MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

Penerbit USIM PRESS

Article

Auditing Gender and Geographic Disparities in Machine Learning Models of KIP Kuliah Award Decisions: A Two-Cohort Temporal Evaluation

Muhammad Thoriq[1,2], Sakinah Ali Pitchay[1,3] and Farida Ridzuan[1,3]

[1] Faculty of Science and Technology, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia.

[2] Department of Informatics, Universitas Adzkia, Padang 25175, West Sumatra, Indonesia.

[3] Cybersecurity and Systems Research Unit, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia.

Correspondence should be addressed to:
Sakinah Ali Pitchay; sakinah.ali@usim.edu.my

Article Info
Article history:
Received: 15 April 2026
Accepted: 15 October 2026
Published: 30 October 2026

Academic Editor:
Azira Khalil

Malaysian Journal of Science, Health & Technology

MJoSHT2026, Volume 12, Issue No. 2
eISSN: 2601-0003

https://doi.org/10.33102/mjosht.597

Copyright © Muhammad Thoriq et al. This is an open access article distributed under the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

***Abstract*— Universities increasingly consider machine learning (ML) to help screen applications for need-based scholarships. Such models learn from past award decisions and may carry forward the group differences they contain, yet they are usually tested on the same intake used to build them. In practice, a model built on one year is applied to the next. This study examines whether a fairness audit of one intake still holds for the following intake. All applications for the Kartu Indonesia Pintar Kuliah (KIP Kuliah) scholarship at a private university in Indonesia in 2023 (n = 708) and 2024 (n = 829) were analysed. Logistic regression, random forest, and XGBoost models were trained on the 2023 cohort and tested once on the 2024 cohort; gender and geographic origin were withheld from the models and used only to compare their selections with past decisions. The models predicted past decisions only slightly better than chance, and their apparent loss of accuracy in 2024 was largely due to a lower acceptance rate rather than to poorer applicant rankings. No consistent gender gap was found. When the models filled a fixed quota, applicants from municipal (city) schools were selected 19–23 percentage points more often than those from regency schools, about twice the gap in the 2024 decisions themselves, because parental occupation acted as an indirect signal of school location. This gap depended on the selection rule and did not hold under every setting. The proposed contribution is a two-intake audit protocol that separates disparities inherited from past decisions from those added by a model. Institutions using such models should re-audit them before every intake.**

***Keywords*—machine learning; algorithmic fairness; decision support; educational financial assistance; temporal validation; KIP Kuliah.**

I. INTRODUCTION

Kartu Indonesia Pintar Kuliah (KIP Kuliah) is Indonesia's main instrument for helping secondary-school graduates from economically disadvantaged households enter higher education. Unlike merit scholarships, it is need-based: applicants must show economic hardship through government social registries or supporting documents, and the admitting institution verifies each record before proposing recipients within its allocation [1–3]. Targeting errors are a persistent problem in Indonesian social assistance [4], and evaluations of the KIP programme have reported inaccurate beneficiary identification and inconsistent distribution [5]. Institutions,

therefore, have practical reasons to seek consistent, data-driven support when many applications must be screened within a short admissions window.

The idea examined in this study is simple. A fairness audit of a selection model is usually conducted once, on the intake data used to build the model. In practice, however, a model built on last year's applicants would be used to rank this year's, so the question that matters to an institution is whether the group differences found in one intake still hold in the next. This study, therefore, builds models on one intake of a national need-based scholarship, audits them on the following intake, and compares every gap produced by a model with the gap already present in the institution's own past decisions. This makes it possible to tell whether a model inherits a disparity from past decisions or adds one of its own, which a single-intake audit cannot do.

A growing body of work applies machine learning (ML) to scholarship and financial-aid selection, using decision trees, naïve Bayes, k-nearest neighbours, random forests, gradient boosting, and ensembles [6–11]. These studies establish technical feasibility, but they share two limitations. First, models are trained and tested on a random partition of a single intake. A model used in practice would be trained on past cohorts and applied to the next one, whose applicant mix, acceptance rate, and administrative conditions may differ; random splits cannot reveal the resulting dataset shift [12]. Second, the target is the institution's historical decision, yet subgroup outcomes are rarely quantified, even when fairness is named as a motivation.

Research on algorithmic fairness in education shows why the second limitation matters. Models fitted to historical records can reproduce the disparities embedded in those records [13, 14]. Group fairness criteria such as demographic parity and equal opportunity formalise different notions of equal treatment and cannot, in general, be satisfied simultaneously [15-18]. Empirical studies have audited student performance, dropout, and admission models using subgroup slicing [19], examined whether protected attributes should be included in a model [20, 21], and documented accuracy–fairness trade-offs [22–24]. A recent mapping study, nevertheless, finds that fairness work in education focuses on performance and dropout outcomes [25]. Need-based award decisions, in which the label itself is an administrative judgement, remain little studied.

Two further gaps motivate this design. The first is temporal. Selection conditions rarely stay fixed between intakes: the number of applicants, their backgrounds, and the number of awards can all change. At the case-study institution, applications rose from 708 to 829 between 2023 and 2024 while the allocation fell from 211 to 195 awards. Comparable shifts occur elsewhere in the region. In Malaysia, for example, the planned number of recipients of the PTPTN eKasih Children's Scholarship, which, like KIP Kuliah, relies on a national poverty registry, was raised from an expected 5,800 to 10,000 in 2026 [26]. Malaysia's higher-education blueprint names equity for students from socio-economically disadvantaged backgrounds as one of its five system aspirations [27], ML has already been applied to Malaysian government scholarship applications [6], and both the ASEAN and the Malaysian guidelines on AI governance list fairness, transparency, and accountability among their principles [28, 29]. Fairness properties measured at one point can fail to hold under such shifts, as shown for census-based benchmarks across years and regions [30] and for clinical models deployed in new settings [31, 32]. The educational audits cited above, however, assess fairness using a single cohort or a random split of pooled data, so their results may not generalise to the next intake. The second gap concerns attribution. A disparity in model selection rates does not show whether the model created it or merely reproduced a gap already present in historical decisions. When the label is an institutional award decision, that distinction bears directly on accountability: an inherited disparity points to the decision process, whereas an amplified one points to the model [33, 34].

Earlier work at the case-study institution modelled a single 2024 cohort, comparing classifiers and interpreting a random forest with SHAP [35-37]. Those studies used random partitions and did not treat fairness as the object of inference. The present study uses two consecutive cohorts and asks three questions: (RQ1) how well do models trained on the 2023 cohort reproduce award decisions in the 2024 cohort, relative to within-cohort estimates; (RQ2) what disparities in selection and error rates arise across gender and geographic origin, and how do they compare with the disparities in the historical decisions; and (RQ3) how stable are these conclusions across cohort, decision-threshold policy, model family, and sampling uncertainty?

