UNDER PEER REVIEW
Policy Article

Harnessing Data Analytics to Bridge Socioeconomic Gap in Nepal's Education System

***Abstract*—This research aims to bridge the socioeconomic disparity gap in Nepal's education system through the integration of data analytics and explainable artificial intelligence (XAI) with foundational social theories. Despite progress in enrollment rates, systemic inequities in access, retention, and learning outcomes persist among marginalized communities defined by caste, gender, and geography. Current research and policy frameworks suffer from a critical disconnect: they rely on outdated statistical methods and fail to integrate contextual social theories with modern machine learning. To address this, we propose a mixed-methods design combining quantitative modeling using national datasets (EMIS, NLSS) with qualitative insights from stakeholders. Predictive models such as Random Forest and XGBoost will be trained and interpreted using SHAP (SHapley Additive exPlanations) to identify key drivers of educational disparity. Theoretical frameworks—Sen's Capability Approach and Bourdieu's Cultural Capital Theory—are computationally operationalized to contextualize findings. The study will generate policy-relevant tools, including a resource allocation framework, SHAP-based simulation dashboards, and early-warning indicators for student dropouts. This work contributes to academia by pioneering the integration of XAI with equity theories, supports policymakers with actionable, data-driven interventions, and empowers marginalized communities through evidence-based advocacy.**

key words**—Educational equity, data analytics, explainable AI, SHAP, machine learning, Nepal, policy modeling, mixed methods.**

I. Introduction

Nepal has achieved significant progress in education access, with a gross enrollment rate of 95% as of 2024 [7]. However, systemic disparities in quality, retention, and learning outcomes persist, particularly among marginalized populations defined by caste, gender, and rural geography [10]. These inequities challenge the nation's progress toward Sustainable Development Goal 4 (SDG 4), which calls for inclusive and equitable quality education for all.

Despite the availability of rich national datasets such as the Education Management Information System (EMIS) and the Nepal Living Standards Survey (NLSS), their potential for predictive and policy-aligned analytics remains underutilized. Over 90% of existing studies on Nepali education rely on basic statistical methods like OLS regression, which fail to capture non-linear dynamics or interaction effects [9]. Moreover, there is a notable absence of integration between advanced machine learning techniques and social science theories that could provide deeper contextual understanding.

This research proposes a novel, interdisciplinary approach that bridges this gap by combining explainable AI (XAI), spatial analysis, and sociological theory to model, interpret, and address educational disparities in Nepal. By integrating SHAP-based interpretability with Sen's Capability Approach and Bourdieu's Cultural Capital Theory, this work aims to deliver not only technical innovation but also socially grounded, policy-relevant insights.

II. Literature Review

A. Educational Disparities in Nepal and LMICs

Educational inequity remains a persistent challenge in Nepal. Bhattarai [10] highlights disparities in access and learning outcomes among marginalized communities defined by caste, gender, and rural geography. Similar trends are reported in other LMICs, where socioeconomic status and geographic location significantly influence educational attainment [11]. Most of these studies rely on descriptive statistics or linear models, which fail to capture complex interactions between socioeconomic variables.

B. Data Analytics and Machine Learning in Education

The application of predictive analytics and machine learning has gained traction in education. For a given dataset with features $X = \{x_1, x_2, \ldots, x_n\}$ representing student and household attributes, and a target variable Y representing dropout risk or academic performance, predictive models aim to learn a function:

$$\hat{Y} = f(X; \theta) \tag{1}$$

where f is the model (e.g., Random Forest, XGBoost) and θ represents model parameters. Ensemble models, such as Random Forest, aggregate multiple decision trees to reduce variance and improve accuracy:

$$\hat{Y}_{RF} = \frac{1}{T} \sum_{t=1}^{T} \hat{Y}^{(t)} \tag{2}$$

where T is the number of trees in the forest. XGBoost improves predictive performance via gradient boosting:

$$\hat{Y}_{XGB}^{(k)} = \hat{Y}_{XGB}^{(k-1)} + \eta \cdot h_k(X) \tag{3}$$

with learning rate η and base learner $h_k(X)$ at iteration k [19].

C. Explainable AI (XAI) for Policy and Decision-Making

Explainable AI techniques, such as SHAP, decompose model predictions into additive contributions of each feature x_i:

$$\hat{Y} = \phi_0 + \sum_{i=1}^{n} \phi_i \tag{4}$$

where ϕ_0 is the expected model output and ϕ_i is the SHAP value representing the contribution of feature x_i [19]. This allows policymakers to understand **which socioeconomic factors drive dropout risk**, enabling targeted interventions.

