here. The target is G3, the final-period grade on a 0–20 scale.

Caveat. The data come from Portuguese schools in 2005–2006. They do not represent Indian students, schools or conditions. They are used only as a reproducible setting in which to demonstrate a group-level audit, and nothing measured on them is generalised to India.

Table 2. Characteristics of the dataset, verified on the working copy.

Characteristic	Verified value
Dataset name	UCI Student Performance (ID: 320)
Creator	Paulo Cortez, University of Minho, Portugal
Original source	Two secondary schools, Portugal (GP & MS)
Collection period	2005–2006
File used	student-mat.csv (Mathematics)
Instances (verified on load)	395
Total attributes	33 (4 categorical, 13 binary, 16 numerical)
Missing values (verified)	0
Duplicate rows (verified)	0
Target variable	G3 — Final period grade (0–20, continuous)
Task type	Regression
Group attribute 1	sex: binary (F = Female, M = Male)
Group attribute 2	address: binary (U = Urban, R = Rural)
DOI	10.24432/C5TG7T
Licence	CC BY 4.0
Represents Indian students?	No — Portuguese educational context only

4.2 Preprocessing and data split

The same steps were applied to every model. Missing values and duplicates were checked first. Binary variables were label-encoded, and the four nominal variables (Mjob, Fjob, reason and guardian) were one-hot encoded with the first level dropped to avoid collinearity. G3 was separated as the target and all other attributes formed the feature matrix, so the attributes sex and address remained among the predictors. The original values of sex and address were also kept aside for the fairness evaluation.

G1 and G2, the first- and second-period grades, were kept as predictors in the main analysis. They come before G3 in time and would normally be available to a school at prediction time, and the dataset documentation notes their strong correlation with G3. To see whether the fairness findings depend on them, a separate experiment without G1 and G2 is reported in Section 5.7.

The data were split 80/20 with a fixed random seed of 42, giving 316 training and 79 test students. The test set contained 38 female and 41 male students, and 60 urban and 19 rural students. For Linear Regression only, each feature was standardised using training-set statistics,

$$z = \frac{x_{\text{test}} - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{1}$$

and the same transformation was applied to the test set. The tree-based models received unscaled features.

4.3 Models

Three regressors span the range from a transparent linear baseline to non-linear ensembles (Table 3). They were implemented with scikit-learn (Pedregosa et al., 2011) and XGBoost (Chen & Guestrin, 2016). Beyond the held-out evaluation, five-fold cross-validation was run on the full 395 students (shuffled, seed 42), and the reported values are the mean and standard deviation of the five fold-level scores.

Table 3. Model configurations.

Model	Key hyperparameters	Scaling	Random seed
Linear Regression	Default (scikit-learn)	StandardScaler	N/A
Random Forest Regressor	n_estimators=100, max_depth=None	None	42
XGBoost Regressor	n_estimators=100, learning_rate=0.1, max_depth=6	None	42

4.4 Performance metrics

For n students with observed grade y_i and predicted grade $\hat{y}_i$, the mean absolute error, root mean squared error and coefficient of determination are

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

MAE is the average size of the error in grade points and is less sensitive to a few large errors, RMSE penalises large errors more heavily, and R^2 is the share of variance in G3 that the model explains.

4.5 Fairness metrics

Let there be G groups, and let n_g be the number of test students in group g. The group-level MAE is the MAE computed only over that group,

$$\mathrm{MAE}_g = \frac{1}{n_g}\sum_{i \in g} |y_i - \hat{y}_i| \tag{5}$$

Performance disparity (PD) is the gap between the worst-served and best-served group,

$$\mathrm{PD} = \max_g \mathrm{MAE}_g - \min_g \mathrm{MAE}_g \tag{6}$$

The spread of group MAEs is summarised by their standard deviation, and this is normalised by the mean group MAE to give a coefficient of variation that can be compared across models with different error levels:

$$\mathrm{SD}_{\mathrm{MAE}} = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\mathrm{MAE}_g - \mu_{\mathrm{MAE}}\right)^2}, \quad \mu_{\mathrm{MAE}} = \frac{1}{G}\sum_{g=1}^{G}\mathrm{MAE}_g \tag{7}$$

$$\mathrm{CV} = \frac{\mathrm{SD}_{\mathrm{MAE}}}{\mu_{\mathrm{MAE}}} \tag{8}$$

Here the population form of the standard deviation is used (division by G). With two groups, as for sex and for address, this makes $\mathrm{SD}_{\mathrm{MAE}}$ exactly half of PD, and CV equals PD divided by twice the mean group MAE. The three metrics are therefore closely related, and all are reported so that readers can see the same disparity on absolute and relative scales. A smaller value means more even accuracy across groups.

Qualification. These metrics describe how unevenly prediction error is spread across groups. They do not capture every dimension of educational justice, and a low PD does not by itself make a system fair in a wider social sense.

4.6 Bias-mitigation method

Inverse-frequency reweighting was used as a pre-processing method. Training students were assigned to one of $K = 4$ joint groups (F-U, F-R, M-U, M-R; sex × address), so that both attributes are addressed together. If group k holds a share $p_k = n_k/n$ of the n training students, each of its members receives a weight proportional to $1/p_k$, scaled so that the weights sum to n. Writing $k(i)$ for the group of student i,

$$w_i = \frac{1}{p_{k(i)}} \cdot \frac{n}{\sum_{j=1}^{n} \frac{1}{p_{k(j)}}} = \frac{n}{K n_{k(i)}} \tag{9}$$

The weighted training objective is

$$\min_f \sum_{i=1}^{n} w_i \ell\big(y_i, f(x_i)\big) \tag{10}$$

where ℓ is the squared error for Linear Regression, which becomes weighted least squares, and the model's own loss for Random Forest and XGBoost. Weights were passed through each model's *sample_weight* argument. Table 4 gives the group sizes and resulting weights. Rural students, who are the smaller groups, receive about four times the weight of the largest group. All other settings and the test set were the same as for the baseline models.

Table 4. Joint groups in the training set and their inverse-frequency weights.

Group	Training students	Share p_k	Weight w_i
F-U (female, urban)	136	0.430	0.581
F-R (female, rural)	34	0.108	2.324
M-U (male, urban)	111	0.351	0.712
M-R (male, rural)	35	0.111	2.257

Note. $n = 316$, $K = 4$; the weights sum to 316.

