Machine learning models in the enhancement of PSE in high-dimensional socioeconomic data: a review

Gene Marck B. Catedrilla[1,2], Joey Aviles[1]
[1]Angeles University Foundation (AUF), Angeles, Philippines
[2]College of Computer Studies, Laguna State Polytechnic University, Los Baños, Philippines

Article Info

Article history:

Received Aug 12, 2025
Revised Dec 29, 2025
Accepted Jan 11, 2026

Keywords:

Ensemble model
High-dimensional data
Machine learning models
Socioeconomic evaluation
Systematic review

ABSTRACT

This study reviews the use of machine learning (ML) techniques to improve propensity score (PS) estimation in high-dimensional socioeconomic data. Traditional logistic regression (LR) often performs poorly under nonlinear and complex covariate structures, leading to bias and model misspecification. Across the reviewed studies, ensemble methods such as random forests (RF) and gradient boosting, and deep learning models consistently achieved better covariate balance, lower bias, and greater flexibility than conventional approaches, while classification-based methods improved performance in imbalanced datasets. The review also highlights practical considerations, including calibration, transparent reporting, and integration with doubly robust estimators to strengthen causal inference. The findings show that ML-based propensity score estimation (PSE) can substantially enhance the validity and reliability of socioeconomic evaluations, provided that its implementation is carefully guided by appropriate expertise and best-practice standards.

Corresponding Author:

Gene Marck B. Catedrilla
Angeles University Foundation (AUF)
Angeles, Philippines
Email: 24-2093-423@auf.edu.ph or catedrilla.genemarck@lspu.edu.ph

1. INTRODUCTION

Impact evaluations assess the causal effects of interventions by comparing observed outcomes with a counterfactual scenario-what would have occurred in the absence of the intervention. The difference between these outcomes represents the intervention's impact, which may be measured at multiple levels and is not limited to long-term effects, contrary to common assumptions. Nonetheless, impact evaluation remains the primary approach capable of providing credible evidence of long-term outcomes. When properly designed and executed, impact evaluations can inform policy decisions, shape public opinion, and improve program operations [1].

Impact evaluation is embedded within the broader shift toward evidence-based policymaking, which emphasizes outcomes over inputs in public policy and organizational decision-making. This results-oriented approach supports the monitoring of national and international targets and strengthens accountability, budget allocation, and program design [1]. By providing credible evidence of performance, impact evaluations determine whether programs have achieved or are achieving their intended objectives and quantify improvements in beneficiaries' quality of life attributable to the intervention. However, a persistent challenge in socioeconomic evaluation is the lack of high-quality baseline data, which undermines the reliability of causal estimates [2].

Propensity score (PS) methods are widely applied in socioeconomic evaluation to address incomplete or missing baseline data. By creating balanced comparison groups, PS methods enable more credible causal inference even when baseline information is limited [3]. PS also reduce multidimensional covariate information into a single scalar measure, facilitating covariate balance when full baseline data are unavailable [4]. As noted by Austin *et al.* [5], PS approaches are particularly valuable in large-scale socioeconomic evaluations where consistent baseline data collection is difficult. Cham and West [6] further demonstrated that machine learning (ML) models can improve propensity score estimation (PSE) under missing data conditions by capturing nonlinear and nonadditive relationships that traditional logistic regression (LR) fails to model effectively. These advantages are especially important in high-dimensional socioeconomic datasets [7]. Ensemble methods and deep neural networks (DNNs) have likewise shown strong performance in handling missing data and complex covariate structures [7], [8].

Recent advances in causal inference extend beyond conventional PSE through the integration of ML with frameworks such as double/debiased machine learning (DML), meta-learning approaches (e.g., T-learners, S-learners, X-learners), and heterogeneous treatment effect (HTE) models. These methods enable estimation not only of average treatment effects but also of subgroup-specific impacts, which is critical for public policy and development programs [9]-[11]. Situating ML-based PSE within this broader causal inference landscape highlights its relevance for addressing complex real-world evaluation challenges.

Despite the growing adoption of ML-based PSE, existing studies remain fragmented, with limited synthesis across socioeconomic domains, insufficient comparison across ML model families, and inadequate discussion of practical issues such as calibration, fairness, and interpretability. This gap constrains researchers' ability to select appropriate ML models for high-dimensional socioeconomic evaluation and motivates the need for a comprehensive review. This paper aims to review literature on the viability of ML models in predicting and estimating PSs. Specifically, this literature review will focus on the following:
1. Explore the development of ML model applications in PS analysis.
2. Highlight the practical implications of ML in predicting PSs for researchers and accreditors in socioeconomic evaluation.
3. Provide a summary of how ML models and PS can improve the effectiveness of socioeconomic evaluation.

2. LITERATURE REVIEW

2.1. Propensity score and its usage

PS methodology provides a framework for achieving covariate balance in observational studies by adjusting for systematic differences between treated and control groups. Estimated PSs are commonly applied through stratification into subclasses, matching treated and control units with similar scores, or inverse probability of treatment weighting (IPTW), each aiming to approximate randomized experimental conditions. To address limitations of traditional parametric models used in PSE, ML algorithms have been increasingly adopted to improve flexibility and robustness in modeling complex treatment assignment mechanisms [12].