The proposed contribution of this study is a pre-specified two-cohort audit protocol for scholarship decision-support models. It separates model-induced from inherited disparity, tests whether disparity conclusions persist under temporal transfer and alternative threshold policies and examines whether retained predictors act as proxies for the audit attributes excluded from training. The study does not claim to measure normative eligibility or to propose a deployable selection system. Although the case is a single Indonesian institution, the protocol requires only historical decisions, pre-decision attributes, and audit attributes, and can therefore be applied to other quota-based allocation decisions such as scholarships, admissions, and social-assistance targeting. Section II describes the data and methods, Section III reports and discusses the results, and Section IV concludes.

II. Material and Method

A. Study Design and Data

The study uses a retrospective two-cohort design (Figure 1). The data comprise all KIP Kuliah applications to Universitas Adzkia, a private university in Padang, West Sumatra, for the 2023 and 2024 intakes, exported from the national KIP Kuliah registration system. The target is the recorded institutional outcome, Accepted or Not Accepted. It is treated as a historical award decision that may reflect documentary verification, allocation limits, and contemporaneous implementation conditions, not as an independent measure of eligibility. The 2023 cohort is used for all model development; the 2024 cohort is used once, as a forward temporal test. The 2025 cohort is reserved for future validation: it was not opened for this study, so that the locked models can later be tested on an intake that played no part in their design.

The national eligibility rules were the same in both years. Applicants qualify through a secondary-school KIP card; registration in DTKS or a related social-assistance programme; P3KE decile 3 or below; residence in a social-care institution; or, failing these, documented gross parental income of at most IDR 4,000,000 per month or IDR 750,000 per capita together with a certificate of economic hardship [2, 3]. Recipients are determined by the national agency, on the institution's proposal, after verification. The institution's allocation varies from year to year and was fully allocated, so the number of recipients equals the allocation: 211 in 2023 and 195 in 2024. In current practice, the allocation combines a regional quota administered by the Higher Education Service Institution (LLDIKTI), part of which is tied to programme accreditation, with a second allocation channel (the 'faction' quota). The split between channels in 2023 and 2024 was not available to the authors, and the dataset does not identify the channel through which each recipient was awarded.

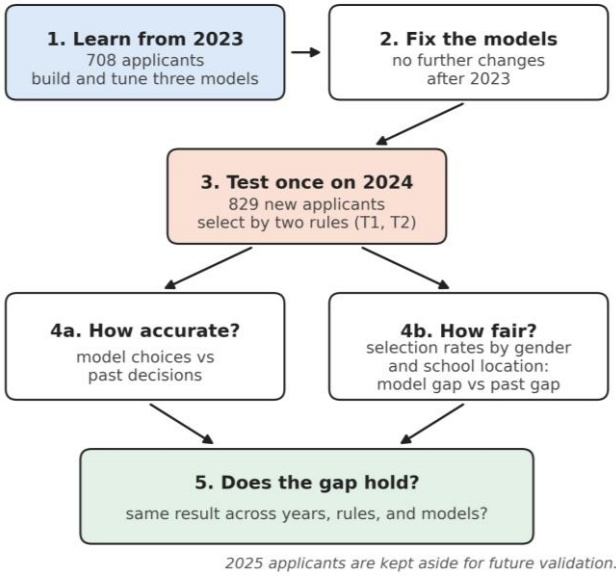


Figure 1. Study design in five steps. Models are built and tuned on the 2023 applicants only, fixed, and then applied once to the 2024 applicants. Their selections are compared with past decisions for accuracy and for gaps between groups, and each gap is checked for whether it holds across years, selection rules, and models. The 2025 applicants are kept aside for future validation.

Direct identifiers (names, national identity and family-card numbers, student numbers, addresses, telephone numbers, and e-mail addresses) were removed before analysis. A one-way hash of the identity number was used only to detect applicants who applied in both years and was then discarded. Use of the de-identified records was authorised by the Institute for Research and Community Service (LPPM) of Universitas Adzkia (permit no. 744/UA.1.2/PI/2026), which restricts use to the stated research and prohibits onward disclosure. TABLE I summarises the cohorts.

TABLE I. COHORT PROFILE OF KIP KULIAH APPLICANTS

Characteristic	2023	2024
Applicants, n	708	829
Accepted, n (%)	211 (29.8)	195 (23.5)
Female / male, n	537 / 171	617 / 212
Accepted: female / male, %	27.2 / 38.0	24.0 / 22.2
School in West Sumatra / other provinces, n	674 / 34	789 / 40
Accepted: West Sumatra / other, %	29.8 / 29.4	23.2 / 30.0
School in municipality / regency, n	158 / 550	170 / 659
Accepted: municipality / regency, %	48.1 / 24.5	31.2 / 21.5
Registered in DTKS, n (%)	538 (76.0)	649 (78.3)
Registered in P3KE, n (%)	415 (58.6)	543 (65.5)
Parental income reported, n (%)	136 (19.2)	144 (17.4)
Also applied in 2023, n	–	37

How to read: the table compares the two intakes; the acceptance rows show the share of each group that was accepted in past decisions. DTKS = integrated social welfare data; P3KE = extreme-poverty targeting data (deciles 1–7). Percentages of accepted are historical acceptance rates within each group.

B. Predictors and Audit Attributes

Candidate predictors were limited to information recorded at application and relevant to economic need: DTKS registration, P3KE status (not registered or decile 1–7), father's and mother's occupation and vital status, number of dependants, and parental income brackets. Consistent with the national rules, income is reported only by applicants outside the registry pathways (97.8% of DTKS-registered applicants left it blank, against none of those registered in neither DTKS nor P3KE); it is therefore structurally not applicable, rather than missing at random and was entered as bracket midpoints with an explicit not-reported indicator instead of being imputed. Programme choice was excluded from the primary model because it may encode programme capacity, and three programmes were first offered in 2024; it is examined in a sensitivity analysis. School name, date of birth, a free-text achievement field, registration date, and a sparsely completed household-asset block that follows the same completion pattern were excluded for re-identification, leakage, or data-quality reasons.

Gender and geographic origin were fixed as the primary audit attributes before modelling. Neither is a direct indicator of economic need; both are recorded before the decision, and both may be linked to differential representation or treatment. They were excluded from the predictor matrix and used only for auditing. Because 95.2% of applicants attended schools in West Sumatra, no single other province reached the minimum group size. Geographic origin was therefore operationalised in two ways: province (West Sumatra versus all other provinces pooled) and school-location type (municipality or kota versus regency or *kabupaten*). The second operationalisation was adopted after descriptive profiling of the cohorts and before any model was fitted. DTKS/P3KE status and income are legitimate-need criteria; disparities between them are reported diagnostically rather than treated as fairness violations.

Reference groups were fixed by one rule: the most populous group in 2023 (female, West Sumatra, regency). Estimates are reported only for groups with at least 30 records; equal opportunity additionally requires at least 10 historical

Accepted cases per group, and odds-based measures at least 10 cases in each outcome class.

C. Preprocessing and Models

All learned transformations were implemented in a single pipeline and fitted only on the 2023 training folds. Categorical variables were one-hot encoded, with levels below 1% pooled and unseen 2024 levels mapped to zero vectors. The number of dependants was parsed to an integer and capped at 10. Missing categorical values were retained as an explicit "not reported" level, and missing numeric values were median-imputed with an indicator. No oversampling was applied; class imbalance was handled by class weights.