4.7 Uncertainty of the disparity estimates

With 79 test students, and only 19 in the rural group, a single value of PD can move a lot from one sample to the next. To gauge this, a paired bootstrap was run on the saved test predictions. In each of $B = 10{,}000$ resamples (seed 42), 79 test students were drawn with replacement, and PD was recomputed for the baseline and the reweighted model on the same resampled students. The change in disparity for each resample is

$$\Delta\mathrm{PD} = \mathrm{PD}_{\mathrm{rew}} - \mathrm{PD}_{\mathrm{base}} \tag{11}$$

and the 2.5th and 97.5th percentiles of its distribution form a 95% percentile interval. The share of resamples with $\Delta\mathrm{PD} < 0$ is also reported. This procedure captures variation from which students happen to be in the test set. It does not capture variation from retraining the models on different training samples or seeds, so the intervals are, if anything, too narrow. Figure 1 summarises the full pipeline.

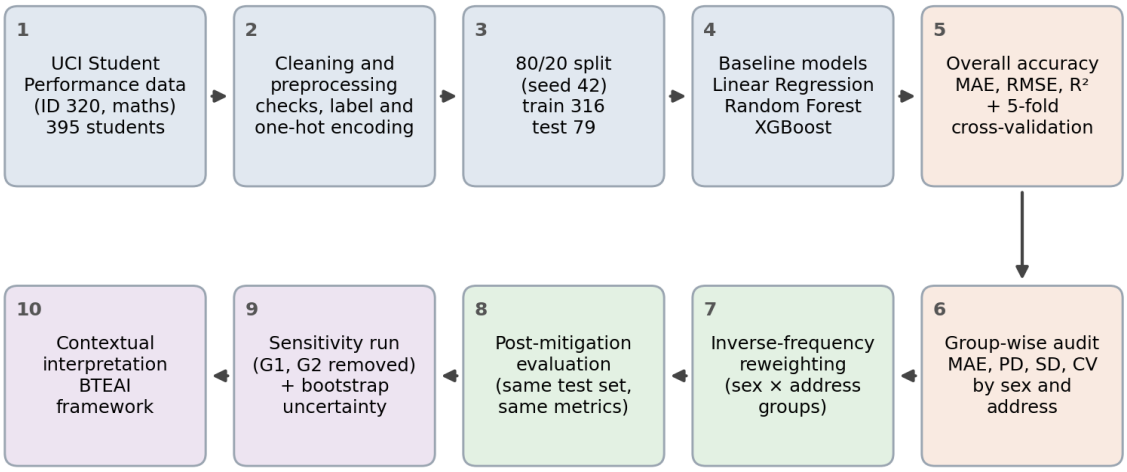


Figure 1. Research methodology workflow.

5. Experimental Results

Every number in this section was computed from the dataset with the pipeline in Section 4 and a fixed seed of 42. None is simulated or illustrative. The analysis was run in a Python notebook, and the saved test predictions were used for the bootstrap analysis.

5.1 Distribution of the target variable

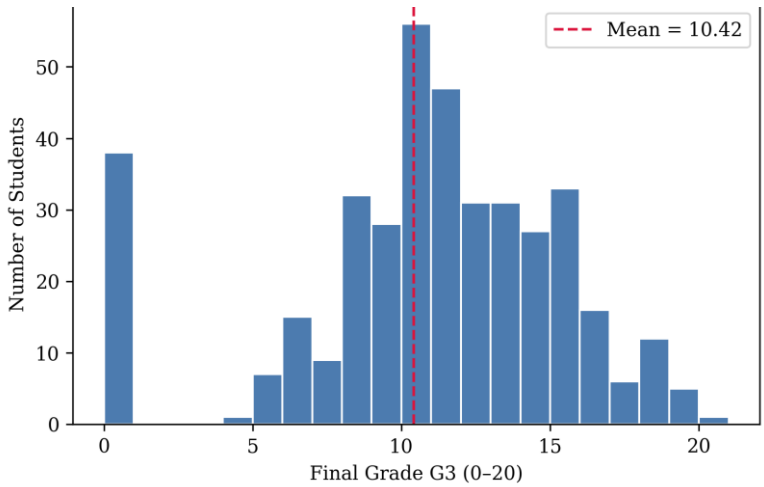


Figure 2. Distribution of the final mathematics grade (G3), N = 395. The dashed line marks the mean (10.42).

Most G3 values lie between 8 and 14. There is also a separate cluster at zero, about 38 students, who either did not take the final assessment or scored zero in it. This zero cluster is described in the original dataset documentation (Cortez & Silva, 2008) and is a hard case for any regression model, since it lies far from the rest of the distribution.

5.2 Five-fold cross-validation

Table 5. Five-fold cross-validation performance with all predictors (mean ± SD over folds).

Model	MAE	RMSE	R^2
Linear Regression	1.378 ± 0.165	2.057 ± 0.176	0.792 ± 0.042

Model	MAE	RMSE	R^2
Random Forest	1.020 ± 0.135	1.610 ± 0.337	0.868 ± 0.058
XGBoost	1.018 ± 0.134	1.634 ± 0.307	0.864 ± 0.056

Random Forest and XGBoost were about equally accurate in cross-validation and clearly better than Linear Regression (R^2 of about 0.86–0.87 against 0.79). This fits their ability to model non-linear interactions among the mixed categorical and numeric predictors. The fold-to-fold spread for the ensembles is not small, especially for RMSE, which may reflect folds that contain several zero-grade students.

5.3 Held-out test set: overall performance

Table 6. Overall performance of the baseline models on the held-out test set (n = 79).