PS methods are widely used across disciplines to support causal inference from observational data. In healthcare and epidemiology, they are applied to evaluate treatment effectiveness and safety using real-world data sources such as electronic health records and insurance claims [13]. In social science research, PSs are used to assess the impacts of social programs, educational interventions, and workforce training initiatives by balancing baseline characteristics between participants and non-participants [14]. Economists employ PS analysis to estimate the causal effects of policy interventions, including unemployment benefits, minimum wage policies, and development programs, on labor and income outcomes [15]. Beyond these areas, PS techniques are increasingly applied in marketing analytics to evaluate advertising and loyalty programs [16], as well as in environmental science, public policy evaluation, and other quasi-experimental research settings.

2.2. Various common and conventional methods on getting propensity score

The traditional way of estimating PSs mostly uses statistical models, with LR being the most common method. LR models the log-odds of treatment assignment as a linear function of the observed variables. Once we estimate these PSs, we apply them through several established methods. Stratification, or subclassification, involves dividing the study population into groups, often quintiles, based on the estimated PSs. We then compare outcomes between treated and control units within each group. The overall treatment effect is usually calculated as a weighted average across these groups [17].

Matching techniques pair each treated unit with one or more control units that have very similar PSs, such as nearest-neighbor matching and caliper matching. This creates a matched sample where covariate distributions are balanced. IPTW assigns weights to everyone based on their PS. Treated units receive a weight of *1/e(X)*, while control units get a weight of *1/(1-e(X))*. These weights create a pseudo-population where the distribution of covariates does not depend on treatment assignment. This allows for estimating the average treatment effect in the population (ATE) or the average treatment effect on the treated (ATT) through weighted analyses. Although these methods are well-known and widely used in statistical software, they rely heavily on correctly specifying the LR model. Misspecification, such as omitting relevant confounders or failing to include necessary interaction or non-linear terms, can result in residual confounding and biased effect estimates. Additionally, conventional LR finds it challenging to handle high-dimensional covariate data [18].

2.3. Machine learning and propensity score

PSs can be estimated using ML algorithms to address limitations of traditional LR, particularly under nonlinear, nonadditive, and high-dimensional covariate structures. ML-based approaches offer greater flexibility by automatically capturing complex relationships and interactions without requiring explicit model specification, and they are generally more effective when many potential confounders are present. However, increased model flexibility also introduces risks, as highly complex algorithms—such as DNNs and flexible tree ensembles—may overfit treatment assignment models, leading to suboptimal covariate balance and biased causal estimates. Consequently, careful model implementation, tuning, and validation are essential when applying ML to PSE [19].

Tree-based ensemble methods, including gradient boosting machines (GBM) and random forests (RF), are among the most commonly used ML approaches for PSE. These methods demonstrate strong performance by accommodating nonlinearities and interactions while maintaining robustness through aggregation across multiple decision trees [20]. Penalized regression methods, such as Lasso and Ridge regression, extend LR by introducing regularization to improve stability in high-dimensional settings. Lasso performs variable selection by shrinking some coefficients to zero, enhancing model parsimony, while Ridge regression stabilizes estimates by shrinking coefficients without exclusion, which is particularly beneficial under multicollinearity [21].

Neural networks offer high representational capacity for modeling complex treatment–covariate relationships but are less frequently applied in PSE due to their sensitivity to sample size, tuning requirements, and risk of overfitting [22]. Their limited interpretability and reliance on extensive hyperparameter optimization further complicate validation in applied socioeconomic studies [23].

Ensemble learning approaches that combine multiple algorithms through cross-validation—such as super learner frameworks integrating LR, GBM, Lasso, and support vector machines—provide a flexible and robust alternative to single-model estimation. These methods often outperform individual learners, particularly in high-dimensional and heterogeneous datasets, by balancing predictive performance with improved covariate balance [19].

2.4. Other considerations for ML-based propensity scoring

Hyperparameter tuning—such as adjusting learning rate, tree depth, and regularization strength—is critical when applying ML to PSE. Techniques such as k-fold cross-validation help control overfitting and support covariate balance; however, optimizing treatment assignment prediction accuracy alone is insufficient and may even be detrimental to causal validity [24].

Regardless of the ML algorithm used, post-estimation assessment of covariate balance between treated and control groups remains essential after applying PSs through matching, weighting, or stratification. Standard diagnostics include standardized mean differences (SMD), with values below 0.1 typically indicating acceptable balance, variance ratios, and visual tools such as love plots. Inadequate balance indicates failure of the PS model—irrespective of its complexity—and necessitates model or method refinement [5]. Combining ML-based PS methods with a separate outcome regression model enables doubly robust estimation, ensuring consistent causal estimates if either the PS model or the outcome model is correctly specified [25].

Recent literature further emphasizes transparency and fairness in ML-based PSE, particularly in socioeconomic applications. Complex ML models may obscure treatment assignment mechanisms, reducing interpretability and stakeholder trust. Explainable artificial intelligence (XAI) tools, including SHAP values, LIME, and interpretable tree-based models, help clarify model behavior and enhance transparency in policy evaluation [26], [27]. Fairness-aware ML approaches additionally support the identification and mitigation of biased treatment assignment across demographic subgroups, reducing the risk that PSE reinforces existing inequities [28]. These considerations are increasingly critical as ML-based methods gain adoption in government and development program evaluations.