Three classifiers were compared: L2-regularised logistic regression (LR) as a transparent baseline, random forest (RF) [38], and XGBoost [39]. Hyperparameters were selected by nested cross-validation on 2023 (outer: stratified 5-fold, repeated five times; inner: stratified 5-fold grid search), using the area under the precision–recall curve (PR-AUC) as the criterion because the Accepted class is the minority [40]. The outer folds yielded out-of-fold (OOF) probabilities for every 2023 applicant, averaged over the five repetitions. Each final model was then tuned on the full 2023 cohort, refitted, and applied once to 2024. RF and XGBoost were refitted with five random seeds; the primary results use the first seed, and the spread across seeds is reported as a sensitivity analysis. The final configurations were LR with C = 0.01; RF with 500 trees, maximum depth 5, and square-root feature sampling; and XGBoost with 200 trees, maximum depth 2, and learning rate 0.1. The resulting T1 thresholds were 0.374, 0.396, and 0.472; grids and fold-level selections are given in the supplementary material.

D. Decision Thresholds

Two threshold policies were pre-specified. T1 (performance-oriented) is the single global threshold that maximises F1 on the 2023 OOF probabilities. T2 (capacity-matched) selects the highest-scoring q of applicants, where q equals the 2023 acceptance proportion (29.8%); it mimics allocation under a fixed quota and uses 2024 scores but not 2024 labels. A descriptive threshold sweep (0.05–0.95) is reported as Supplementary Fig. S1. Group-specific thresholds were not considered because they would require the audit attributes at decision time.

A short example shows the difference. In 2024, there were 829 applicants and 195 awards. Under T2, a model recommends the 247 applicants with the highest scores (29.8% of 829), regardless of those scores. Under T1, it recommends every applicant whose score exceeds the cut-off learned in 2023 (0.374 for LR); in 2024, this meant 770 applicants (92.9%). T1 therefore fixes the required score, whereas T2 fixes the number of applicants selected, as a quota does.

E. Evaluation Measures

Predictive performance is summarised by PR-AUC (primary), F1, balanced accuracy, precision, recall, specificity, ROC-AUC, and the Brier score. Accuracy is reported only alongside the majority-class baseline (0.702 in 2023; 0.765 in 2024); the full set of measures, including accuracy and the Brier score, is given in the supplementary material. The temporal optimism gap for a metric m is $\Delta_m = m_{2023,\mathrm{OOF}} - m_{2024}$. Because PR-AUC depends on prevalence, it is also expressed descriptively as lift, PR-AUC divided by prevalence.

For audit attribute A with group g and reference group r, let $\hat{y}$ be the predicted decision and Y the historical decision. The selection rate and the group fairness measures [15, 17] are

$$\mathrm{SR}_g = \mathrm{P}(\hat{y} = 1 \mid \mathrm{A} = \mathrm{g}) \quad (1)$$

$$\mathrm{DPD}_g = \mathrm{SR}_g - \mathrm{SR}_r, \ \mathrm{DIR}_g = \frac{\mathrm{SR}_g}{\mathrm{SR}_r} \quad (2)$$

$$\mathrm{EOD}_g = \mathrm{TPR}_g - \mathrm{TPR}_r, \ \mathrm{FPRD}_g = \mathrm{FPR}_g - \mathrm{FPR}_r \quad (3)$$

$$\mathrm{AOD}_g = \tfrac{1}{2}(\mathrm{EOD}_g + \mathrm{FPRD}_g) \quad (4)$$

where TPR and FPR are the true- and false-positive rates relative to the historical decision. A disparate impact ratio below 0.80 is used only as a flag. To separate inherited from model-induced disparity, the historical gap and the disparity change are defined as

$$\mathrm{DPD}_g^{\mathrm{hist}} = \mathrm{P}(\mathrm{Y} = 1 \mid \mathrm{A} = \mathrm{g}) - \mathrm{P}(\mathrm{Y} = 1 \mid \mathrm{A} = \mathrm{r}) \quad (5)$$

$$\Delta\mathrm{DP}_g = \mathrm{DPD}_g^{\mathrm{model}} - \mathrm{DPD}_g^{\mathrm{hist}} \quad (6)$$

When the interval for ΔDP excludes zero, the model is described as amplifying (larger gap, same sign), attenuating (smaller gap, same sign), or reversing (opposite sign) the historical disparity; otherwise, it is described as reproducing it. Equal opportunity and related error-rate measures are interpreted relative to past decisions, not to true eligibility.

F. Uncertainty, Stability, and Proxy Analysis

Intervals are 95% percentile intervals from 2,000 bootstrap resamples, stratified by outcome for predictive measures and by audit group for fairness measures [41], with the fitted model held fixed. Model comparisons and ΔDP use paired resamples. For each model family, each disparity was assigned to one of four pre-specified classes, checked in this order. *Inconclusive*: the group-size rule is not met. *Stable*: the interval excludes zero with the same sign in the 2023 OOF and 2024 test predictions under both T1 and T2. *Cohort-dependent*: the interval excludes zero in one cohort only, or the sign differs between cohorts. *Threshold-dependent*: within a cohort, the interval excludes zero under one policy only. Disparities whose intervals include zero throughout are also *inconclusive*. A disparity is considered stable overall if at least two of the three model families are stable; otherwise, the most frequent class is reported.

Excluding an attribute from training does not prevent a model from using correlated inputs [13, 20]. Put simply, removing gender or location from a model does not stop it from guessing them: other inputs, such as a parent's occupation, may reveal whether a family lives in a town or a rural district. Such inputs serve as proxies, allowing a model to reproduce a gap even when the attribute itself is withheld. Proxy strength was therefore estimated as the cross-validated ROC-AUC of LR and RF classifiers predicting each audit attribute from the predictor matrix, where a value of 0.5 means that the attribute cannot be guessed at all. Sensitivity analyses repeat the audit (S1) without the 37 repeat applicants; (S2) with programme choice added; (S3) without the income block; (S4) across five random seeds; and (S5) with the number selected set to the 2024 allocation of 195. Because the allocation is known to the

institution before selection, S5 corresponds to the operational rule; T2 was pre-specified with the 2023 proportion to avoid using any 2024 information. If a disparity is classified as stable, reweighing [42] is examined as an exploratory mitigation, and its performance trade-off is reported.

G. Software and Reproducibility

Analyses were run on Google Colab with Python 3.13, scikit-learn 1.6.1 [43], XGBoost 3.4.1, pandas 2.2.3, and NumPy 2.1.3, using a global seed of 2026. Demographic parity differences were cross-checked against Fairlearn 0.14.0 [44]. The analysis plan was fixed before model fitting. The analysis notebook and the data dictionary are available from the first author on request. The application records cannot be shared publicly under the terms of the data access permit; access to the de-identified data requires approval from LPPM Universitas Adzkia.