D. Comparative Analysis of Past Studies

Table I summarizes key studies in the context of educational equity, datasets, methods, and gaps:

TABLE I: Comparison of Past Studies on Educational Analytics in LMICs

Study	Dataset	Method	Focus	Gap
Bhattarai (2024) [10]	EMIS, NLSS	OLS Regression	Dropout risk	No non-linear modeling, limited interpretability
Baker (2014) [12]	US K-12 data	Decision Trees, Educational Data Mining	At-risk students	Black-box models, no social theory
Lundberg & Lee (2017) [19]	Synthetic, Education datasets	XGBoost + SHAP	Model interpretability	Not applied to LMICs, no policy integration
Global Partnership (2019) [11]	Multi-country LMIC data	Descriptive Statistics	Equity gaps	No predictive modeling or XAI

III. Research Gap

A. Theoretical–Methodological Divide

Current research in educational data analytics lacks integration between modern machine learning and social science theories. While models like Random Forest and XGBoost offer high predictive accuracy, their "black-box" nature limits interpretability for policymakers. Conversely, theories such as Sen's Capability Approach [16] and Bourdieu's Cultural Capital [15] provide rich frameworks for understanding inequality but lack computational operationalization in the Nepali context.

Furthermore, longitudinal analysis of how socioeconomic variables interact with policy reforms over time is absent. No study has applied causal machine learning techniques—such as Difference-in-Differences (DiD)—to evaluate the impact of education policies in Nepal, creating a disconnect between policy formulation and evidence-based evaluation.

B. Policy–Data Disconnect

Despite the existence of national data repositories, policymakers lack accessible tools for localized decision-making. There is no interactive "what-if" simulation framework to assess the impact of interventions such as teacher redistribution or scholarship targeting. Policies remain reactive, based on annual reports rather than predictive analytics, leading to delayed and inefficient responses.

C. Sen's Capability Approach and Quantitative Operationalization

Sen's Capability Approach frames educational equity as the freedom to achieve valued outcomes [16]–[18]. To integrate this with predictive modeling, capabilities can be operationalized using indices derived from observed variables.

Capability Score: For a student i in dimension d (e.g., literacy, numeracy, retention), the capability can be estimated as:

$$C_{i,d} = w_d \cdot f(x_{i,d}) \tag{5}$$

where:

- $x_{i,d}$ is the observed outcome or proxy (e.g., exam score, attendance) [18].
- $f(\cdot)$ is a normalization function to scale outcomes between 0 and 1 [17].
- w_d is the weight assigned to each dimension based on policy priority or expert judgment [18].

Overall Capability Index: Aggregating across multiple dimensions D:

$$CI_i = \sum_{d=1}^{D} C_{i,d} = \sum_{d=1}^{D} w_d \cdot f(x_{i,d}) \tag{6}$$

This index can serve as a **dependent variable or target** in predictive modeling, allowing Random Forest/XGBoost to predict capability deprivation risk.

Integration with SHAP: Applying SHAP to models predicting CI_i identifies **which socioeconomic and school-level features most influence capability deprivation**, making the analysis interpretable for policymakers [19].

IV. Proposed Solution

A. Mixed-Methods Design

This study employs a sequential mixed-methods research design, integrating quantitative modeling with qualitative inquiry to ensure both statistical rigor and contextual depth. The approach unfolds in three distinct phases:

1) **Quantitative Phase:** Leveraging 25 years of national education data (2000–2025) from the Education Management Information System (EMIS) and the Nepal Living Standards Survey (NLSS), we train interpretable machine learning models—specifically Random Forest and XGBoost—to predict student dropout risks. SHAP (SHapley Additive exPlanations) is applied post-hoc to interpret model outputs and identify the most influential socioeconomic and geographic drivers of educational disparity.
2) **Qualitative Phase:** To ground the quantitative findings in lived experience, we conduct semi-structured interviews with 42 key stakeholders, including teachers, parents, school administrators, and education policymakers. Thematic analysis of these interviews enables validation of model insights and uncovers nuanced barriers to access and retention that may not be captured in structured datasets.

3) **Integration Phase:** Findings from both streams are synthesized through participatory workshops involving education officials and community representatives. These sessions facilitate co-design of equitable policy interventions and allow for simulation-based testing of potential strategies before real-world implementation.