3. METHODS

This review paper focuses on the literature review process. This process involves selecting and quantifying existing studies that apply ML models in PS analysis. Therefore, different tools for searching scholarly databases are the main materials used.

3.1. Scope and focus, and search strategy

This literature review examines the application of ML models in PS analysis, with a specific focus on addressing missing baseline data in high-dimensional socioeconomic datasets. The review synthesizes peer-reviewed journal articles, conference papers, and related academic studies that evaluate ML–based approaches for PSE, methodological implementation, and performance assessment.

Relevant literature was retrieved from major academic databases, including SCOPUS, Google Scholar, IEEE Xplore, SpringerLink, ScienceDirect, and the ACM Digital Library. The Publish or Perish (PoP) tool was additionally used to identify supplementary studies. Litmaps was employed to assess source relevance and citation connectivity. Search terms were combined systematically to refine retrieval, as summarized in Table 1.

Table 1. Keywords for searching related literature

Category	Keywords
PS analysis	PS, PS analysis
ML models	ML, ML models, ML algorithms
Socioeconomic assessment	Socioeconomic, socioeconomic assessment, socioeconomic evaluation
High-dimensional data	High-dimension data, high dimensional data

3.2. Inclusion and exclusion criteria

The review prioritizes studies based on relevance and methodological quality. Publications were primarily restricted to those released within the last ten years; however, the time frame was extended to up to fifteen years when necessary to ensure sufficient coverage of relevant work and maintain generalizability. Only peer-reviewed journal articles and reputable conference proceedings published in English were included, while non-peer-reviewed sources such as editorials, blog posts, and preprints were excluded.

Eligible studies explicitly examined the use of ML algorithms—including tree ensembles, penalized regression, super learner frameworks, and neural networks—for PS analysis. This included applications involving PSE, weighting, or matching under both simple and high-dimensional covariate structures. Studies applying ML to alternative causal inference methods without a primary focus on propensity modeling, as well as those using PSs in less complex evaluation settings, were excluded.

3.3. Conceptual framework

This review drew from major scholarly databases, including SCOPUS, Google Scholar, IEEE Xplore, SpringerLink, ScienceDirect, and the ACM Digital Library. The study selection process followed the PRISMA framework, as illustrated in the PRISMA flowchart adapted from [23] as shown in Figure 1.

The initial search identified 1,245 records from academic databases and an additional 95 records from grey literature and reference lists. After removing 180 duplicates, 1,160 unique records remained for title and abstract screening. Of these, 960 records were excluded for failing to meet the review objectives, resulting in 200 full-text articles assessed for eligibility. Following full-text evaluation, 85 articles were excluded for not applying ML techniques, 35 for lacking a direct focus on PS analysis, and 20 for non-socioeconomic applications. The final qualitative synthesis therefore included 60 studies.

3.4. Reporting the review

The findings were organized in accordance with the stated research objectives. The review process followed the PRISMA framework to ensure methodological transparency and reproducibility, thereby strengthening the credibility of the selected literature. Study synthesis involved systematic evaluation of relevance, methodological approaches, and applications of ML in PS analysis. The review emphasized identifying methodological trends, commonly used algorithms, and performance patterns across datasets of varying dimensionality, with the objective of highlighting recent advances, methodological gaps, and directions for future research at the intersection of ML and causal inference.