III. Results and Discussion

A. Temporal Transfer of Predictive Performance (RQ1)

In plain terms, the models predicted past decisions only slightly better than chance, and the drop in one accuracy measure between the two years was mainly due to a lower acceptance rate rather than the models becoming worse. All three models ranked applicants better than chance, but only modestly (TABLE II). On the 2023 out-of-fold predictions, PR-AUC was 0.377–0.390 against a prevalence of 0.298 and ROC-AUC was 0.604–0.616. On the 2024 cohort, PR-AUC was 0.291–0.303 against a prevalence of 0.235 and ROC-AUC was 0.571–0.583. Paired bootstrap differences in 2024 PR-AUC between models ranged from −0.011 to −0.005, and every interval included zero. The ensembles therefore did not outperform the regularised linear baseline, for which nested tuning selected the strongest penalty in the grid (C = 0.01).

TABLE II. PREDICTIVE PERFORMANCE: 2023 OUT-OF-FOLD VERSUS 2024 FORWARD TEST

Model	PR-AUC 2023 / 2024	Δ PR-AUC (95% CI)	ROC-AUC 2023 / 2024	F1 2024 T1 / T2	BA 2024 T1 / T2
LR	0.379 / 0.303	0.076 (0.012, 0.138)	0.605 / 0.575	0.381 / 0.335	0.510 / 0.553
RF	0.377 / 0.297	0.080 (0.015, 0.143)	0.604 / 0.571	0.386 / 0.335	0.531 / 0.553
XGB	0.390 / 0.291	0.099 (0.034, 0.159)	0.616 / 0.583	0.385 / 0.353	0.568 / 0.567
No-skill	0.298 / 0.235	–	0.500 / 0.500	–	0.500

How to read: higher values mean closer agreement with past decisions, and the No-skill row shows what random guessing would score. 2023 values are out-of-fold (OOF) predictions from nested cross-validation; 2024 values come from a single forward test. BA = balanced accuracy; T1 = F1-optimal threshold from 2023 OOF; T2 = capacity-matched selection (q = 0.298). No-skill PR-AUC equals prevalence.

PR-AUC fell by 0.076–0.099 between the 2023 out-of-fold estimates and the 2024 test, and each interval excluded zero. This decline should not be read as a loss of ranking ability. PR-AUC scales with prevalence, which fell from 29.8% to 23.5% because a smaller allocation (211 to 195 awards) was spread over more applicants (708 to 829). Expressed as lift over prevalence (a descriptive aid added after the analysis), PR-AUC was 1.27–1.31 in 2023 and 1.24–1.29 in 2024. The prevalence-invariant ROC-AUC fell by 0.030–0.033, and its intervals included zero. The apparent optimism thus reflects a shift in the acceptance rate rather than degraded discrimination [12].

The threshold policy mattered more than the model choice. The F1-optimal thresholds chosen on the 2023 cohort selected 92.9% (LR), 80.3% (RF), and 49.1% (XGB) of 2024 applicants, and F1 under T1 fell significantly for every model (Δ = 0.089–0.103). Under T2, the selection rate was fixed at 29.8%, and balanced accuracy did not differ between cohorts (Δ = −0.009 to 0.035; all intervals included zero). A threshold tuned for F1 on one intake is therefore fragile when the acceptance rate changes, whereas quota-style selection transfers more predictably.

The modest discrimination is itself a finding about the label. Pre-decision attributes recorded at the time of application explain only a small part of historical decisions. In 2024, applicants registered in DTKS were accepted at almost the same rate as unregistered applicants (23.4% versus 23.9%). Outcomes, therefore, appear to depend substantially on information that the application records do not capture, such as documentary verification or allocation constraints. Earlier single-cohort analyses at the same institution used random partitions and different variable sets [36, 37], so their estimates are not directly comparable with the forward-temporal estimates reported here.

B. Historical and Model-Induced Disparities (RQ2)

In short, the models did not favour either gender, and the province results were too uncertain to interpret. They did, however, select applicants from municipal schools more often than those from regency schools, by a wider margin than the 2024 decisions themselves. Figure 2 shows these differences as selection rates, and Fig. 3 shows the gaps with their uncertainty.

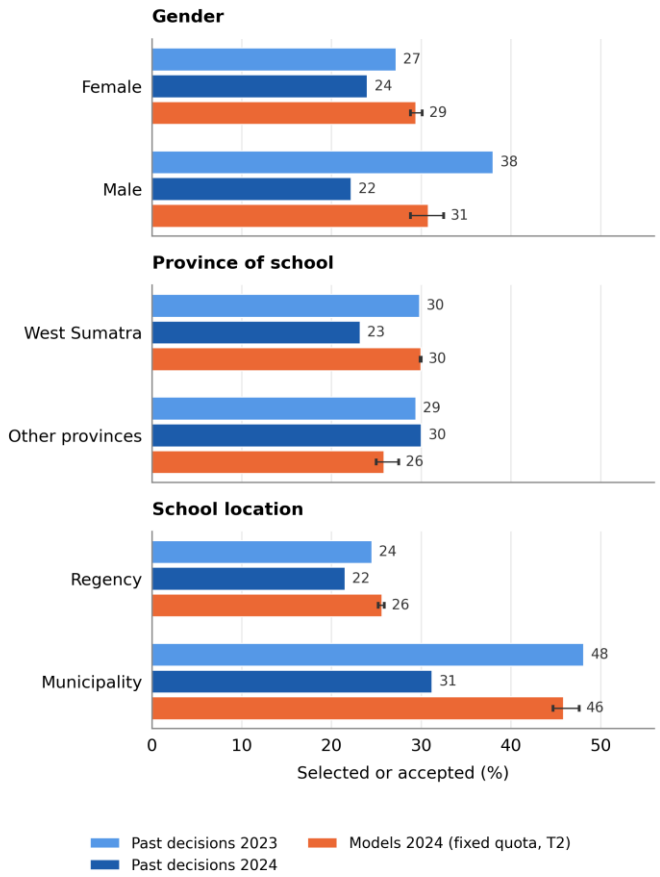


Figure 2. Share of applicants accepted in past decisions (2023 and 2024) and selected by the models in 2024 when filling a fixed quota (T2), by group. Orange bars show the mean of the three models, and the black lines span the lowest to the highest model. The models select 29.8% of applicants overall, more than the 23.5% accepted in 2024, so the levels differ; what matters is the gap between the two groups in each panel. The gender and province gaps are small and uncertain, whereas the gap between municipal and regency schools is about 20 points under the models and 10 points in the 2024 decisions.

Gender. Historical decisions favoured male applicants in 2023 (DPD^{hist} = 0.108; 95% CI, 0.026 to 0.192) but not in 2024 (−0.018; 95% CI, −0.082 to 0.047). No model produced a gender gap in selection rates in either cohort: every DPD interval included zero (TABLE III; Figure 3). In 2023, the models therefore removed most of the historical male advantage, and ΔDP intervals excluded zero for four of the six configurations (ΔDP = −0.096 to −0.146). This is consistent with the absence of a gender proxy among the predictors (Section III-D). Equal-opportunity differences were mostly negative, meaning a lower true-positive rate for historically accepted men, but they reached significance only in isolated cells (2023 XGB-T2: −0.192, −0.330 to −0.052; 2024 LR-T1: −0.094, −0.205 to −0.0005).

