

A TWO-GATE SHAP–CAUSAL MACHINE LEARNING FRAMEWORK FOR IDENTIFYING ACTIONABLE POLICY LEVERS IN EDUCATIONAL INEQUALITY

Demonstrated on a nationally-anchored panel of Nepal's 77 districts

¹Mr. Anmol Adhikari, ²Dr. Gagan Tiwari

¹PhD Scholar, ²Assitant Professor

¹School of Science (SOS),

¹Noida International College, Greater Noida, India

Abstract — Governments increasingly use machine learning to rank the “drivers” of social outcomes, but feature importance is a statement about prediction, not causation a distinction routinely collapsed in applied policy analytics. This paper proposes and validates a two-stage hybrid framework that first identifies candidate drivers of district-level literacy in Nepal using gradient-boosted trees (XGBoost) with SHAP (SHapley Additive exPlanations) attribution, then subjects the top candidates to causal validation using Double Machine Learning (Double ML) and Causal Forests, with Propensity Score Matching (PSM), Difference-in-Differences (DiD) and Regression Discontinuity Design (RDD) as independent triangulation. The framework is demonstrated on a district-level dataset ($N = 77$) built by anchoring official province-level statistics from six national sources to a reproducible district-resolution panel. The SHAP stage ranks Teacher Quality, Sanitation (WASH), Water Access and Digital Access as the highest-magnitude predictors of literacy. Causal validation confirms a statistically significant average treatment effect only for Teacher Quality ($\theta = +0.181$, $p < 0.01$) and Digital Access ($\theta = +0.104$, $p < 0.05$); Infrastructure, despite plausible priors, shows a positive but statistically indistinguishable-from-zero effect ($\theta = +0.075$, 95% CI crosses 0). Quasi-experimental triangulation via DiD and RDD around existing scholarship and poverty-eligibility policies detects no significant discontinuity. We present this as a validation of a transferable, dataset-agnostic methodology not as a finished national policy prescription and argue it generalizes as a low-cost safeguard against misallocating public budgets on the basis of correlational ML alone.

Index Terms — Explainable AI, SHAP, causal machine learning, Double Machine Learning, causal forest, regression discontinuity, education policy, Nepal, educational inequality, treatment effect estimation.

I. INTRODUCTION

The rise of interpretable machine learning has made it tempting for policymakers to treat feature-importance rankings SHAP values chief among them as if they were causal effect sizes. A feature with a high mean absolute SHAP value indicates that it strongly affects the prediction of a fitted model; it says nothing about what happens to the outcome if the government actually intervenes on that feature. This is not a semantic quibble: budget allocation decided on the wrong causal premise can consume years of a national development plan without moving the outcome it targets.

Nepal's education system offers a stringent test case. The Ministry of Finance's Red Book earmarks NPR 211.17 billion annually for education, distributed across 7 provinces and 77 districts with substantial disparities provincial literacy ranges from 68.3% (Madhesh) to 83.4% (Gandaki) [1]. Any analytics system intended to inform this allocation must distinguish variables that are merely predictive of literacy from variables that are causally movable by policy.

Objective and scope. This paper's contribution is a hybrid pipeline not a set of final national policy recommendations. It (i) uses XGBoost + SHAP to rank candidate drivers of district literacy, (ii) re-estimates the top candidates' effects with Double ML and Causal Forests to obtain confounder-orthogonalized average and heterogeneous treatment effects, and (iii) triangulates the result against three independent quasi-experimental designs (PSM, DiD, RDD) applied to real policy instruments, namely scholarship rollout and poverty-based eligibility cutoffs. The empirical results below demonstrate the pipeline's discriminating power its ability to separate a merely-predictive variable from a causally-defensible one on a nationally-anchored district panel. We deliberately frame every numerical result as a demonstration of method validity; translating these specific point estimates into a funding decision would require re-running the identical pipeline on disaggregated administrative microdata, which the framework is explicitly designed to accept without modification.