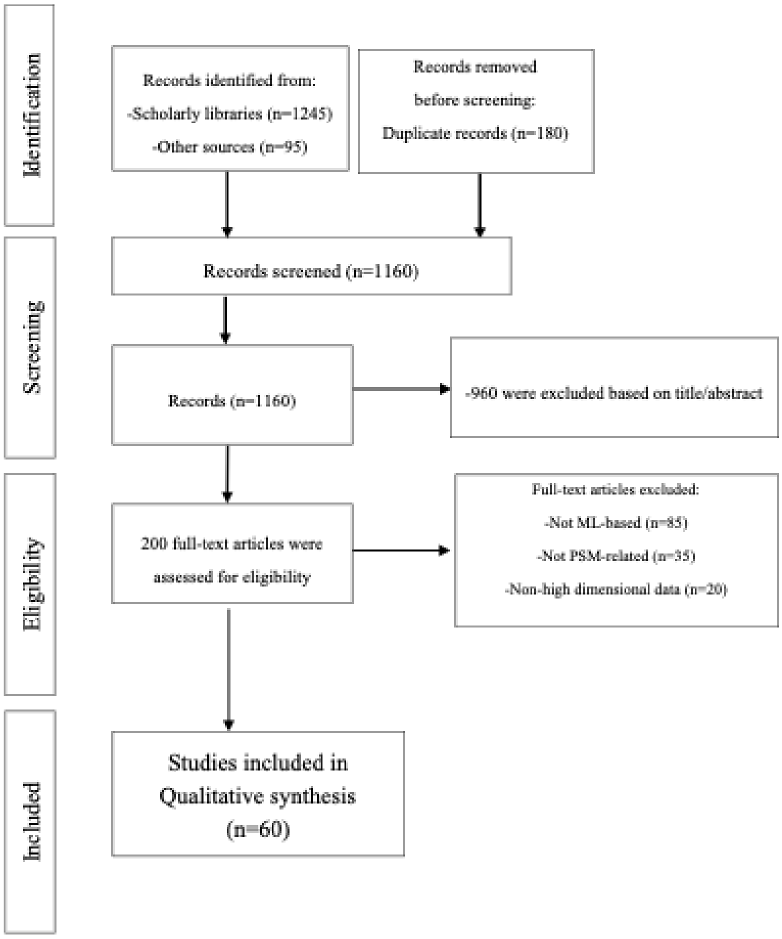

Figure 1. Literature review process (PRISMA flowchart; [29])

4. RESULTS AND DISCUSSION

In evaluating the effectiveness of ML models for predicting and estimating PSs, this review brings together 60 research papers from various digital libraries. It shows that no single source was preferred. The discussion follows the stated objectives to maintain coherence and relevance.

4.1. Application of ML models in PSE

The review indicates that LR remains one of the most commonly used approaches for estimating PSs; however, its widespread application has been accompanied by increased model misspecification, which adversely affects the accuracy of treatment probability estimates [30]. To address these limitations, more flexible ML–based methods have been increasingly adopted. Across the reviewed studies, ensemble learning algorithms and neural networks demonstrated superior performance in achieving covariate balance and reducing bias, particularly in high-dimensional socioeconomic datasets [31]. Ensemble methods—including gradient boosted trees, RFs, and bagged trees—consistently outperformed LR by improving covariate balance, reducing bias, and maintaining valid confidence intervals, especially under nonlinear covariate structures [31]. DNN further showed strong capability in managing complex high-dimensional PSE, often surpassing both LR and other ML approaches in predictive accuracy and stability [32].

Classification-based approaches, particularly classification tree analysis (CTA), also emerged as effective alternatives for PSE. CTA demonstrated improved accuracy over LR in settings characterized by imbalanced covariates, owing to its ability to capture nonadditive effects and variable interactions [33].

4.2. Practical implications of ML in predicting propensity score for socioeconomic evaluation

ML models demonstrate clear advantages over traditional LR in predicting PSs within complex socioeconomic datasets, particularly when modeling nonlinear relationships and interactions commonly observed in real-world data [34]. In multilevel observational settings, nonparametric ML methods have also been shown to outperform parametric LR approaches [35]. Across the reviewed studies, ML-based PSE consistently achieved improved covariate balance, reduced bias, and more stable confidence intervals, especially under conditions of nonlinearity and non-additivity [7], [12], [19].

Accurate interpretation of treatment effects is central to socioeconomic evaluation. Model misspecification—frequently encountered in LR—can distort causal estimates, whereas ML methods flexibly capture complex treatment–covariate relationships, thereby reducing bias. With appropriate hyperparameter tuning, ML-based approaches improve estimation of treatment probabilities and sample-level effects [7], [24]. Moreover, integrating ML-based PSE with doubly robust methods, which combine propensity modeling with independent outcome regression, provides additional protection against model misspecification and enhances the reliability of causal inference [25]. Despite their methodological advantages, many empirical studies provide limited reporting on model diagnostics, performance metrics, and hyperparameter tuning, underscoring the need for standardized best practices and evaluation frameworks in applied socioeconomic research [35]. While ML improves the validity and reliability of PSE in high-dimensional settings, its effectiveness depends on careful implementation, consistent methodology, and transparent reporting [36].

Computational feasibility is also a critical consideration in applied contexts. Socioeconomic institutions often face resource constraints, and certain ML approaches—particularly deep learning and large ensemble models—require substantial computational resources and tuning effort. Consequently, model selection should balance predictive performance with computational cost, implementation complexity, and available expertise [12], [37], especially in development agencies and public-sector evaluations.

4.3. Key findings and synthesis

ML-based PSE provides substantial methodological and practical advantages over conventional LR, particularly in high-dimensional and nonlinear socioeconomic datasets. Across the reviewed studies, ML approaches more effectively capture complex treatment–covariate relationships, reduce model misspecification, and achieve improved covariate balance. Ensemble methods, including RFs, GBMs, and bagged trees, consistently outperform traditional parametric models in predictive accuracy and bias reduction [31], [34], [38], [39]. DNNs demonstrate strong performance in highly complex and multivariate settings, highlighting their adaptability to large and heterogeneous datasets and their capacity to model nonadditive effects and high-order interactions [32], [39], [40].

Integrating ML-based PSE with doubly robust estimation methods further strengthens causal inference by providing protection against misspecification of either the treatment or outcome model [22], [28]. In multilevel and hierarchical observational settings, nonparametric ML approaches outperform standard LR in achieving covariate balance and reducing bias, underscoring their value for complex socioeconomic evaluations [35]. Causal tree–based algorithms are particularly effective in settings with severe covariate imbalance, a common feature of socioeconomic data, due to their ability to capture nonlinearities and heterogeneous assignment mechanisms [33].