Model	MAE	RMSE	R^2
Linear Regression	1.647	2.378	0.724
Random Forest	1.165	1.923	0.820
XGBoost	1.240	2.137	0.777

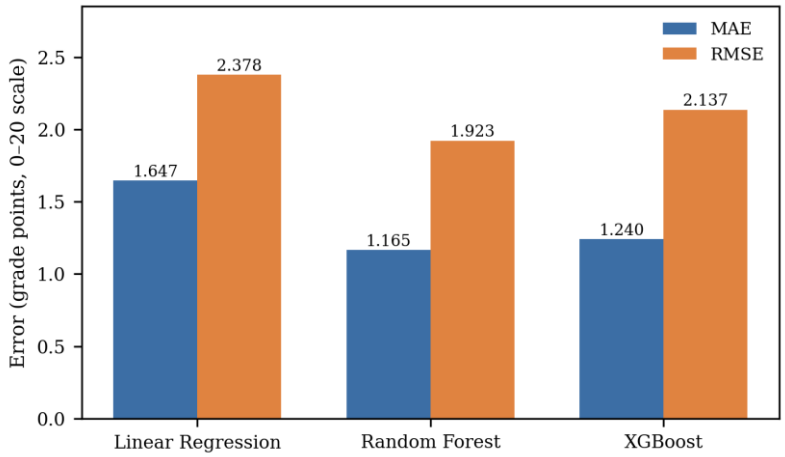


Figure 3. Overall MAE and RMSE of the three baseline models on the held-out test set.

Random Forest had the lowest MAE (1.165) and the highest R^2 (0.820), so by every aggregate metric it is the strongest baseline. The next two sections show that this ranking does not carry over to group-level fairness.

5.4 Group-wise prediction error of the baseline models

Table 7. Group-wise prediction error of the baseline models on the held-out test set.

Model	Attribute	Group	n	Group MAE	Group RMSE
Linear Regression	sex	Female (F)	38	1.878	2.741
Linear Regression	sex	Male (M)	41	1.432	1.984
Linear Regression	address	Urban (U)	60	1.545	2.296
Linear Regression	address	Rural (R)	19	1.966	2.622
Random Forest	sex	Female (F)	38	1.408	2.341
Random Forest	sex	Male (M)	41	0.940	1.430
Random Forest	address	Urban (U)	60	1.179	1.987
Random Forest	address	Rural (R)	19	1.122	1.707

Model	Attribute	Group	*n*	Group MAE	Group RMSE
XGBoost	sex	Female (F)	38	1.422	2.424
XGBoost	sex	Male (M)	41	1.072	1.831
XGBoost	address	Urban (U)	60	1.293	2.321
XGBoost	address	Rural (R)	19	1.074	1.404

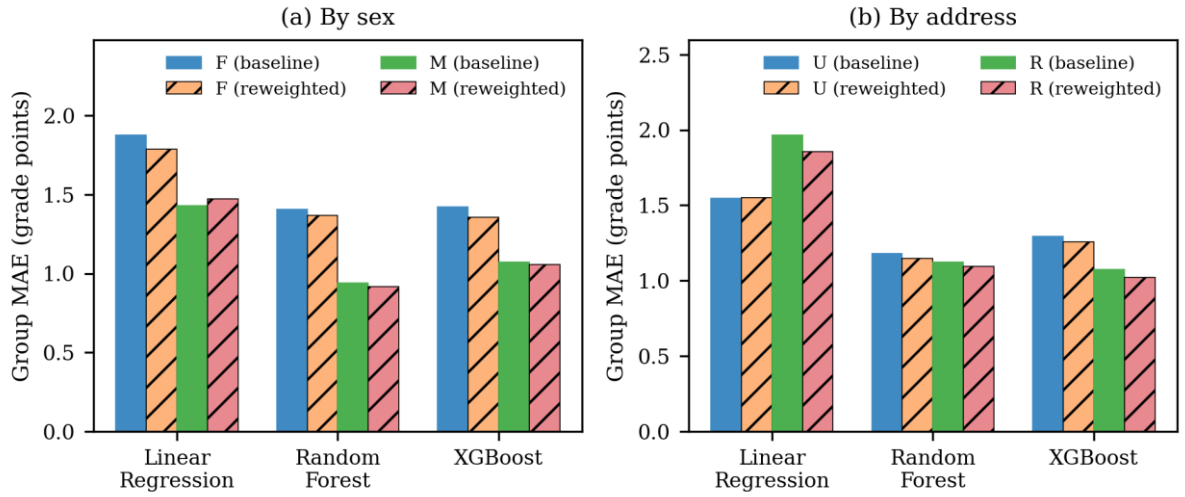


Figure 4. Group-wise MAE for (a) sex and (b) address groups, for the baseline models and their reweighted counterparts.

One pattern holds for all three baseline models: female students had a higher group MAE than male students (Random Forest 1.408 against 0.940, XGBoost 1.422 against 1.072, Linear Regression 1.878 against 1.432). The address pattern was less consistent. Linear Regression had higher error for rural students (1.966 against 1.545 for urban), while Random Forest and XGBoost had similar or lower error for rural students (for example XGBoost, 1.074 against 1.293). The presence and direction of a group gap therefore depend on the model and cannot be read off from the aggregate metric.

The direction of the errors adds a further detail. In the saved test predictions, the mean signed error (predicted minus observed grade) was negative for female students under all three models (−0.41 for Linear Regression, −0.19 for Random Forest, −0.16 for XGBoost), meaning their grades were on average under-predicted. For male students it was −0.22, +0.02 and +0.13. For rural students the tree ensembles over-predicted on average (+0.20 for Random Forest, +0.36 for XGBoost), while they under-predicted urban students slightly (−0.17 and −0.12). If predictions were used to allocate support, over-prediction for a group would mean fewer students of that group being noticed, so the sign of the error matters as well as its size. With 19 rural students these averages are themselves uncertain.

5.5 Fairness metrics of the baseline models

Table 8. Fairness metrics of the baseline models.

Model	Group attribute	Overall MAE	PD	SD_{MAE}	CV
Linear Regression	sex	1.647	0.446	0.223	0.135
Linear Regression	address	1.647	0.421	0.211	0.120
Random Forest	sex	1.165	0.468	0.234	0.199
Random Forest	address	1.165	0.057	0.029	0.025
XGBoost	sex	1.240	0.350	0.175	0.140
XGBoost	address	1.240	0.219	0.110	0.093

Main finding. Random Forest, the most accurate model overall, had the largest sex-based disparity of the three (PD = 0.468, CV = 0.199) and, at the same time, the smallest address-based disparity (PD = 0.057, CV = 0.025). XGBoost had the smallest sex-based PD (0.350) without being the most accurate. Selecting a model on aggregate accuracy therefore does not guarantee that it treats any given group best, and which group is best treated depends on the attribute examined.