B. Policy-Relevant Outputs

The research is designed to generate actionable tools and frameworks that support evidence-based decision-making in education policy. Key deliverables include:

1) A **data-driven resource allocation framework** that prioritizes districts and wards with the highest predicted risk of dropout, enabling targeted deployment of teachers, scholarships, and infrastructure.
2) An interactive **SHAP-based simulation dashboard** that allows policymakers to explore the projected impact of various interventions—such as increasing female teacher representation or expanding transportation access—on educational equity outcomes.
3) **Early-warning indicators** for at-risk students, derived from predictive models and interpretable features, which can be integrated directly into Nepal's EMIS to enable proactive support mechanisms at the school level.

V. Research Objectives

1) **Analyze Education Disparities:** Investigate the influence of caste, income, and geography on access and outcomes using Sen's and Bourdieu's frameworks.
2) **Develop Predictive Models:** Train interpretable ML models to identify key drivers of disparity and generate policy simulations.
3) **Propose Policy Solutions:** Design equitable interventions and early-warning mechanisms.
4) **Design an Analytics Framework:** Propose a scalable M&E system and capacity-building programs for local stakeholders.

VI. Theoretical and Methodological Framework

A. Theoretical Foundations

1) **Capability Approach (Sen, 2000):** Educational access is framed as freedom to achieve valued outcomes. Barriers such as distance to school, resource scarcity, and discrimination limit "conversion factors."
2) **Cultural Capital Theory (Bourdieu, 2000):** Exam design favoring dominant languages, teacher bias, and unequal access to STEAM programs perpetuate exclusion.

TABLE II: Methodological Triangulation Framework

Phase	Analytical Focus	Policy Relevance
Quantitative	Predictive modeling, SHAP, DiD	Priority districts, structural inequities
Qualitative	Barrier analysis, stakeholder interviews	Root causes, implementation challenges
Integration	Scenario testing, co-design	Evidence-based solutions, roadmaps

VII. Data Description and Challenges

A. Data Sources

This study leverages national-level datasets to analyze educational disparities in Nepal:

- **Education Management Information System (EMIS):** Provides annual school-level statistics including enrollment, dropout rates, teacher-student ratios, infrastructure, and exam performance across districts.
- **Nepal Living Standards Survey (NLSS):** Household-level survey capturing socioeconomic indicators such as income, parental education, household size, geographic location, and access to basic amenities.

B. Variables and Criteria

Key variables are selected based on relevance to educational equity and predictive modeling:

- **Demographic:** Student age, gender, caste/ethnicity, geographic region.
- **Socioeconomic:** Household income, parental education, access to electricity and internet, occupation of guardians.
- **School-level:** Student-teacher ratio, number of teachers, distance to school, infrastructure index, exam scores.
- **Outcome:** Student dropout status or retention over academic years.

C. Data Challenges in Nepal

Nepal presents unique challenges in data acquisition and quality:

1) **Incomplete records:** Many schools have missing enrollment or performance data due to limited reporting capacity.
2) **Non-standardized data:** Variations in district-level reporting and inconsistent coding of variables.
3) **Access restrictions:** Government datasets often require special permissions and are not publicly available in raw form.
4) **Sparse longitudinal data:** Historical records are inconsistent, making it difficult to track student-level progress over time.

D. Data Preprocessing

To prepare the datasets for predictive modeling:

- Handle missing values using imputation techniques (mean/mode or predictive imputation).
- Encode categorical variables such as caste, gender, and district using one-hot encoding or label encoding.
- Normalize continuous variables for algorithms sensitive to scale (e.g., XGBoost feature scaling is optional but Random Forest handles unscaled features).
- Merge household-level and school-level data using student and school identifiers.

VIII. Expected Outcomes

A. Academic Contributions

1) First integration of SHAP-based ML with Sen's Capability Approach in Nepal.
2) Doctoral thesis and publications in Q1 journals on XAI for educational equity.

B. Policy Implementation Tools

1) Interactive dashboard with QGIS integration for geospatial prioritization.
2) Policy toolkit with implementation guidelines and teacher training modules.

C. Scalability

The framework is designed for replication in other low- and middle-income countries (LMICs) facing similar educational equity challenges.

IX. Preliminary Results / Conceptual Outputs

Although real data outputs are not yet available, Figure 1 presents a conceptual workflow of the research process.