Collectively, these findings reinforce the consensus that ML-based PSE offers a more flexible, accurate, and robust foundation for treatment assignment modeling than traditional approaches [38], [39]. Beyond methodological performance, ML-based PSE supports equity-focused evaluation through integration with heterogeneous treatment effect models, enabling identification of differential intervention impacts across subgroups defined by gender, socioeconomic status, or geographic context [40]–[43]. Effective application of these methods nevertheless requires careful implementation, including hyperparameter tuning, model calibration, and rigorous diagnostic assessment. Transparent reporting remains critical in high-dimensional settings to ensure interpretability, reproducibility, and policy relevance, as emphasized in recent methodological guidance [44]–[47]. When applied with appropriate methodological and ethical safeguards, ML-based PSE enhances the rigor, credibility, and precision of socioeconomic evaluations [3], [12], [48].

This review contributes by synthesizing ML-PSE evidence across public health, economics, education, and social policy. Unlike prior work focused on individual algorithms, it compares performance patterns across multiple ML families and socioeconomic contexts, extending earlier analyses by Cannas and Arpino [34], Tu [31], and Guzman-Alvarez *et al.* [7]. It further integrates emerging perspectives on fairness-aware ML [26], explainable artificial intelligence [27], and multilevel modeling [35], providing clearer guidance on when and how ML-based methods outperform traditional PS approaches in policy-relevant settings [38], [49], [50]. The performance analysis of models used in propensity estimation is summarized in Table 2.

Table 2. Analysis of model performance used in propensity estimation

Author/s	Best model	Performance	Remarks
Tu [31]	Gradient boosting (GBM)	Lowest MSE across all simulation scenarios	GBM consistently outperformed RF, bagging, and multinomial LR in generalized PSE.
Cannas and Arpino [34]	RF (PSW)	Best ASAM (covariate balance); top bias reduction in PSW	RF produced strongest overall balance; NN also strong but slightly below RF.
Ferri-García and Rueda [28]	RF (large sample) / LR (small sample)	Lowest MSE in most conditions (RF)	RF removed most bias as volunteer sample size increased; GBM second-best in large samples.
Greene *et al.* [20]	ML-based GPS (CDF method)	Bias = –0.045 to 0.028	Very low absolute bias; excellent stratification quality for ordinal exposures.
Šinkovec *et al.* [21]	Ridge LR (tuned)	Lowest RMSE among compared methods	Tuning improved stability and reduced estimation error in small/sparse samples.
Zou *et al.* [8]	Kernel ML (proposed method)	ATE mean ≈ 0.500, CI coverage = 95.0%	Most accurate and stable ATE estimates; far better coverage than RF or LASSO.
Ferri-García and Rueda [28]	GBM (with all predictors)	Lowest MSE for large samples	GBM achieved second-best MSE overall and strongest when many predictors used.
Guo *et al.* [37]	DNNs	Most stable PS predictions	DNNs outperform LR and kernel methods in high-dimensional nonlinear data.
Salditt and Nestler [35]	BART-RE (super learner)	SL weight = 0.47–0.60 (highest)	BART-RE consistently dominates SL; indicates best performance in multilevel settings.

5. CONCLUSION AND RECOMMENDATIONS

This review synthesized evidence from 60 studies applying ML-PSE in socioeconomic evaluation. Across diverse empirical and simulated contexts, the findings consistently indicate that ML-PSE provides substantial methodological advantages over traditional logistic regression, particularly when data exhibit nonlinearity, high dimensionality, and incomplete baseline information. Ensemble learning approaches, including RF, gradient boosting, and bagged trees, repeatedly demonstrated superior performance in achieving covariate balance, reducing bias, and improving predictive accuracy. DNNs further showed strong capacity to model complex, nonadditive relationships and frequently outperformed conventional methods in challenging socioeconomic settings, underscoring the potential of flexible learning architectures for causal adjustment in complex policy data.

Despite these advantages, the review highlights that successful implementation of ML-PSE depends critically on careful methodological practice. Appropriate hyperparameter tuning, model calibration, diagnostic assessment, and transparent reporting are essential to ensure robustness and credibility of results. Interpretability remains a key challenge, particularly for highly complex models such as DNNs; however, advances in explainable artificial intelligence and fairness-aware ML provide promising pathways to address transparency and accountability concerns. These considerations are especially salient in public policy and socioeconomic research, where equity, trust, and interpretability are integral to decision-making. Overall, gradient boosting methods, RF, DNNs, and Bayesian additive regression trees emerged as the most reliable approaches for improving bias reduction, covariate balance, and coverage probabilities in high-dimensional socioeconomic data, supporting the use of ML-PSE as a robust alternative to traditional methods under complex data-generating conditions.

The review also identifies important avenues for future research. Greater empirical validation using real-world socioeconomic datasets is needed, as much of the existing evidence remains simulation-based. Integrating ML-PSE with heterogeneous treatment effect modeling frameworks, such as causal forests and meta-learners, offers significant potential to uncover differential impacts across population subgroups defined by gender, income, or geography, thereby supporting more equitable and targeted policy design. Further development of fairness-aware PS methods is warranted to mitigate algorithmic bias in treatment assignment, alongside systematic evaluation of computational efficiency and scalability to inform adoption in resource-constrained institutional settings. Finally, the establishment of standardized reporting guidelines and best-practice frameworks will be essential to promote transparency, reproducibility, and responsible use of ML-PSE as these methods continue to gain prominence in socioeconomic evaluation.