5.6 Effect of inverse-frequency reweighting

Table 9 compares each baseline model with its reweighted version on the same test set. All changes are point estimates; their uncertainty is examined in Section 5.8. Percentage change is (reweighted − baseline)/baseline × 100.

Table 9a. Before and after reweighting: Linear Regression.

Metric	Baseline	Reweighted	Change	% change
Overall MAE	1.647	1.625	−0.022	−1.3%
Overall RMSE	2.378	2.420	+0.042	+1.8%
Overall R^2	0.724	0.714	−0.010	−1.3%
PD (sex)	0.446	0.318	−0.128	−28.7%
PD (address)	0.421	0.307	−0.114	−27.2%
SD_{MAE} (sex)	0.223	0.159	−0.064	−28.7%
SD_{MAE} (address)	0.211	0.153	−0.057	−27.2%
CV (sex)	0.135	0.098	−0.037	−27.6%
CV (address)	0.120	0.090	−0.030	−25.0%

Table 9b. Before and after reweighting: Random Forest.

Metric	Baseline	Reweighted	Change	% change
Overall MAE	1.165	1.134	−0.032	−2.7%
Overall RMSE	1.923	1.914	−0.009	−0.5%
Overall R^2	0.820	0.821	+0.002	+0.2%
PD (sex)	0.468	0.451	−0.016	−3.5%
PD (address)	0.057	0.051	−0.006	−10.2%
SD_{MAE} (sex)	0.234	0.226	−0.008	−3.5%
SD_{MAE} (address)	0.029	0.026	−0.003	−10.2%
CV (sex)	0.199	0.198	−0.002	−0.8%
CV (address)	0.025	0.023	−0.002	−7.7%

Table 9c. Before and after reweighting: XGBoost.

Metric	Baseline	Reweighted	Change	% change
Overall MAE	1.240	1.201	−0.040	−3.2%
Overall RMSE	2.137	2.049	−0.088	−4.1%
Overall R^2	0.777	0.795	+0.018	+2.3%
PD (sex)	0.350	0.300	−0.050	−14.3%
PD (address)	0.219	0.234	+0.015	+6.7%
SD_{MAE} (sex)	0.175	0.150	−0.025	−14.3%
SD_{MAE} (address)	0.110	0.117	+0.007	+6.7%
CV (sex)	0.140	0.124	−0.016	−11.4%
CV (address)	0.093	0.103	+0.010	+10.7%

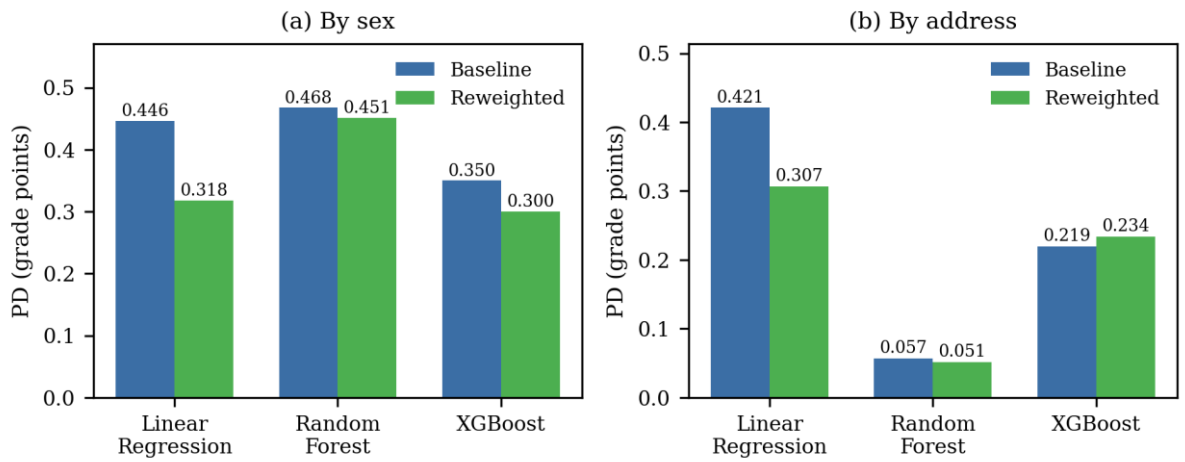


Figure 5. Performance disparity (PD) before and after inverse-frequency reweighting, by (a) sex and (b) address.

Reweighting lowered sex-based PD for all three models (Linear Regression −28.7%, Random Forest −3.5%, XGBoost −14.3%). It lowered address-based PD for Linear Regression (−27.2%) and Random Forest (−10.2%) and raised it slightly for XGBoost (+6.7%). Overall accuracy was not harmed in a consistent way: MAE fell for Random Forest (−2.7%) and XGBoost (−3.2%) and, marginally, for Linear Regression (−1.3%), although the Linear Regression RMSE rose by 1.8%. On this dataset, then, the reduction in disparity did not require a visible accuracy sacrifice, which is in line with the point made by Blanzeisky and Cunningham (2022) that fairness and accuracy are not opposed in every setting. The Linear Regression result also has a simple reading. It is the least flexible model, so a change in the weights moves its fitted line further than it moves the fit of a forest or a boosted ensemble, which may explain why its change in PD is the largest.

5.7 Sensitivity experiment: excluding G1 and G2

Table 10. Sensitivity experiment: baseline models trained without G1 and G2.