Province. The pooled out-of-province group was small (34 and 40 applicants, of whom 10 and 12 were accepted), and its intervals were wide. The historical DPD was −0.004 in 2023 and 0.068 in 2024, and neither was distinguishable from zero. Every model DPD and EOD interval also included zero (TABLE IV). With groups of this size, the audit can neither detect nor rule out a moderate disparity.

TABLE III. GENDER AUDIT, 2024 FORWARD TEST (COMPARISON: MALE; REFERENCE: FEMALE)

Source	SR M / F	DPD (95% CI)	DIR	EOD (95% CI)	ΔDP (95% CI)
Historical	0.222 / 0.240	−0.018 (−0.082, 0.047)	0.92	–	–
LR T1	0.925 / 0.930	−0.006 (−0.047, 0.035)	0.99	−0.094 (−0.205, −0.0005)	0.012 (−0.067, 0.088)
LR T2	0.288 / 0.301	−0.014 (−0.084, 0.059)	0.95	−0.051 (−0.204, 0.103)	0.004 (−0.087, 0.098)
RF T1	0.797 / 0.806	−0.008 (−0.073, 0.053)	0.99	−0.112 (−0.252, 0.009)	0.010 (−0.082, 0.100)
RF T2	0.311 / 0.293	0.018 (−0.055, 0.091)	1.06	−0.023 (−0.179, 0.142)	0.036 (−0.058, 0.130)
XGB T1	0.505 / 0.486	0.018 (−0.058, 0.094)	1.04	−0.083 (−0.238, 0.078)	0.037 (−0.060, 0.132)
XGB T2	0.325 / 0.288	0.037 (−0.038, 0.107)	1.13	−0.022 (−0.178, 0.138)	0.055 (−0.042, 0.146)

How to read: each row gives the share selected among male and female applicants in 2024 and the gap between them; an interval that includes zero means the gap may be due to chance, and ΔDP shows whether a model's gap is larger or smaller than the gap in past decisions. SR = selection rate; DPD = demographic parity difference; DIR = disparate impact ratio; EOD = equal opportunity difference; ΔDP = DPD(model) − DPD(historical). The corresponding 2023 out-of-fold estimates are given in the supplementary material.

School location. Location showed the clearest pattern. Applicants from municipal schools were historically more likely to be accepted in both cohorts (DPD^{hist} = 0.236, 0.151 to 0.321, in 2023; 0.096, 0.021 to 0.174, in 2024). Under capacity-matched selection, every model reproduced this direction in both cohorts, with 2024 DPD ranging from 0.188 to 0.225 and DIR from 1.72 to 1.89 (TABLE IV). The historical gap narrowed between cohorts, but the model gap did not. Models trained on 2023 carried the larger 2023 pattern into 2024, so ΔDP in 2024 was positive for every T2 configuration (0.091–0.128), with intervals excluding zero for RF-T2 (0.128; 0.016 to 0.239) and XGB-T1 (0.115; 0.007 to 0.219). In the 2023 out-of-fold predictions, by contrast, model gaps were smaller than the historical gap (ΔDP = −0.081 to −0.183), significantly so under T1. The additional municipal selections were concentrated among applicants who had not been accepted historically: under T2, the 2024 false-positive-rate difference was 0.189–0.263 with all intervals above zero, whereas the T2 EOD intervals included zero.

TABLE IV. GEOGRAPHIC AUDIT, 2024 FORWARD TEST

Source	SR g / r	DPD (95% CI)	DIR	EOD (95% CI)	ΔDP (95% CI)
Province: other (g) vs West Sumatra (r)					
Historical	0.300 / 0.232	0.068 (−0.068, 0.208)	1.29	–	–
LR T1	0.875 / 0.932	−0.057 (−0.176, 0.039)	0.94	−0.206 (−0.484, 0.043)	−0.125 (−0.328, 0.068)
LR T2	0.250 / 0.300	−0.050 (−0.186, 0.092)	0.83	0.128 (−0.180, 0.432)	−0.118 (−0.278, 0.043)
RF T1	0.725 / 0.807	−0.082 (−0.231, 0.053)	0.90	−0.108 (−0.391, 0.142)	−0.150 (−0.359, 0.046)
RF T2	0.250 / 0.300	−0.050 (−0.184, 0.093)	0.83	0.040 (−0.248, 0.348)	−0.118 (−0.292, 0.057)
XGB T1	0.375 / 0.497	−0.122 (−0.272, 0.035)	0.75	−0.190 (−0.494, 0.112)	−0.190 (−0.392, 0.008)
XGB T2	0.275 / 0.299	−0.024 (−0.158, 0.131)	0.92	−0.071 (−0.342, 0.233)	−0.092 (−0.280, 0.104)
Location: municipality (g) vs regency (r)					
Historical	0.312 / 0.215	0.096 (0.021, 0.174)	1.45	–	–
LR T1	0.941 / 0.926	0.016 (−0.029, 0.053)	1.02	−0.026 (−0.109, 0.047)	−0.081 (−0.170, 0.004)
LR T2	0.447 / 0.259	0.188 (0.108, 0.272)	1.72	0.153 (−0.006, 0.306)	0.091 (−0.016, 0.198)
RF T1	0.876 / 0.785	0.092 (0.034, 0.150)	1.12	0.049 (−0.052, 0.151)	−0.004 (−0.101, 0.086)
RF T2	0.476 / 0.252	0.225 (0.139, 0.304)	1.89	0.101 (−0.059, 0.260)	0.128 (0.016, 0.239)
XGB T1	0.659 / 0.448	0.211 (0.124, 0.293)	1.47	0.168 (0.016, 0.315)	0.115 (0.007, 0.219)
XGB T2	0.453 / 0.258	0.195 (0.111, 0.276)	1.76	0.150 (−0.005, 0.305)	0.099 (−0.009, 0.205)

Read as TABLE III. In the location panel, positive values mean that applicants from municipal schools were selected more often than those from regency schools. Abbreviations as in TABLE III. The out-of-province group had n = 40 (12 accepted) in 2024. DIR values outside 0.80–1.25 are flagged in the supplementary table.