ACKNOWLEDGMENTS

The authors would like to extend their gratitude to the Angeles University Foundation, and Laguna State Polytechnic University for the support in the conducting this study. Center for Lakes Sustainable Development (CLSD) and Bureau of Fish and Aquatic Resources (BFAR) for allowing the authors to use their socioeconomic data in the next phase of this project. Lastly, the Department of Science and Technology Science Education Institute (DOST-SEI) for the financial report.

FUNDING INFORMATION

The authors would like to acknowledge the DOST-SEI for the research grant for the whole conduct of the project.

AUTHOR CONTRIBUTIONS STATEMENT

The task in completing this research is equally divided among the authors.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Gene Marck B. Catedrilla	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓
Joey Aviles		✓		✓	✓			✓		✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

This review is the first phase of this project which does not involved categorical or numerical dataset. Hence, all the data used in this paper is available in the doi section of the references.

REFERENCES

[1] H. White and D. A. Raitzer, “Impact evaluation of development interventions: a practical guide,” Manila, Philippines, Dec. 2017. doi: 10.22617/TCS179188-2.
[2] J. Gao, Y. C. Zhang, and T. Zhou, “Computational socioeconomics,” *Physics Reports*, vol. 817, pp. 1–104, Jul. 2019, doi: 10.1016/j.physrep.2019.05.002.
[3] K. Narita, J. D. Tena, and C. Detotto, “Causal inference with observational data: a tutorial on propensity score analysis,” *Leadership Quarterly*, vol. 34, no. 3, p. 101678, Jun. 2023, doi: 10.1016/j.leaqua.2023.101678.
[4] P. R. Rosenbaum and D. B. Rubin, “Propensity scores in the design of observational studies for causal effects,” *Biometrika*, vol. 110, no. 1, pp. 1–13, Feb. 2023, doi: 10.1093/biomet/asac054.
[5] P. C. Austin, N. Jembere, and M. Chiu, “Propensity score matching and complex surveys,” *Statistical Methods in Medical Research*, vol. 27, no. 4, pp. 1240–1257, Apr. 2018, doi: 10.1177/0962280216658920.
[6] H. Cham and S. G. West, “Propensity score analysis with missing data,” *Psychological Methods*, vol. 21, no. 3, pp. 427–445, Sep. 2016, doi: 10.1037/met0000076.
[7] A. Guzman-Alvarez, X. Qin, and P. W. Scott, “Deep neural networks for propensity score estimation,” *Multivariate Behavioral Research*, vol. 57, no. 1, pp. 164–165, Jan. 2022, doi: 10.1080/00273171.2021.2011695.
[8] B. Zou, X. Mi, P. J. Tighe, G. G. Koch, and F. Zou, “On kernel machine learning for propensity score estimation under complex confounding structures,” *Pharmaceutical Statistics*, vol. 20, no. 4, pp. 752–764, Jul. 2021, doi: 10.1002/pst.2105.
[9] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, “Metalearners for estimating heterogeneous treatment effects using machine learning,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, no. 10, pp. 4156–4165, Mar. 2019, doi: 10.1073/pnas.1804597116.
[10] Z. Sadique, R. Grieve, K. Diaz-Ordaz, P. Mouncey, F. Lamontagne, and S. O’Neill, “A machine-learning approach for estimating subgroup- and individual-level treatment effects: an illustration using the 65 trial,” *Medical Decision Making*, vol. 42, no. 7, pp. 923–936, Oct. 2022, doi: 10.1177/0272989X221100717.
[11] D. Kabata and M. Shintani, “On propensity score misspecification in double/debiased machine learning for causal inference: ensemble and stratified approaches,” *Communications in Statistics: Simulation and Computation*, vol. 54, no. 5, pp. 1283–1293, May 2025, doi: 10.1080/03610918.2023.2279022.
[12] Y. Shang, Y. H. Chiu, and L. Kong, “Robust propensity score estimation via loss function calibration,” *Statistical Methods in Medical Research*, vol. 34, no. 3, pp. 457–472, Mar. 2025, doi: 10.1177/09622802241308709.
[13] A. Dang, “Real-world evidence: a primer,” *Pharmaceutical Medicine*, vol. 37, no. 1, pp. 25–36, Jan. 2023, doi: 10.1007/s40290-022-00456-6.
[14] S. I. Kumbhar and C. M. Jagtap, “Impact assessment of training and education programmes on the workers quality of work life and quality of life,” *Scholarly Research Journal for Humanity Science and English Language*, vol. 9, no. 47, pp. 11485–11494, Oct. 2021, doi: 10.21922/srjhsel.v9i47.7688.
[15] M. Gibson, W. Hearty, and P. Craig, “The public health effects of interventions similar to basic income: a scoping review,” *The Lancet Public Health*, vol. 5, no. 3, pp. e165–e176, Mar. 2020, doi: 10.1016/S2468-2667(20)30005-0.
[16] A. Lemmens and S. Gupta, “Managing churn to maximize profits,” *Marketing Science*, vol. 39, no. 5, pp. 956–973, Sep. 2020, doi: 10.1287/mksc.2020.1229.