Model	MAE	RMSE	R^2	PD (sex)	PD (address)	SD_{MAE} (sex)	CV (sex)
Linear Regression	3.395	4.196	0.141	0.162	0.442	0.081	0.024
Random Forest	3.096	3.900	0.258	0.072	0.053	0.036	0.012
XGBoost	3.174	3.979	0.228	0.399	0.055	0.199	0.063

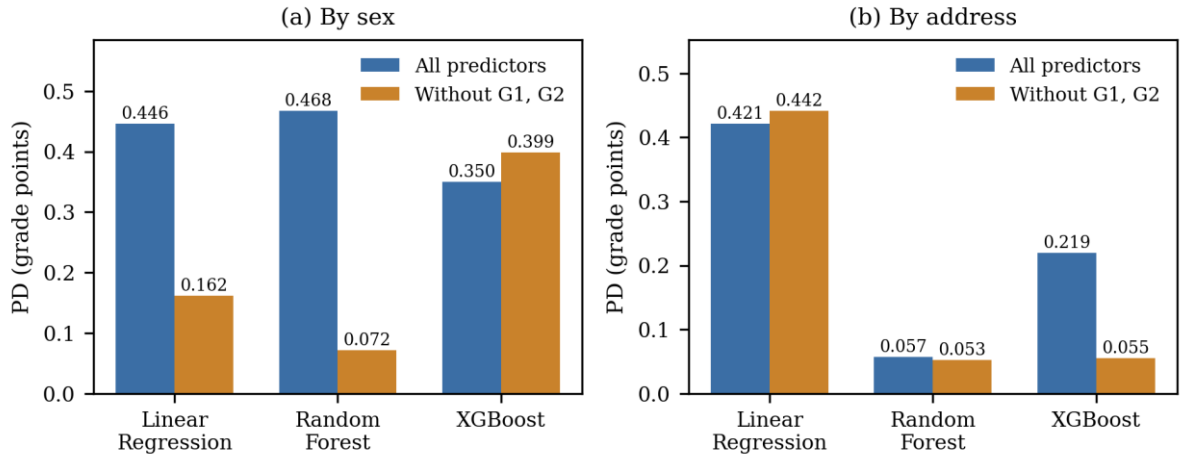


Figure 6. Baseline PD with all predictors and without G1 and G2, by (a) sex and (b) address.

Dropping the two prior grades cut accuracy sharply for every model (R^2 fell to between 0.14 and 0.26), which agrees with the documented strength of their link to G3. The disparity pattern also changed. Random Forest stayed low on both attributes (PD of 0.072 for sex and 0.053 for address). Linear Regression's sex-based PD fell from 0.446 to 0.162, but its address-based PD stayed high (0.421 to 0.442). XGBoost's sex-based PD went the other way, from 0.350 to 0.399. This does not show why; the models and features changed together, and we did not test the mechanism. What it does show is that a fairness result measured under one feature set need not hold under another.

The choice of metric matters too. The relative measure CV (sex) fell for all three models once G1 and G2 were removed (for XGBoost from 0.140 to 0.063), because the average error grew so much, even though XGBoost's absolute gap widened. A disparity that looks larger in grade points can look smaller relative to a much larger overall error, and the two readings answer different questions.

5.8 Uncertainty of the disparity changes

Table 11. Paired bootstrap results for PD (10,000 resamples of the test set): point estimate and 95% percentile interval.

Model	Attribute	PD, baseline	PD, reweighted	ΔPD	Share ΔPD < 0
Linear Regression	sex	0.446 [0.022, 1.238]	0.318 [0.018, 1.148]	−0.128 [-0.268, 0.215]	0.76
Linear Regression	address	0.421 [0.019, 1.391]	0.307 [0.017, 1.368]	−0.114 [-0.300, 0.278]	0.64
Random Forest	sex	0.468 [0.028, 1.181]	0.451 [0.025, 1.166]	−0.016 [-0.122, 0.096]	0.60
Random Forest	address	0.057 [0.012, 0.823]	0.051 [0.011, 0.869]	−0.006 [-0.108, 0.143]	0.41
XGBoost	sex	0.350 [0.020, 1.157]	0.300 [0.015, 1.069]	−0.050 [-0.247, 0.157]	0.65
XGBoost	address	0.219 [0.012, 0.886]	0.234 [0.012, 0.844]	+0.015 [-0.244, 0.219]	0.54

Note. ΔPD = PD(reweighted) − PD(baseline), computed on the same resampled students.

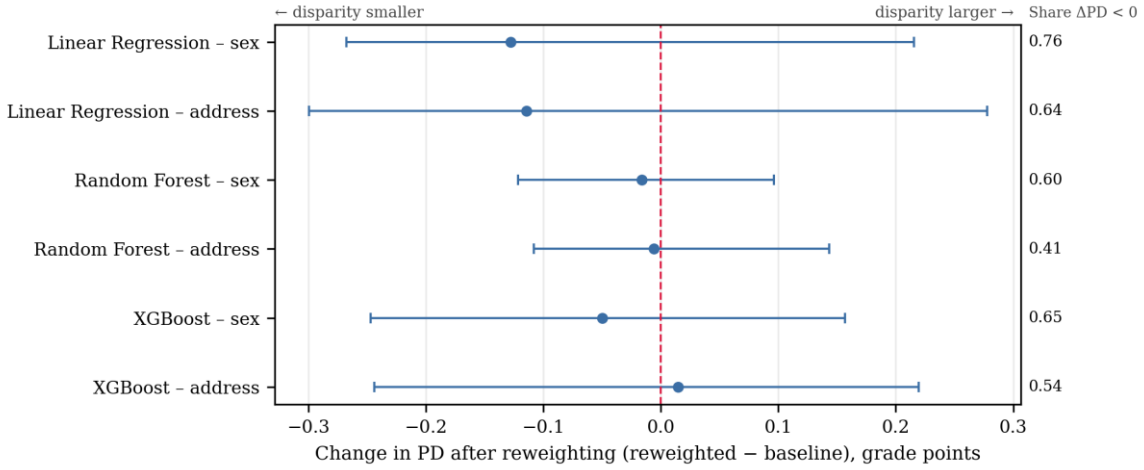


Figure 7. Change in PD after reweighting with 95% bootstrap intervals. Points to the left of zero indicate a smaller disparity. The column on the right gives the share of resamples with ΔPD < 0.

The intervals are wide. Every 95% interval for ΔPD contains zero, and the share of resamples in which reweighting reduced PD ranged from 0.60 to 0.76 for sex and from 0.41 to 0.64 for address. The reductions in Table 9 are therefore consistent with a real but modest improvement and also with no change at all; a 79-student test set cannot separate the two. The intervals for PD itself have lower limits close to zero. This is partly an effect of the statistic, since a range computed from two noisy group means is never negative and is biased upward, so an interval for PD should be read together with the interval for ΔPD and not on its own.

6. The Bharatiya Trustworthy Educational AI (BTEAI) Framework

This section proposes the Bharatiya Trustworthy Educational AI (BTEAI) framework, a conceptual guide for designing, evaluating and deploying AI-based educational systems in India. It draws on general trustworthy-AI principles (High-Level Expert Group on Artificial Intelligence, 2019; UNESCO, 2021), on the fairness literature reviewed in Section 2, and on Bharatiya educational thought, NEP 2020 priorities and India's socioeconomic and linguistic diversity.