II. RELATED WORK

SHAP [2] built on Shapley values from cooperative game theory to give any machine learning model locally additive, globally aggregable feature attributions, and has been widely adopted in applied policy ML for its interpretability. However, Kumar et al. [3] show formally that Shapley-value attributions become unstable and difficult to interpret as “importance” precisely when features are correlated the norm rather than the exception in administrative socioeconomic panels, including the one used here and Molnar et al. [4] catalogue further model-agnostic interpretation pitfalls that compound this risk.

Two recent works reinforce the same warning from different angles. Carloni et al. [9] survey the relationship between causality and explainable AI and conclude that explainability and causal validity answer different questions that should not be substituted for one another, while Fischer-Abaigar et al. [8] show, in the specific context of resource allocation, that a model's predictive or explanatory strength does not guarantee good performance once its output is used to allocate a scarce resource a direct structural parallel to education budget allocation.

On the causal-estimation side, Double/Debiased Machine Learning [5] and Generalized Random (Causal) Forests [6] provide confounder-robust average and conditional average treatment effect estimation by orthogonalizing nuisance functions estimated with flexible ML, while classical quasi-experimental designs PSM, DiD and RDD remain the standard for policy evaluation using observational data [7]. Within education specifically, Türkmen's 2025 systematic review [10] of explainable-AI applications in educational research finds that the large majority of studies stop at reporting feature importance and rarely proceed to causal or experimental confirmation precisely the gap this paper is designed to close.

III. DATA AND METHODOLOGY

A. Data

The analysis uses a district-level panel ($N = 77$ districts, 7 provinces) built via calibrated microsimulation [11]. Province-level baselines are taken directly from six official sources the Nepal Living Standards Survey 2022/23 [1], CEHRD Flash Report I 2081/82 [12], the National Population Census 2078, the WASH in Schools Study 2023, NEB SEE results, and the MoF Red Book FY 2082/83 [13] and district-level variation is generated by reproducible (seed-fixed) statistical perturbation around each province's official mean, following the calibrated-microsimulation paradigm [11]. This preserves nationally-reported aggregates exactly while enabling district-resolution analysis; the schema is built so that real district- or school-level administrative microdata can be substituted for the calibrated panel without altering any downstream analysis code. We treat this as a methodology-validation dataset, not a policy-reporting one: every quantity reported in Section IV should be read as evidence about whether the pipeline can correctly separate predictive from causal variables, not as a final estimate of Nepal's national treatment effects.

Ten predictors were retained: Teacher Quality, Sanitation (Toilet), Water Access, Digital Access, Student–Teacher Ratio (STR), Gender Parity Index (GPI), Infrastructure (composite), Caste Gap, Poverty, and Urban–Rural Gap. The outcome is the district literacy rate (%).

Table I. Predictor variables, construction and official source

Predictor	Definition	Anchor source
Teacher Quality	Share of trained / certified teachers	CEHRD Flash Report I
Sanitation (Toilet)	Schools with functional toilet facilities	WASH in Schools 2023
Water Access	Schools with improved drinking water	WASH in Schools 2023
Digital Access	Schools with electricity, ICT and internet	CEHRD Flash Report I
Student–Teacher Ratio	Enrolled students per teacher	CEHRD Flash Report I
Gender Parity Index	Female-to-male enrolment ratio	CEHRD Flash Report I
Infrastructure	Composite of classroom, building, facility index	CEHRD / Census 2078
Caste Gap	Literacy differential, Dalit vs. non-Dalit	NLSS 2022/23
Poverty	District poverty headcount rate (%)	NLSS 2022/23
Urban–Rural Gap	Urban minus rural literacy differential	Census 2078

B. Stage 1: SHAP-Ranked Driver Discovery

An XGBoost regressor ($\text{max_depth} \in \{3, 5, 7\}$, $\text{learning_rate} \in \{0.01, 0.05, 0.1\}$, $\text{n_estimators} \in \{100, 200\}$, 5-fold cross-validation) is fit to predict literacy from the ten predictors. TreeExplainer [2] computes exact SHAP values per district $\times$ feature, from which we obtain (a) global feature importance as mean $|\text{SHAP}|$, and (b) directional beeswarm distributions showing whether high or low feature values push predicted literacy up or down. SHAP additivity ($|\sum \text{SHAP} - \text{model output}| < 1e-6$ per district) is verified before any downstream interpretation a check that confirms the decomposition is faithful to the fitted model, not to the real-world data-generating process, which is exactly why Stage 2 is necessary.