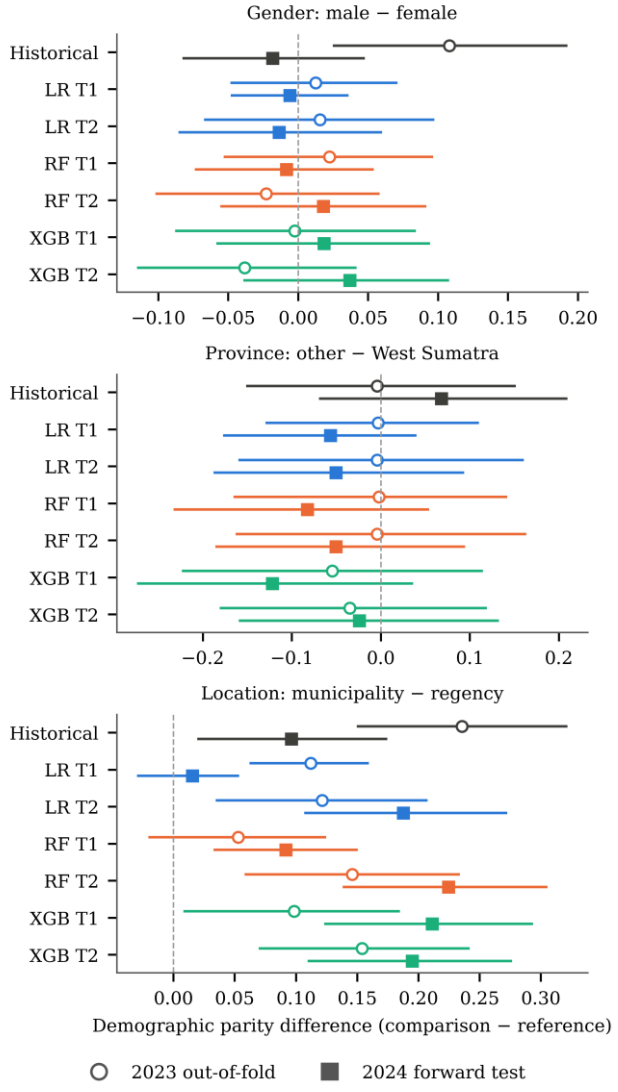


Figure 3. Gap in selection rates between groups (comparison minus reference) with 95% bootstrap intervals, for past decisions and for each model and selection rule. Zero means no gap; an interval that crosses the dashed zero line means the gap could be due to chance. Open markers: 2023 (out-of-fold); filled markers: 2024 (forward test).

C. Stability of Disparity Conclusions (RQ3)

Put simply, no gap held consistently across both years, all three models, and both selection rules; the school-location gap came closest. No disparity met the pre-specified overall stability criterion (TABLE V). The location disparity in selection rates came closest. It was stable for XGBoost and significant under T2 for all three models in both cohorts, but LR-T1 in 2024 and RF-T1 in 2023 did not exclude zero, so the overall class is threshold-dependent. Both exceptions occur where T1 selected 74–93% of applicants. At that level, almost everyone is selected, thereby compressing group differences (Supplementary Fig. S1). The gender equal-opportunity difference was cohort-dependent. For gender selection rates, the inconclusive class means that every interval included zero, although group sizes were adequate, so no disparity was detected; for province, the small comparison group left the intervals too wide to rule out a moderate disparity. All province measures and the location equal-opportunity difference were inconclusive. Because no disparity was classed as stable, the pre-specified exploratory mitigation (reweighing) was not triggered.

The location disparity under T2 was robust across all sensitivity analyses. Point estimates of the 2024 DPD were 0.193–0.238 without repeat applicants (S1), 0.136–0.225 with programme choice added (S2), 0.195–0.210 without the income block (S3), 0.225–0.254 (RF) and 0.173–0.217 (XGB) across five seeds (S4), and 0.163–0.222 when exactly the 2024 allocation of 195 was selected (S5). Adding programme choice also created a gender gap under T2 that was absent in the primary models (DPD = 0.100–0.126). First-choice programmes at the institution are strongly gender-segregated: 87% of applicants to primary teacher education were women, compared with 35% in industrial engineering. This supports the decision to exclude programme choice from the primary predictor set. It also shows a route by which allocation rules can create disparity: where part of the quota is tied to programme accreditation, gender-segregated programme choice can translate into gender differences in awards.

TABLE V. STABILITY CLASSIFICATION OF AUDITED DISPARITIES

Attribute	Measure	LR	RF	XGB	Overall
Gender	DPD	I (ns/ns | ns/ns)	I (ns/ns | ns/ns)	I (ns/ns | ns/ns)	Inconclusive
Gender	EOD	CD (ns/ns | −*/ns)	I (ns/ns | ns/ns)	CD (ns/−* | ns/ns)	Cohort-dependent
Province	DPD	I (ns/ns | ns/ns)	I (ns/ns | ns/ns)	I (ns/ns | ns/ns)	Inconclusive
Province	EOD	I (ns/ns | ns/ns)	I (ns/ns | ns/ns)	I (ns/ns | ns/ns)	Inconclusive
Location	DPD	TD (+*/+* | ns/+*)	TD (ns/+* | +*/+*)	S (+*/+* | +*/+*)	Threshold-dependent
Location	EOD	I (ns/ns | ns/ns)	I (ns/ns | ns/ns)	CD (ns/ns | +*/ns)	Inconclusive

How to read: each cell shows, for one model, whether the gap was statistically clear in each year and under each rule; the class summarises whether the gap held throughout. Cell = class (2023 OOF T1/T2 | 2024 T1/T2). +* / −* = interval excludes zero (sign of comparison − reference); ns = interval includes zero. S = stable; TD = threshold-dependent; CD = cohort-dependent; I = inconclusive.

D. Proxy Structure and Need-Based Diagnostics

Parents' occupation allowed the models to infer school location, but not gender. Location was the only audit attribute with a material proxy. LR and RF predicted municipal schooling from the predictors with ROC-AUC of 0.705 (0.653 to 0.754) and 0.719 (0.668 to 0.769), compared with 0.50 (0.45 to 0.55) for gender and 0.56–0.62 for province. The largest logistic coefficients were for parental occupation: households of fishers and farmers were associated with regency schools, whereas the 'other' and 'not working' categories were associated with municipal schools. The models can thus reconstruct school location through livelihood indicators, and withholding location from training did not prevent the disparity [20, 21].

Compared with the historical decisions, all models concentrated their selections on registered and lower-decile applicants. Under T2 in 2024, 34–35% of DTKS-registered applicants were selected, compared with 12–16% of unregistered applicants, whereas their historical acceptance rates were 23.4% and 23.9%, respectively. Selection rates for P3KE decile 1 were 61–74% compared with a historical rate of 34%, whereas deciles 5–7 were rarely selected (0–16%). This is a policy-relevant differentiation in line with the programme's need criteria, which admit P3KE households only up to decile 3, not a fairness violation. It shows, however, that the models weight the registries more heavily than past decisions did, while selecting regency applicants less often, although their livelihood profiles are typical of rural need.

E. Practical Implications for Institutions

Three findings matter for scholarship decision support. First, a fairness audit of a single intake can be misleading. An audit of 2023 alone would have highlighted a historical male advantage that the models did not reproduce, and an audit of the 2024 decisions alone would have shown a modest location gap without revealing that models trained on the previous intake tended to widen it. Fairness properties that do not transfer across time have been reported in other domains [30–32]; the present results show the same behaviour in a need-based award setting. Second, the historical gap and the model gap should be reported together, because separating inherited from model-induced disparity changed the interpretation of both the gender and the location results. Third, the location gap arose without location being an input, through occupational proxies, and was concentrated among applicants whose past decisions had not been accepted.