The framework does not claim that classical Indian texts contained a formal definition of algorithmic fairness. Its claim is narrower: that long-standing values centred on the learner's dignity, circumstances and all-round development give an ethical setting in which computational fairness tools, such as those used in Section 5, can be applied, checked and overseen. It has eight components arranged as a cycle (Figure 8, Table 12). The cycle is meant to be repeated whenever the data, the features or the deployment context change.

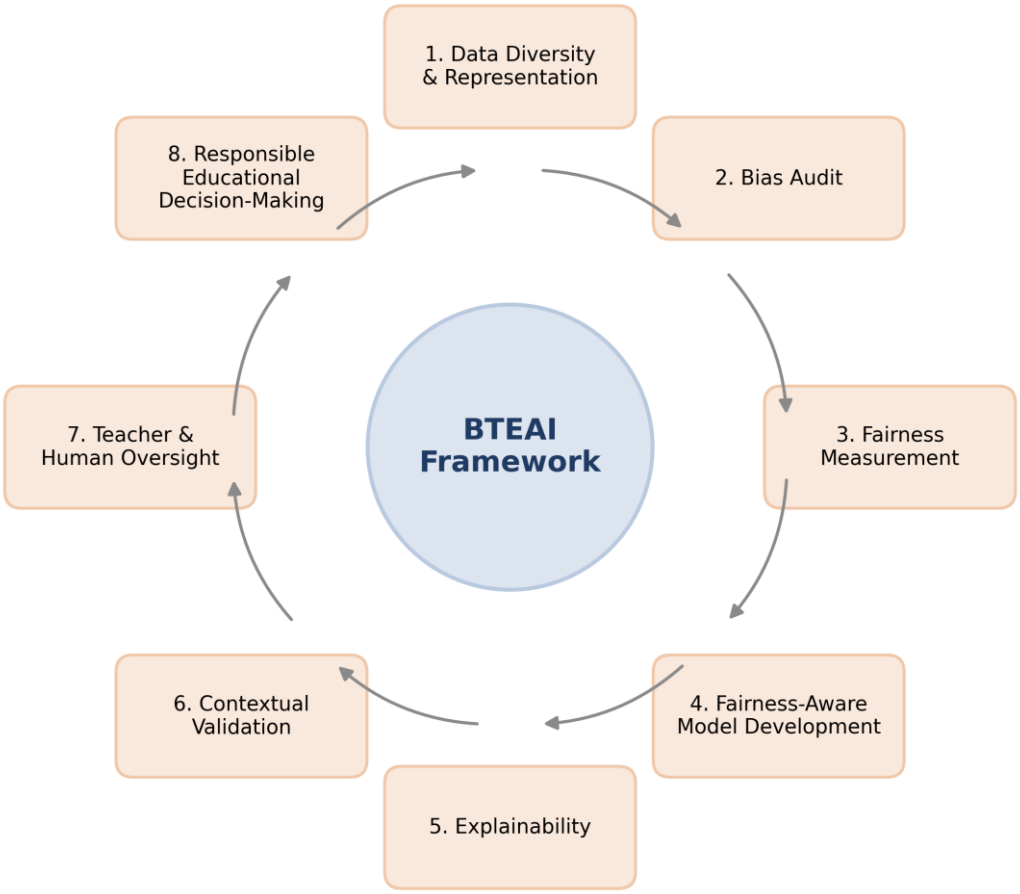


Figure 8. The Bharatiya Trustworthy Educational AI (BTEAI) framework as an eight-component cycle.

Component 1: Data diversity and representation

A model meant for Indian use has to be trained on data that reflect the diversity of Indian learners. Data that under-represent rural students, students from low-income households, linguistic minorities or students with disabilities will disadvantage them in any model built on it. This is the same concern that the address-based results in Sections 5.4 and 5.5 raise on a small scale. The component calls for documented data provenance, an explicit check of who is represented, and new data collection in under-represented regions before a fairness-sensitive model is deployed.

Component 2: Bias audit

A structured audit should come before deployment. Baker and Hawn (2022) observe that descriptive research can turn an algorithm's bias from unknown to known, after which the problem, its location and its effect are clear enough to correct. The audit in Section 5 does this for the sex and address groups. A full audit would also ask whether the training data record earlier unequal treatment, whether feature engineering encodes group membership indirectly, and whether tuning choices affect groups differently.

Component 3: Fairness measurement

Fairness should be measured and reported, not just asserted. The framework asks for group-wise MAE, PD, SD_{MAE} and CV, in the format of Tables 7 to 11, for every group attribute relevant to the setting. In India these go beyond sex and address to socioeconomic status, type of school (government, private, Kendriya Vidyalaya, Navodaya), state or region, home language and social category, where it is lawful and ethical to record them. No single metric covers every aspect of fairness, so metrics have to be chosen for the decision the model supports.

Component 4: Fairness-aware model development

Where a disparity is found, mitigation should be considered. Section 5.6 shows that reweighting is cheap and simple, and that on this dataset it lowered the point estimates of disparity with no visible accuracy cost. Section 5.8 also shows that the effect cannot be confirmed on a sample this small, so the framework treats reweighting as a first step whose effect is to be verified on adequate data. In higher-stakes uses, in-processing methods such as constrained optimisation or fairness-regularised losses may be more suitable, and the choice should reflect the size of the disparity and the cost of any trade-off.

Component 5: Explainability

Teachers and administrators who rely on a prediction need to understand it. Post-hoc explanation methods such as SHAP for tree models and LIME for individual predictions can show why a particular student received a particular prediction. This matters especially given Section 5.7, where a change in which features the model could use also changed its fairness properties. It matters most where teachers have little ML training but carry the professional responsibility for decisions. The Indian ideal of *viveka*, discriminating judgement, applies directly: a recommendation should be understood before it is acted on.

Component 6: Contextual validation

A model validated on one dataset, including the UCI benchmark used here, should not be moved to another setting without re-validation. A model tested on urban, English-medium schools in one state may behave differently in rural schools that teach in a regional language, use different assessment formats and serve students with different socioeconomic backgrounds. Both accuracy and fairness metrics should be re-evaluated on local data before operational use. Where local data are too few, participatory data collection with teachers, students and community representatives should be preferred.