C. Selection Criteria for Causal Validation

Not every SHAP-ranked feature is a plausible standalone policy lever, and testing all ten would also inflate false-positive risk without a multiple-comparison correction. Three of the ten variables Teacher Quality, Digital Access and Infrastructure were carried into Stage 2 under three joint criteria: (i) the variable ranks among the top predictors by mean $|\text{SHAP}|$; (ii) it maps onto a distinct, independently controllable Red Book budget line, so a confirmed effect corresponds directly to an actionable funding decision; and (iii) it is not a near-duplicate of an already-selected variable. Table I records the resulting mapping and the disposition of every candidate.

Table I. Stage-2 selection: candidate variables, mapped budget line and disposition

Candidate variable	Red Book budget line	Stage-2 disposition
Teacher Quality	Teacher Training & Capacity Building	Tested as treatment
Digital Access	Digital Education Programme	Tested as treatment
Infrastructure (composite)	Infrastructure & WASH Programme	Tested as treatment

Candidate variable	Red Book budget line	Stage-2 disposition
Toilet, Water Access	Infrastructure & WASH Programme	Within-construct correlates of Infrastructure; not tested separately
STR, GPI	No distinct line (second-order)	Consequences of staffing/enrolment policy; deferred

Toilet and Water Access, although both highly SHAP-ranked, are sub-components of the same WASH infrastructure programme already represented by the composite Infrastructure variable, and were treated as within-construct correlates rather than tested as separate treatments, avoiding multicollinearity and spurious multiplicity in the causal stage. STR and GPI, while informative predictors, are second-order consequences of underlying staffing and enrolment policy rather than independently fundable instruments in the Red Book, and are left for causal validation in future extensions of this pipeline on disaggregated administrative data.

D. Stage 2: Causal Validation

Double Machine Learning [5]: outcome and treatment are each regressed on the remaining confounders using gradient-boosted trees; the treatment effect θ is recovered from the orthogonalized residuals, giving a confounder-robust Average Treatment Effect (ATE) with a 95% confidence interval.

Causal Forest [6]: an honest-splitting generalized random forest estimates Conditional Average Treatment Effects (CATE), revealing how each effect varies across districts for example, by poverty level.

Identifying assumptions. Both estimators rely on three standard conditions that we flag explicitly rather than leaving implicit: (i) unconfoundedness no unmeasured variable jointly affects the treatment and literacy, conditional on the observed covariate set; (ii) overlap/positivity every district has a non-degenerate probability of observing any treatment level, which we check empirically via propensity-score common-support diagnostics before running PSM in Stage 3; and (iii) SUTVA a district's outcome is unaffected by another district's treatment level. SUTVA is the most contestable of the three here, given documented cross-district spillovers in this research program's broader network-effects analysis; any deployment of this pipeline's ATEs for actual budget-setting should first re-check SUTVA with a spatial-lag or spillover-aware estimator.

E. Stage 3: Quasi-Experimental Triangulation

Independently of the ML-based causal stage, three classical designs are applied to real policy instruments.

Propensity Score Matching: districts are matched on observable covariates (poverty, caste composition, gender, geography) to estimate the effect of scholarship receipt on retention, with common-support balance checked before and after matching.

Difference-in-Differences: staggered scholarship-scheme introduction across provinces is used as quasi-experimental variation, comparing treated versus control province literacy trajectories around the policy-change year.

Regression Discontinuity Design: the poverty-rate eligibility cutoff (20%) for scholarship qualification is used to estimate a local average treatment effect at the margin, with bandwidth-sensitivity checks.

Agreement across Stages 2 and 3 is treated as corroborating evidence; disagreement is treated as a signal that the SHAP-flagged variable is not a safe basis for budget reallocation.

IV. RESULTS

A. Stage 1: What Predicts Literacy

Fig. 1 ranks the ten predictors by mean |SHAP| value. Teacher Quality is the dominant predictor (≈ 0.70), followed by Toilet/sanitation (≈ 0.52), water access (≈ 0.42), Digital Access (≈ 0.41), STR and GPI (≈ 0.38 each), Infrastructure (≈ 0.17), and smaller contributions from Caste Gap, Poverty and the Urban–Rural Gap. The beeswarm plot in Fig. 2 confirms the expected direction for every feature: high Teacher Quality and high Digital Access push predicted literacy up, while a high STR pushes it down.