For institutions, these findings translate into a short routine. Before each intake, the model should be re-audited on the most recent intake whose decisions are final, and the selection rate of each audited group should be reported beside the rate in past decisions. Selection should use a fixed number of places matched to the allocation rather than a score cut-off carried over from an earlier year, because cut-offs did not transfer. Inputs that can reveal protected or geographic attributes, such as parental occupation, should be checked in every audit. Model output should direct human attention rather than decide; in this case, reviewers would look more closely at municipal applicants recommended by a model but not supported by the documentary record, and would record the reasons for each final decision so that the reasons can be traced [33, 34]. These steps correspond to the fairness, transparency, accountability, and human-involvement principles of the ASEAN and Malaysian AI governance guidelines [28, 29], and they would allow a Malaysian institution to show that a model supports, rather than undermines, the equity aspiration of the national higher-education blueprint [27]. Registry-based awards such as the PTPTN eKasih Children's Scholarship share the conditions studied here, a poverty registry and an allocation that changes over time [26], so the protocol could be applied to them with the same data requirements; whether the same patterns appear there is an empirical question. None of these results shows that the models would improve the fairness or quality of award decisions.

F. Limitations

The study has several limitations. Most importantly, the target is the institution's past decision, not a measure of who truly needed the award. Past decisions may contain their own biases, so the results show how the models reproduce or change those decisions, not whether they identify eligible applicants; equal-opportunity and error-rate measures describe agreement with past decisions only. Some groups were small: only 34 and 40 applicants came from schools outside West Sumatra, so the

province's results are unreliable and cannot rule out a moderate gap. Discrimination was modest, and recorded attributes explained little of the outcome; the disparities, therefore, concern a weak decision-support signal, which is nonetheless the realistic case for this institution. The data come from one institution and two intakes, which limits generalisation to the analytical level, and the Malaysian and wider ASEAN relevance discussed above is argued rather than tested. The school-location operationalisation was adopted after descriptive profiling and before model fitting, and the lift ratio was added after the analysis as a descriptive aid. National eligibility rules did not change between the cohorts [2, 3], so the temporal results reflect changes in applicant mix and institutional allocation rather than in eligibility criteria. The historical label also merges awards made through different allocation channels that the data does not distinguish. Many cells were tested, and conclusions rest on the pre-specified stability rules rather than on individual significant cells. Confirmatory evaluation on the 2025 cohort, together with analysis of explanation stability, is left to subsequent work.

IV. CONCLUSION

This study examined gender and geographic disparities in KIP Kuliah award decision models across two consecutive cohorts. Models trained on 2023 ranked 2024 applicants only modestly better than chance, and ROC-AUC did not change significantly between cohorts. The fall in PR-AUC reflected a lower acceptance rate, and F1-optimal thresholds failed to account for this shift, whereas capacity-matched selection did. No stable gender disparity was found, and the models did not reproduce the male advantage present in the 2023 decisions. Province-level disparities could not be assessed reliably due to small sample sizes. Applicants from municipal schools were selected at higher rates across all capacity-matched configurations in both cohorts. The models reconstructed this pattern from occupational proxies and carried it into a cohort in which the historical gap had narrowed. Under the pre-specified rules, however, the disparity was classed as threshold-dependent rather than stable.

The proposed contribution is an audit protocol that separates inherited from model-induced disparity and tests whether disparity conclusions survive changes in intake and in the selection rule. Applied here, it showed that single-intake fairness estimates are not reliable evidence for the next intake. Because the protocol needs only routinely recorded decisions and attributes, it can be used by other institutions and programmes that allocate a fixed number of awards. Institutions that consider such models should re-audit them before every intake on the most recent cohort with final decisions, report each group's selection rate under the model beside its rate in past decisions, select by a quota-matched rule rather than a fixed score cut-off, screen inputs for proxies of protected and geographic attributes, and keep final decisions with human reviewers who record their reasons. Future work will evaluate the locked models on the 2025 cohort, assess the stability of the models' explanations, and embed these audit outputs into an accountable human-review workflow.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

ACKNOWLEDGMENT

The authors thank the Institute for Research and Community Service (LPPM), Universitas Adzkia, for permission to access the de-identified KIP Kuliah application records (permit no. 744/UA.1.2/PI/2026).

DECLARATION OF AI-ASSISTED TOOLS USAGE

The authors conceived the study, defined the research questions within the first author's PhD research, obtained the data and the data access permit, provided the institutional knowledge of the KIP Kuliah selection process, selected and approved the study design and analysis protocol, ran all analyses, interpreted the results, and revised the manuscript with supervisory review. During the preparation of this work, the authors used Claude (Anthropic), a generative AI tool, as a supporting tool to draft and check parts of the Python analysis code, to draft and edit parts of the manuscript text, to prepare figures from the analysis outputs, and to check bibliographic details. All AI-assisted output was reviewed, verified against the analysis outputs, and edited by the authors, who take full responsibility for the content of the manuscript.

REFERENCES

[1] World Bank, *The Promise of Education in Indonesia*. Washington, DC, USA: World Bank, 2020, doi: 10.1596/34807.

[2] Pusat Layanan Pembiayaan Pendidikan, *Pedoman Pendaftaran Kartu Indonesia Pintar Kuliah (KIP Kuliah) Merdeka Tahun 2023*, ver. 3.0. Jakarta, Indonesia: Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, 2023. [Online]. Available: https://kip-kuliah.kemdiktisaintek.go.id/uploads/Pedoman-Pendaftaran-KIP-Kuliah-2023-update-15Mei2023-_e79a89.pdf

[3] Pusat Layanan Pembiayaan Pendidikan, *Pedoman Pendaftaran KIP Kuliah Merdeka Tahun 2024*, ver. 1.0. Jakarta, Indonesia: Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, 2024. [Online]. Available: https://kip-kuliah.kemdikbud.go.id/uploads/20240207-Pedoman-Pendaftaran-KIP-Kuliah-2024_dfb2c1.pdf

[4] V. Alatas, A. Banerjee, R. Hanna, B. A. Olken, R. Purnamasari, and M. Wai-Poi, "Self-targeting: Evidence from a field experiment in Indonesia," *J. Polit. Econ.*, vol. 124, no. 2, pp. 371–427, 2016, doi: 10.1086/685299.

[5] P. Ninghardjanti, W. Murtini, A. Hindrayani, and K. B. Sangka, "Evaluation of the Smart Indonesia Card (KIP) program for the vocational schools in Surakarta City," in *Proc. 1st UPY Int. Conf. Educ. Soc. Sci. (UPINCESS 2022)*, Atlantis Press, 2023, pp. 217–226, doi: 10.2991/978-2-494069-39-8_20.

[6] W. D. Ahmad and A. A. Bakar, "Ensemble machine learning model for higher learning scholarship award decisions," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 303–312, 2020, doi: 10.14569/IJACSA.2020.0110540.

[7] H. Hendri *et al.*, "A hybrid data mining for predicting scholarship recipient students by combining K-means and C4.5 methods," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 3, pp. 1726–1735, 2024, doi: 10.11591/ijeecs.v33.i3.pp1726-1735.