Component 7: Teacher and human oversight

No AI prediction or recommendation should replace a teacher's professional judgement. Teachers know about individual students' circumstances, family situations and progress in ways that a tabular model cannot capture. AI output should serve as one input to a teacher-led process and not as a final decision. This is in keeping with the *guru–śiṣya parampar*ā, in which the teacher is personally responsible for the student's development.

Component 8: Responsible educational decision-making

Institutions that use AI need a governance layer that makes decisions accountable, transparent and reversible. They should record which model and version is used, its known limits and fairness properties (as in Tables 7 to 11), where the training data came from, how human review of AI output works, and how students or families can complain if they think they have been harmed. NEP 2020's emphasis on accountability supports this requirement.

Table 12. The eight BTEAI components and the evidence in this study that motivates each.

Component	Core requirement	Link to this study
1. Data diversity and representation	Documented provenance; representation audit; local data collection	Small rural group (19 test students) and unstable address results (Sections 5.4, 5.8)
2. Bias audit	Structured audit before deployment	Group gaps invisible in aggregate metrics (Sections 5.3–5.5)
3. Fairness measurement	Report group MAE, PD, SD_{MAE}, CV for relevant attributes	Tables 7–11
4. Fairness-aware model development	Apply and verify mitigation	Reweighting lowered point estimates but not conclusively (Sections 5.6, 5.8)
5. Explainability	Explanations for teachers and administrators	Fairness changed with the feature set (Section 5.7)
6. Contextual validation	Re-validate on local data	Data are Portuguese; no transfer to India (Section 4.1)
7. Teacher and human oversight	Human override; AI as one input	Signed errors differ by group (Section 5.4)
8. Responsible decision-making	Documentation, review, grievance routes	Need to re-audit when features or data change (Section 5.7)

7. Discussion

RQ1: Overall accuracy

The three models were chosen to cover a range of complexity. Linear Regression was the transparent baseline and the least accurate (test $R^2 = 0.724$). Random Forest and XGBoost did better (0.820 and 0.777), in line with their ability to capture non-linear interactions. Cross-validation gave the same ranking (R^2 of about 0.87 for Random Forest against 0.79 for Linear Regression). Without G1 and G2 all models were much less accurate, and Random Forest remained the best of the three (Table 10).

RQ2: Group-level differences

Group MAE was higher for female than for male students under all three baseline models, while the address gap changed direction between Linear Regression and the ensembles. These gaps should not be read as effects of sex or residence. They reflect how predictors such as study time, family support, internet access and parental education are distributed within each group, and what the model has learned from them. What is clear is that the gaps are not small next to the overall error. Sex-based PD of 0.35 to 0.47 grade points is a large fraction of a model MAE of 1.2 to 1.6, and it does not follow the aggregate ranking.

RQ3: Effects of reweighting

Point estimates fell for sex-based PD in all three models and for address-based PD in two, generally without an accuracy penalty. That runs against the fairness–accuracy trade-off that is often expected (Menon & Williamson, 2018) and agrees with Blanzeisky and Cunningham (2022). The bootstrap analysis calls for restraint, however. Every interval for ΔPD includes zero, so the honest reading is that reweighting may help a little and that this test set cannot show it. The rise in XGBoost's address-based

PD is a reminder that one method optimised for two attributes at once will not improve every model–attribute pair, and that fairness should be reported attribute by attribute and not as one overall claim.

RQ4: Bharatiya contextual principles

The eight BTEAI components address aspects of Indian educational diversity that aggregate fairness metrics cannot reach: language, socioeconomic position, geography and institutional context. Computational tools of the kind used in Section 5 are necessary but not sufficient. Responsible educational AI in India also needs governance, teacher oversight, validation on local data and an ethical orientation drawn from Bharatiya educational values.

What was found here, group gaps that averages hide and fairness properties that shift with the feature set, is relevant to the Indian setting, where diversity by language, income and region is much greater than in the Portuguese sample. A model deployed across Indian schools without a fairness audit, and without a new audit whenever its features or data source change, could produce quite different outcomes for rural and urban students, for low-income and high-income families and for different language communities. That would work against the equity commitments of NEP 2020.

8. Implications

8.1 Theoretical implications

The study contributes to the literature on trustworthy AI, algorithmic fairness and educational ML in four ways. It applies PD, SD_{MAE} and CV, with the formal definitions above, to regression-based grade prediction and does so in a fully reproducible way. It evaluates inverse-frequency reweighting in that setting and, through the bootstrap, separates the observed change from what sampling alone could produce. It shows that ranking models by aggregate accuracy does not reliably predict their group-level fairness and that fairness properties can change with the feature set. And it proposes the BTEAI framework as a way to connect fairness measurement with Bharatiya educational values, NEP 2020 and responsible AI governance.

8.2 Practical implications

Developers and educational technology providers should make group-wise fairness audits a standard part of model evaluation, report them next to accuracy metrics, include uncertainty intervals, and repeat them whenever the feature set or data source changes. Institutions that adopt such systems should put fairness review, teacher training on the limits of AI and clear procedures for human override in place. Teachers should treat predictions as probabilistic estimates drawn from past patterns and not as statements about a student's ability, and should be supported in questioning them, particularly for students whose circumstances are rare in the training data.

8.3 Implications for Indian education

- **Rural–urban and state differences.** Even with recent gains, about one school in three still lacks a computer or internet access nationally, and the gap is far larger in some states (Department of School Education and Literacy, 2025a; Ministry of Education, n.d.). Models trained mainly on data from well-connected settings may perform less consistently for rural students, as the address results here illustrate on a small scale, so rural-group evaluation should be explicit before large-scale deployment.
- **Socioeconomic variation.** Features such as parental education, internet access and paid tutoring, all present in the UCI data, act as proxies for socioeconomic status and may amplify group disparities in a much more unequal setting than the Portuguese sample.

- **Language.** Systems used in multilingual classrooms should be validated separately for students taught in different languages, since assessment formats, grading norms and learning resources may differ.
- **Digital access.** Unreliable connectivity, data costs and limited device access mean that adaptive AI systems must be designed around the digital divide and must not deepen it.