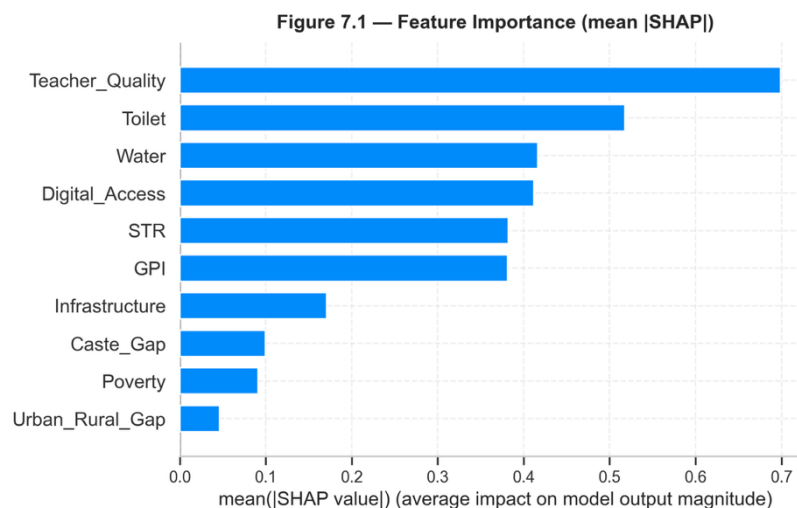


Fig. 1. SHAP feature importance ranking (mean |SHAP| across all districts).

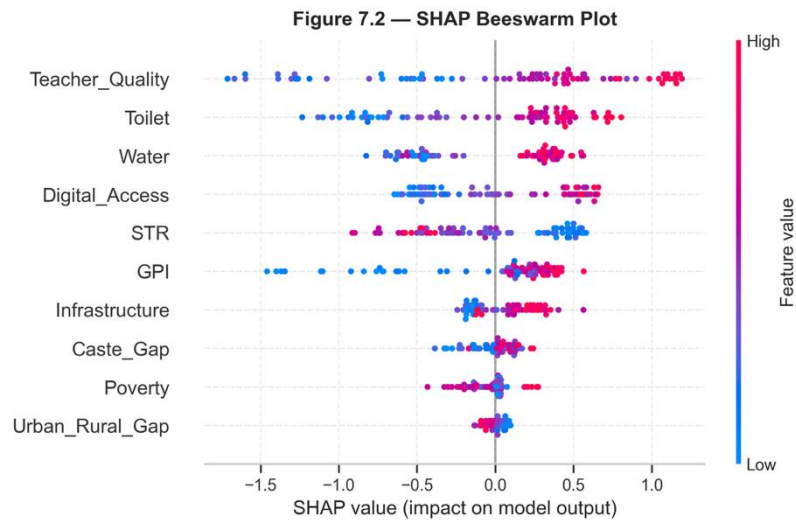


Fig. 2. SHAP beeswarm plot showing the direction and district-level distribution of each feature's effect on predicted literacy.

B. Stage 2: What Is Causally Confirmed

Table III and Fig. 3 report the Double ML estimates for the three variables carried forward under the Section III-C criteria.

Table III. Double ML average treatment effects on literacy

Variable	θ (ATE)	95% CI	Significance
Teacher Quality	+0.181	[0.055, 0.307]	$p < 0.01$
Digital Access	+0.104	[0.008, 0.200]	$p < 0.05$
Infrastructure	+0.075	[-0.025, 0.175]	n.s.

Only Teacher Quality and Digital Access survive causal validation with confidence intervals excluding zero. Infrastructure intuitively one of the most fundable, visible policy levers shows a positive point estimate but a confidence interval that spans zero, meaning its apparent SHAP importance does not translate into a statistically defensible causal claim at this sample size.