[8] M. H. Ismail, T. R. Razak, N. M. Noor, and A. A. Aziz, "Evaluating machine learning algorithms for predicting financial aid eligibility: A comparative study of random forest, gradient boosting and neural network," in *Proc. 18th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, 2024, pp. 1–6, doi: 10.1109/IMCOM60618.2024.10418450.

[9] A. T. Remolado, R. V. Perocho, and C. C. Hormillada, "Predicting scholarship eligibility using machine learning: A data-driven approach," in *Proc. 9th Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, 2025, pp. 421–426, doi: 10.1109/ICITISEE68184.2025.11355053.

[10] T. Suprapti, B. Nurhakim, B. W. A. Hermina, and V. A. S. Simbolon, "A decision tree model with grid search optimization for scholarship recipient classification," *J. Tek. Inform. (JUTIF)*, vol. 6, no. 5, pp. 3800–3813, 2025, doi: 10.52436/1.jutif.2025.6.5.5235.

[11] B. Isnanto and R. Sulaiman, "An empirical comparison of C4.5, naive Bayes, and KNN for scholarship selection," *J. Inf. Syst. Informatics*, vol. 8, no. 3, pp. 3252–3268, 2026, doi: 10.63158/journalisi.v8i3.1617.

[12] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognit.*, vol. 45, no. 1, pp. 521–530, 2012, doi: 10.1016/j.patcog.2011.06.019.

[13] R. S. Baker and A. Hawn, "Algorithmic bias in education," *Int. J. Artif. Intell. Educ.*, vol. 32, no. 4, pp. 1052–1092, 2022, doi: 10.1007/s40593-021-00285-9.

[14] R. F. Kizilcec and H. Lee, "Algorithmic fairness in education," in *The Ethics of Artificial Intelligence in Education*. New York, NY, USA: Routledge, 2022, pp. 174–202, doi: 10.4324/9780429329067-10.

[15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3315–3323.

[16] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017, doi: 10.1089/big.2016.0047.

[17] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Sci. Rep.*, vol. 12, Art. no. 4209, 2022, doi: 10.1038/s41598-022-07939-1.

[18] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–38, 2024, doi: 10.1145/3616865.

[19] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in *Proc. 9th Int. Conf. Learn. Anal. Knowl. (LAK)*, 2019, pp. 225–234, doi: 10.1145/3303772.3303791.

[20] R. Yu, H. Lee, and R. F. Kizilcec, "Should college dropout prediction models include protected attributes?," in *Proc. 8th ACM Conf. Learn. @ Scale*, 2021, pp. 91–100, doi: 10.1145/3430895.3460139.

[21] O. B. Deho, S. Joksimovic, J. Li, C. Zhan, J. Liu, and L. Liu, "Should learning analytics models include sensitive attributes? Explaining the why," *IEEE Trans. Learn. Technol.*, vol. 16, no. 4, pp. 560–572, 2023, doi: 10.1109/TLT.2022.3226474.

[22] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Fair and transparent student admission prediction using machine learning models," *Algorithms*, vol. 17, no. 12, Art. no. 572, 2024, doi: 10.3390/a17120572.

[23] R. A. Opoku, B. Pei, and W. Xing, "Unveiling accuracy-fairness trade-offs: Investigating machine learning models in student performance prediction," *J. Learn. Anal.*, vol. 12, no. 2, pp. 125–139, 2025, doi: 10.18608/jla.2025.8543.

[24] K. Kesgin, S. Kiraz, S. Koşunalp, and B. Stoycheva, "Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning," *Appl. Sci.*, vol. 15, no. 15, Art. no. 8409, 2025, doi: 10.3390/app15158409.

[25] N. Pham, H. Pham Ngoc, and A. Nguyen-Duc, "Fairness for machine learning software in education: A systematic mapping study," *J. Syst. Softw.*, vol. 219, Art. no. 112244, 2025, doi: 10.1016/j.jss.2024.112244.

[26] Bernama, "PTPTN BAeK scholarship expanded to benefit 10,000 students," *Bernama*, Jun. 14, 2026. [Online]. Available: https://www.bernama.com/en/news.php?id=2568298

[27] Ministry of Education Malaysia, *Malaysia Education Blueprint 2015–2025 (Higher Education)*. Putrajaya, Malaysia: Ministry of Education Malaysia, 2015.

[28] Association of Southeast Asian Nations, *ASEAN Guide on AI Governance and Ethics*, Feb. 2024. [Online]. Available: https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf

[29] Ministry of Science, Technology and Innovation, *National Guidelines on Artificial Intelligence Governance and Ethics (AIGE)*. Putrajaya, Malaysia: MOSTI, Sep. 2024. [Online]. Available: https://www.malaysia.gov.my/en/my-initiative/whole-government-digital-services/tadbir-urus/national-guidelines-on-artificial-intelligence-governance-and-ethics-aige

[30] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," in *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 6478–6490.

[31] J. Schrouff *et al.*, "Diagnosing failures of fairness transfer across distribution shift in real-world medical settings," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 19304–19318.

[32] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, "The limits of fair medical imaging AI in real-world generalization," *Nat. Med.*, vol. 30, no. 10, pp. 2838–2848, 2024, doi: 10.1038/s41591-024-03113-4.

[33] B. Memarian and T. Doleck, "Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review," *Comput. Educ. Artif. Intell.*, vol. 5, Art. no. 100152, 2023, doi: 10.1016/j.caeai.2023.100152.

[34] M. Zajko, "Automated government benefits and welfare surveillance," *Surveill. Soc.*, vol. 21, no. 3, pp. 246–258, 2023, doi: 10.24908/ss.v21i3.16107.

[35] M. Thoriq, F. Maulana, Y. S. Eirlangga, N. Hayati, and M. A. Madani, "Implementasi algoritma naïve Bayes dalam prediksi penerimaan mahasiswa penerima beasiswa KIP di Universitas Adzkia," *J. Fasilkom*, vol. 15, no. 1, pp. 108–114, 2025, doi: 10.37859/jf.v15i1.8959.

[36] M. Thoriq, F. Maulana, and A. Q. Ayun, "Perbandingan naïve Bayes dan random forest pada seleksi KIP di Universitas Adzkia," *J. Pustaka AI*, vol. 5, no. 3, pp. 602–610, 2025, doi: 10.55382/jurnalpustakaai.v5i3.1385.

[37] M. Thoriq, F. Maulana, S. A. Pitchay, F. H. M. Ridzuan, and W. Ramanda, "Explainable AI pada seleksi beasiswa KIP berbasis machine learning di Universitas Adzkia," *J. Pustaka AI*, vol. 6, no. 1, pp. 80–91, 2026, doi: 10.55382/jurnalpustakaai.v6i1.1767.

[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[40] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[41] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.

[42] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012, doi: 10.1007/s10115-011-0463-8.

[43] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[44] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and improving fairness of AI systems," *J. Mach. Learn. Res.*, vol. 24, no. 257, pp. 1–8, 2023.