9. Limitations and Future Research

9.1 Limitations

- **Origin of the data.** The data come from two Portuguese schools in 2005–2006 and do not represent Indian students, schools or socioeconomic conditions. No finding is generalised to India.
- **Narrow context.** The data cover two schools, one subject and one academic year, which limits representativeness even within Portugal.
- **Small subgroups and one split.** Results rest on one train–test split of 79 test students, with only 19 rural students. The bootstrap intervals in Section 5.8 are wide, and they reflect test-set sampling only; variation from retraining, other seeds or other splits is not included. Hyperparameters were not tuned.
- **Group attributes were also predictors.** Sex and address were left in the feature set, so the models could use them directly. A model that excludes them, or one that uses them only to set weights, could behave differently and should be compared in later work.
- **Limited group attributes.** Only sex and address were examined. They cannot capture intersectional disadvantage, caste, disability, religion, home language or socioeconomic status, all of which matter in India.
- **Scope of the metrics.** PD, SD_{MAE} and CV measure disparity in prediction error. They do not measure fairness of access or allocation, or wider distributive and procedural justice.
- **Conceptual framework.** BTEAI is proposed on conceptual grounds and has not been validated with Indian data, institutions or practitioners, and the study includes no input from students, teachers or families.

9.2 Future research

- Building open, ethically governed student-performance datasets from Indian schools and universities across states, languages and income groups.
- Evaluating fairness across languages of instruction and regional scripts.
- Extending the analysis to intersectional groups, such as rural female students from low-income families, using the joint-group weighting of Section 4.6.
- Evaluating fairness across income levels, social categories and state contexts with Indian institutional data.
- Repeating the experiments over many seeds and splits, with cross-validated hyperparameters, to obtain intervals that include training variability.
- Combining explanation methods such as SHAP and LIME with fairness analysis, to diagnose feature-dependent shifts like the one seen in Section 5.7.
- Studying interfaces that help teachers review and override AI recommendations, and measuring the effect on decision quality and equity.

- Testing the eight BTEAI components through case studies in Indian institutions and structured evaluation with teachers, administrators and educational technology practitioners.

10. Conclusion

This paper examined algorithmic fairness in AI-based education in two parts: an empirical audit of group-level prediction-error disparity on a reproducible benchmark, and a conceptual framework, BTEAI, that connects computational fairness with Indian educational values and policy.

On the UCI Student Performance data, Random Forest had the best aggregate accuracy (test MAE = 1.165, R^2 = 0.820) and also the largest sex-based disparity of the three models (PD = 0.468), while having the smallest address-based disparity (PD = 0.057). A model chosen on aggregate accuracy alone therefore need not be the one that protects a given group best, and the answer depends on the attribute examined. Inverse-frequency reweighting lowered the point estimate of sex-based disparity for every model, by 3.5% to 28.7%, and address-based disparity for two of three, generally without loss of accuracy. Bootstrap intervals on the 79-student test set were wide and included zero, so these gains should be treated as promising and unconfirmed. Removing the prior-grade features changed the disparity pattern, most visibly for XGBoost, which shows that an audit done under one feature set does not carry over to another.

The dataset is Portuguese, and no result is extended to Indian students. Its value here is as a transparent setting in which to show how an audit is done. The BTEAI framework proposes eight interdependent components: data diversity and representation, bias audit, fairness measurement, fairness-aware model development, explainability, contextual validation, teacher and human oversight, and responsible decision-making. NEP 2020's commitments to access, equity, quality and accountability provide the policy setting in which they can be put into practice.

Teacher oversight remains central. Models should inform the professional judgement of teachers and not replace it. Regular, explicit fairness checks combined with human oversight grounded in the local context are the most credible route to educational AI that serves every learner equitably.

Data and Code Availability

The UCI Student Performance dataset is publicly available at https://doi.org/10.24432/C5TG7T. The analysis code (a Google Colab notebook), the saved test predictions and the generated tables and figures are available from the corresponding author on request.

References

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org. https://fairmlbook.org

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI* (Technical Report MSR-TR-2020-32). Microsoft.

Blanzeisky, W., & Cunningham, P. (2022). Using Pareto simulated annealing to address algorithmic bias in machine learning. *The Knowledge Engineering Review, 37*, Article e5. https://doi.org/10.1017/S0269888922000029

Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems, 30* (pp. 3992–4001).

Central Board of Secondary Education. (2026). *CBSE curriculum framework for computational thinking (CT) and artificial intelligence (AI) for Classes III–VIII (Session 2026–27)* (Circular No. Acad-15/2026, 1 April 2026). CBSE.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cortez, P. (2008). *Student performance* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), *Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008)* (pp. 5–12). EUROSIS.

Department of School Education and Literacy. (2025a). *School infrastructure: Computer and internet access, 2022–23 to 2024–25* [Press Information Bureau release PIB2161543]. Ministry of Education, Government of India. https://dsel.education.gov.in/sites/default/files/update/PIB2161543.pdf

Department of School Education and Literacy. (2025b). *UDISE+ 2024–25: Unified District Information System for Education Plus* [Data release]. Ministry of Education, Government of India.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). ACM. https://doi.org/10.1145/2783258.2783311

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education: Practices, challenges, and debates* (pp. 174–202). Routledge. https://doi.org/10.4324/9780429329067-10

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115. https://doi.org/10.1145/3457607

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence, 5*, Article 100152. https://doi.org/10.1016/j.caeai.2023.100152

Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research, Vol. 81, pp. 107–118). PMLR. https://proceedings.mlr.press/v81/menon18a.html

Ministry of Education. (2020). *National Education Policy 2020*. Government of India.

Ministry of Education. (n.d.). *Lok Sabha unstarred question no. 4920: State/UT-wise percentage of schools having electricity, internet connection, computer and CWSN toilets, 2024–25* [Annexure]. Government of India. https://www.education.gov.in/sites/upload_files/mhrd/files/parliament_annexure_en/LSUQ_4920_en.pdf

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in Neural Information Processing Systems, 28*.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), Article e1355. https://doi.org/10.1002/widm.1355

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization.

Viṣṇu Purāṇa. (n.d.). Book 1, chapter 19, verse 41.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenič, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, *2021*, Article 8812542. https://doi.org/10.1155/2021/8812542