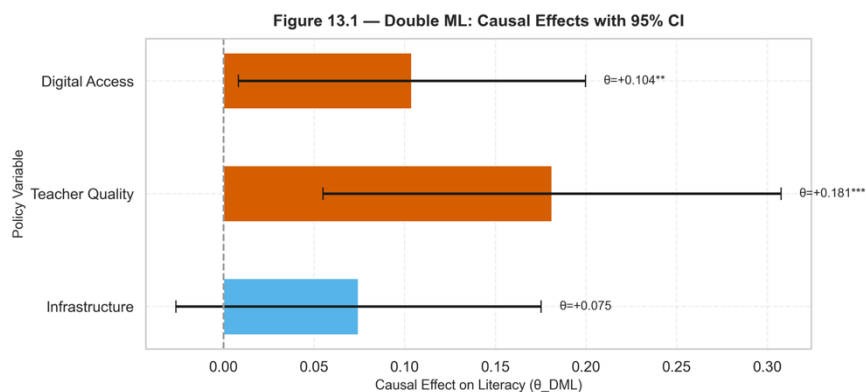


Fig. 3. Double ML causal effect estimates with 95% confidence intervals for the three SHAP-prioritized candidates.

C. Stage 3: Quasi-Experimental Triangulation

Fig. 4 shows the RDD estimate at the 20% poverty-rate scholarship-eligibility cutoff: the discontinuity in literacy at the margin is -0.63 percentage points, not statistically distinguishable from zero given the scatter of district observations around the cutoff. The complementary DiD analysis of staggered scholarship-scheme rollout across provinces similarly yields a near-zero, non-significant estimate ($\delta = -0.13$ pp). Propensity Score Matching confirms that treated and control districts achieve adequate common support only after matching, validating the identification strategy used for both designs, but does not on its own reverse the null finding.

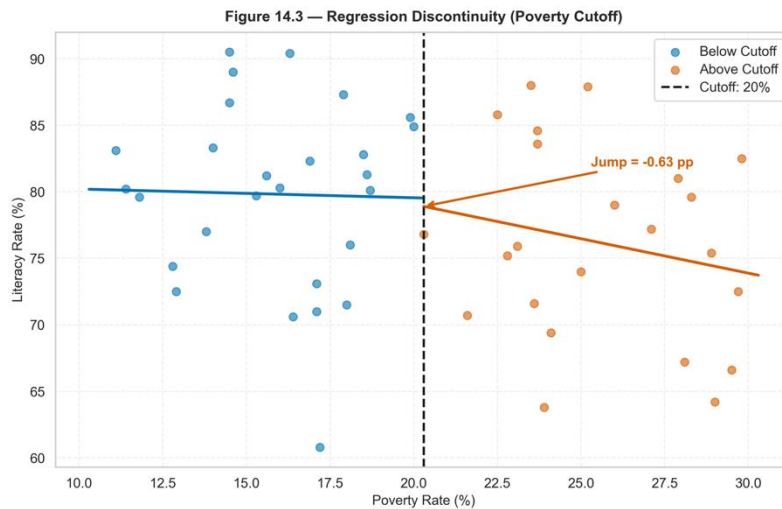


Fig. 4. Regression discontinuity at the 20% poverty-rate scholarship-eligibility cutoff.

V. DISCUSSION

This section interprets the pipeline’s diagnostic output, not a policy verdict. Three findings emerge from comparing the SHAP and causal stages side by side; Table IV summarizes the resulting two-gate decision structure.

Table IV. Predictive rank versus causal verdict: the two-gate decision structure

Variable	SHAP rank	Causal verdict	Two-gate outcome
Teacher Quality	1	Confirmed ($p < 0.01$)	Passes both gates
Digital Access	4	Confirmed ($p < 0.05$)	Passes both gates
Infrastructure	7	Not confirmed	Passes Gate 1, fails Gate 2
Toilet, Water Access	2, 3	Not tested (§III-C)	Queued for validation

First, cross-method convergence occurred for Teacher Quality and Digital Access: both were high-ranked by SHAP and causally confirmed by Double ML, which is the strongest form of evidence this pipeline can produce and is the pattern a real deployment should require before funding a lever.