[17] P. C. Austin, "Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures," *Statistical Methods in Medical Research*, vol. 28, no. 5, pp. 1365–1377, May 2019, doi: 10.1177/0962280218756159.

[18] R. M. James and M. A. Hernán, "Causal Inference: What If," *Causal Inference - What If*, p. 302, 2020, [Online]. Available: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

[19] R. Pirracchio, M. L. Petersen, and M. Van Der Laan, "Improving propensity score estimators' robustness to model misspecification using Super Learner," *American Journal of Epidemiology*, vol. 181, no. 2, pp. 108–119, Jan. 2015, doi: 10.1093/aje/kwu253.

[20] T. J. Greene, S. M. DeSantis, D. W. Brown, A. V. Wilkinson, and M. D. Swartz, "A machine learning compatible method for ordinal propensity score stratification and matching," *Statistics in Medicine*, vol. 40, no. 6, pp. 1383–1399, Mar. 2021, doi: 10.1002/sim.8846.

[21] H. Šinkovec, G. Heinze, R. Blagus, and A. Geroldinger, "To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets," *BMC Medical Research Methodology*, vol. 21, no. 1, p. 199, Dec. 2021, doi: 10.1186/s12874-021-01374-y.

[22] O. Clivio, F. Falck, B. Lehmann, G. Deligiannidis, and C. Holmes, "Neural score matching for high-dimensional causal inference," *Proceedings of Machine Learning Research*, vol. 151, pp. 7076–7110, 2022.

[23] Z. K. Collier and W. L. Leite, "A tutorial on artificial neural networks in propensity score analysis," *Journal of Experimental Education*, vol. 90, no. 4, pp. 1003–1020, Oct. 2022, doi: 10.1080/00220973.2020.1854158.

[24] W. Zhao and S. Livingston, "Allocation predictability of individual assignments in restricted randomization designs for two-arm equal allocation trials," *Statistics in Medicine*, vol. 44, no. 3–4, Feb. 2025, doi: 10.1002/sim.10343.

[25] D. Howdon, W. van den Hout, Y. van der Linden, and K. Spencer, "Replacing performance status with a simple patient-reported outcome in palliative radiotherapy prognostic modelling," *Clinical and Translational Radiation Oncology*, vol. 37, pp. 137–144, Nov. 2022, doi: 10.1016/j.ctro.2022.09.008.

[26] K. T. Rodolfa, H. Lamba, and R. Ghani, "Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy," *Nature Machine Intelligence*, vol. 3, no. 10, pp. 896–904, Oct. 2021, doi: 10.1038/s42256-021-00396-x.

[27] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, Jan. 2024, doi: 10.1007/s12559-023-10179-8.

[28] R. Ferri-García and M. Del Mar Rueda, "Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys," *PLoS ONE*, vol. 15, no. 4, p. e0231500, Apr. 2020, doi: 10.1371/journal.pone.0231500.

[29] V. Truong, "Natural language processing in advertising - a systematic literature review," in *Proceedings - 2022 5th Asia Conference on Machine Learning and Computing, ACMLC 2022*, Dec. 2022, pp. 89–98, doi: 10.1109/ACMLC58173.2022.00024.

[30] Z. Tan, "Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data," *Biometrika*, vol. 107, no. 1, pp. 137–158, Mar. 2020, doi: 10.1093/biomet/asz059.

[31] C. Tu, "Comparison of various machine learning algorithms for estimating generalized propensity score," *Journal of Statistical Computation and Simulation*, vol. 89, no. 4, pp. 708–719, Mar. 2019, doi: 10.1080/00949655.2019.1571059.

[32] M. Peng, Y. Li, C. Wu, and L. Li, "A deep learning approach to nonparametric propensity score estimation with optimized covariate balance," *ArXiv*, 2024, [Online]. Available: http://arxiv.org/abs/2404.04794.

[33] A. Linden and P. R. Yarnold, "Using classification tree analysis to generate propensity score weights," *Journal of Evaluation in Clinical Practice*, vol. 23, no. 4, pp. 703–712, Aug. 2017, doi: 10.1111/jep.12744.

[34] M. Cannas and B. Arpino, "A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting," *Biometrical Journal*, vol. 61, no. 4, pp. 1049–1072, Jul. 2019, doi: 10.1002/bimj.201800132.

[35] M. Salditt and S. Nestler, "Parametric and nonparametric propensity score estimation in multilevel observational studies," *Statistics in Medicine*, vol. 42, no. 23, pp. 4147–4176, Oct. 2023, doi: 10.1002/sim.9852.

[36] M. E. Karim, "High-dimensional propensity score and its machine learning extensions in residual confounding control," *American Statistician*, vol. 79, no. 1, pp. 72–90, Jan. 2025, doi: 10.1080/00031305.2024.2368794.

[37] Y. Guo, V. Y. Strauss, M. Català, A. M. Jödicke, S. Khalid, and D. Prieto-Alhambra, "Machine learning methods for propensity and disease risk score estimation in high-dimensional data: a plasmode simulation and real-world data cohort analysis," *Frontiers in Pharmacology*, vol. 15, Oct. 2024, doi: 10.3389/fphar.2024.1395707.