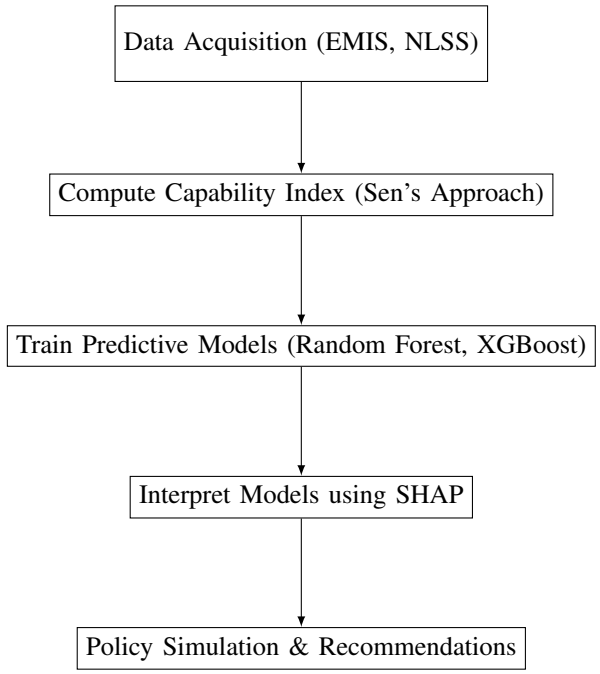

Fig. 1: Conceptual Workflow of the Study: Data → Capability Index → ML Models → SHAP → Policy Recommendations

X. Ethical Considerations

Ethical compliance is critical given the use of student and household data:

1) **Informed Consent:** All participants in interviews and surveys will provide informed consent in accordance with institutional and national guidelines [20].
2) **Data Privacy:** EMIS and NLSS data will be anonymized to protect identities, and access will follow Nepal's data governance regulations.
3) **Minimizing Harm:** Care will be taken to avoid stigmatization of marginalized groups when interpreting or publishing results.
4) **Ethical Approval:** The study will obtain approval from the Institutional Review Board (IRB) of IIMS College and collaborating institutions.

XI. Conclusion

This research bridges a critical gap in both academic and policy domains by integrating advanced data analytics with social theory to address educational inequity in Nepal. By combining SHAP-based interpretability, causal modeling, and stakeholder engagement, the study offers a novel, context-sensitive approach to policy design. The resulting tools and frameworks will empower policymakers, support marginalized communities, and contribute to the global discourse on equitable AI in education.

References

[1] R. S. Baker, "Educational data mining: An advance for intelligent systems in education," *IEEE Intelligent Systems*, vol. 29, no. 3, pp. 78–82, 2014.
[2] S. Bhattarai, "Returns to education in Nepal: An analysis of educational attainment," *Education Economics*, vol. 32, no. 2, pp. 145–167, 2024.
[3] P. Bourdieu, "Cultural capital and social inequality in modern society," *Theory and Society*, vol. 29, no. 5, pp. 567–582, 2000.
[4] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. SAGE Publications, 2017.
[5] Global Partnership for Education, "Nepal education sector analysis," 2019. [Online]. Available: https://www.globalpartnership.org
[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
[7] Ministry of Education, Nepal, "Education statistics of Nepal 2024," 2024. [Online]. Available: https://moe.gov.np
[8] A. Sen, *Development as Freedom*. Oxford University Press, 2000.
[9] TechAxis, "The growing demand for data scientists in Nepal," Mar. 15, 2024. [Online]. Available: https://www.techaxis.com.np/blog
[10] S. Bhattarai, "Returns to education in Nepal: An analysis of educational attainment," *Education Economics*, vol. 32, no. 2, pp. 145–167, 2024.
[11] Global Partnership for Education, "Nepal education sector analysis," 2019. [Online]. Available: https://www.globalpartnership.org
[12] R. S. Baker, "Educational data mining: An advance for intelligent systems in education," *IEEE Intelligent Systems*, vol. 29, no. 3, pp. 78–82, 2014.
[13] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
[14] A. Sen, *Development as Freedom*. Oxford University Press, 2000.
[15] P. Bourdieu, "Cultural capital and social inequality in modern society," *Theory and Society*, vol. 29, no. 5, pp. 567–582, 2000.
[16] A. Sen, *Development as Freedom*. Oxford University Press, 2000.
[17] I. Robeyns, "Selecting capabilities for quality of life measurement," *Social Indicators Research*, vol. 74, pp. 191–215, 2005.
[18] S. Alkire, "Dimensions of human development," *World Development*, vol. 30, no. 2, pp. 181–205, 2002.
[19] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
[20] UNESCO, "Ethics of conducting research on education and learning," UNESCO Policy Guidelines, 2015. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000232045