Second, Infrastructure illustrates the central risk this paper is built to expose a variable that is intuitively actionable and ranks 7th by SHAP importance nonetheless fails to clear a causal-significance bar once confounders are orthogonalized, meaning a policy built solely on the SHAP ranking risks funding a lever this pipeline cannot show actually moves literacy.

Third, the null RDD and DiD results around the modeled poverty-based scholarship-eligibility design illustrate how the same pipeline can also flag a targeting mechanism not just a spending category as potentially mis calibrated relative to the variables identified as genuine causal levers.

Practically, this argues for a two-gate rule in any ML-informed public budgeting process: SHAP, or any comparable feature-attribution method, should be used only to generate a shortlist of candidate levers, never to rank-order a budget; and every shortlisted candidate should pass a confounder-robust causal check Double ML or Causal Forest, cross-validated where possible against a quasi-experimental design tied to an actual policy instrument before funds are committed against it.

Scope and limitations. The specific numerical estimates in Section IV demonstrate that the pipeline can discriminate a causally-defensible variable from a merely-predictive one; they are not presented as Nepal’s national treatment-effect parameters. The underlying panel is anchored to real, officially-reported provincial statistics but generated at district resolution through calibrated microsimulation [11] rather than drawn from raw individual- or school-level administrative records, so effect sizes should be replicated on disaggregated administrative microdata before they inform an actual budget cycle. The identifying assumptions discussed in Section III-D (confoundedness, overlap, SUTVA) also require formal diagnostic testing balance statistics, common-support plots, and placebo/pre-trend checks on that real microdata before deployment. We view this paper’s contribution as the validated architecture for doing so, transferable to any sector where a government is tempted to treat ML feature importance as a funding rationale.

VI. CONCLUSION

This paper validates, on a nationally-anchored education-equity dataset, a hybrid methodology in which explainable ML feature importance and causal treatment-effect estimation are deliberately kept as two separate, sequential gates rather than one. Coupling XGBoost and SHAP with Double ML and Causal Forests, and triangulating against quasi-experimental designs, identified Teacher Quality and Digital Access as the only causally-validated levers among the top SHAP-ranked candidates in this demonstration, and showed that the same pipeline can also surface a potentially mis calibrated targeting design.

We present the resulting “predict, then validate” architecture as a reproducible, sector-agnostic template. Its next test is a re-run on disaggregated administrative microdata not a wider claim about Nepal’s FY 2082/83 budget.

ACKNOWLEDGMENT

The author thanks the Central Bureau of Statistics, the Centre for Education and Human Resource Development, and the Ministry of Finance of Nepal for maintaining the public statistical releases on which the calibration of this study depends.

REFERENCES

- [1] Central Bureau of Statistics, Nepal Living Standards Survey 2022/23, National Statistics Office, Kathmandu, 2024.
- [2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [3] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in *Proc. 37th Int. Conf. on Machine Learning (ICML)*, 2020.
- [4] C. Molnar, G. König, J. Herbringer, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl, "General pitfalls of model-agnostic interpretation methods for machine learning models," in *xxAI — Beyond Explainable AI*, Springer, 2022, pp. 39–68.
- [5] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 2018.
- [6] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *The Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019.
- [7] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, 2015.
- [8] U. Fischer-Abaigar, C. Kern, and F. Kreuter, "The missing link: Allocation performance in causal machine learning," *arXiv:2407.10779*, 2024.
- [9] G. Carloni, A. Berti, and S. Colantonio, "The role of causality in explainable artificial intelligence," *WIREs Data Mining and Knowledge Discovery*, vol. 15, e70015, 2025.
- [10] G. Türkmen, "The review of studies on explainable artificial intelligence in educational research," *Journal of Educational Computing Research*, vol. 63, no. 2, pp. 277–310, 2025.
- [11] F. Bourguignon and A. Spadaro, "Microsimulation as a tool for evaluating redistribution policies," *Journal of Economic Inequality*, vol. 4, no. 1, pp. 77–106, 2006.
- [12] Centre for Education and Human Resource Development, *Flash Report I 2081/82*, Ministry of Education, Science and Technology, Kathmandu, 2025.
- [13] Ministry of Finance, *Red Book, Fiscal Year 2082/83*, Government of Nepal, Kathmandu, 2025.

Copyright & License:



© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.