[38] L. Li *et al.*, "Comparative effectiveness of propensity score estimation methods for inverse probability of treatment weighting analysis with complex survey data: a simulation study," *Journal of Survey Statistics and Methodology*, Apr. 2025, doi: 10.1093/jssam/smaf003.

[39] D. Westreich, J. Lessler, and M. J. Funk, "Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression," *Journal of Clinical Epidemiology*, vol. 63, no. 8, pp. 826–833, Aug. 2010, doi: 10.1016/j.jclinepi.2009.11.020.

[40] V. Borisov, T. Leemann, K. Sebler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: a survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7499–7519, Jun. 2024, doi: 10.1109/TNNLS.2022.3229161.

[41] N. Kreif, K. D. Ordaz, R. Moreno-Serra, A. Mirelman, T. Hidayat, and M. Suhrcke, "Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia," *Health Services and Outcomes Research Methodology*, vol. 22, no. 2, pp. 192–227, Jun. 2022, doi: 10.1007/s10742-021-00259-3.

[42] J. E. Brand, J. Xu, B. Koch, and P. Geraldo, "Uncovering sociological effect heterogeneity using tree-based machine learning," *Sociological Methodology*, vol. 51, no. 2, pp. 189–223, Aug. 2021, doi: 10.1177/0081175021993503.

[43] P. Rehill and N. Biddle, "Heterogeneous treatment effect estimation with high-dimensional data in public policy evaluation," 2024.

[44] W. Hu, X. Sun, Q. Liu, L. Wu, and L. Wang, "Uncertainty calibration for counterfactual propensity estimation in recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 6, pp. 3781–3793, Jun. 2025, doi: 10.1109/TKDE.2025.3552658.

[45] N. Bearth and D. Ballinari, "Improving the finite sample estimation of average treatment effects using double/debiased machine learning with propensity score calibration," 2024.

[46] L. Lourenço, L. Weber, L. Garcia, V. Ramos, and J. Souza, "Machine learning algorithms to estimate propensity scores in health policy evaluation: a scoping review," *International Journal of Environmental Research and Public Health*, vol. 21, no. 11, p. 1484, Nov. 2024, doi: 10.3390/ijerph21111484.

[47] K. Furmańczyk and W. Rejchel, "Prediction and variable selection in high-dimensional misspecified binary classification," *Entropy*, vol. 22, no. 5, p. 543, May 2020, doi: 10.3390/E22050543.

[48] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data*, vol. 15, no. 5, pp. 1–46, Oct. 2021, doi: 10.1145/3444944.
[49] K. Wang, L. Rosman, and H. Lu, "Machine learning versus logistic regression for propensity score estimation: a benchmark trial emulation against the PARADIGM-HF randomized trial," *medRxiv*. p. 2025.06.16.25329708, Jun. 17, 2025, doi: 10.1101/2025.06.16.25329708.
[50] J. Tazare *et al.*, "Transparency of high-dimensional propensity score analyses: guidance for diagnostics and reporting," *Pharmacoepidemiology and Drug Safety*, vol. 31, no. 4, pp. 411–423, Apr. 2022, doi: 10.1002/pds.5412.

BIOGRAPHIES OF AUTHORS

Gene Marck B. Catedrilla is a faculty member at the College of Computer Studies, Laguna State Polytechnic University – Los Baños Campus. He earned his Bachelor of Science in Information Technology (BSIT) from LSPU Los Baños and completed his Master in Information Technology (MIT) at the Technological Institute of the Philippines – Manila. Currently, he is pursuing his Doctor in Information Technology (DIT) at Angeles University Foundation. His academic background is grounded in information technology and research, with specific interests in data analytics, artificial intelligence, and the integration of technology in education and governance. He has actively contributed to institutional and student research initiatives, with recent works focusing on the use of ML for interpreting social and health-related data. In addition to teaching, he is involved in curriculum development and capacity-building workshops in Python programming, data visualization, and automated data science tools. His current research reflects a commitment to using data-driven methods to address complex, real-world problems. He has mentored numerous undergraduate research projects and participated in university-funded studies exploring the transformative role of technology in public and academic sectors. He can be contacted at email: catedrilla.genemarck@lspu.edu.ph.

Dr. Joey Aviles is a seasoned academic and IT education leader, currently serving as Faculty Member and Service Unit Coordinator at the College of Computer Studies, and Program Chair for IT Education Programs at the Graduate School of Angeles University Foundation (AUF). He also leads AUF's AI Integration Framework, driving institutional innovation in teaching, operations, and research. He has held teaching and leadership roles at Panpacific University and Al Farabi Colleges in Saudi Arabia. He is a member of the Philippine Society of IT Educators (PSITE), the Council of IT Education Deans and Heads, and a former Regional Quality Assessment Team (RQAT) evaluator. An active researcher, Dr. Aviles has presented and published in IEEE, ACM, and Scopus-indexed venues, focusing on AI, digital transformation, and IT in education. He also holds certifications from Microsoft and Adobe, further enriching his blend of practical and academic expertise. He can be contacted at email: aviles.joey@auf.edu.ph.