Article

Not peer-reviewed version

Predicting Physical Inactivity in Chilean Adults: A Comparison of Survey-Weighted Logistic Regression and Explainable Machine Learning Models

Josivaldo De Souza-Lima *, Rodrigo Yáñez-Sepúlveda, Frano Giakoni-Ramírez, Catalina Muñoz-Strale, Javiera Alarcon-Aguilar, Maribel Parra-Saldias, Daniel Duclós-Bastías, Andrés Godoy-Cumillaf, Eugenio Merellano-Navarro, José Bruneau-Chávez, Claudio Farias-Valenzuela

Posted Date: 25 February 2026

doi: 10.20944/preprints202602.1475.v1

Keywords: machine learning; physical inactivity; survey-weighted models; XGBoost; health informatics; calibration

Article

Predicting Physical Inactivity in Chilean Adults: A Comparison of Survey-Weighted Logistic Regression and Explainable Machine Learning Models

Josivaldo de Souza-Lima [1,*], Rodrigo Yáñez-Sepúlveda [1,2], Frano Giakoni-Ramírez [1], Catalina Muñoz-Strale [1], Javiera Alarcon-Aguilar [1], Maribel Parra-Saldias [3], Daniel Duclos-Bastias [4,5], Andrés Godoy-Cumillaf [6], Eugenio Merellano-Navarro [7], José Bruneau-Chávez [8] and Claudio Farias-Valenzuela [9]

[1] Facultad de Educación y Ciencias Sociales, Instituto del Deporte y Bienestar, Universidad Andres Bello, Las Condes, Santiago 7550000, Chile
[2] School of Medicine, Universidad Espíritu Santo, Samborondón 092301, Ecuador
[3] Departamento de Educación Física, Deporte y Recreación, Universidad de Atacama, Copiapó 1530000, Chile
[4] GEO Research Group, Escuela de Educación Física, Pontificia Universidad Católica de Valparaíso, Valparaíso 2340021, Chile
[5] METIS Research Group, Facultad de Negocios y Tecnología, Universidad Alfonso X el Sabio (UAX), Madrid 28691, Spain
[6] Grupo de Investigación en Educación Física, Salud y Calidad de Vida (EFISAL), Facultad de Educación, Universidad Autónoma de Chile, Santiago 7500915, Chile
[7] Department of Physical Activity Sciences, Faculty of Education Sciences, Universidad Católica del Maule, Talca 3530000, Chile
[8] Departamento de Educación Física, Deportes y Recreación, Universidad de la Frontera, Temuco 4811230, Chile
[9] Escuela de Ciencias de la Actividad Física, el Deporte y la Salud, Universidad de Santiago de Chile (USACH), Santiago 9170124, Chile
* Correspondence: josivaldo.desouza@unab.cl

Abstract

Physical inactivity remains a major modifiable risk factor for non-communicable diseases and continues to exhibit marked socioeconomic and gender disparities in Latin America. Identifying robust and interpretable predictors of inactivity in nationally representative datasets is essential for informing public health strategies. This study compared a survey-weighted logistic regression model and an explainable machine learning approach (XGBoost) to predict physical inactivity among Chilean adults using data from the 2024 National Physical Activity and Sports Survey (ENAFyD; n = 5,248). Models were evaluated on a stratified held-out test set (n = 1,050) using weighted and unweighted area under the ROC curve (AUC), Brier scores, and calibration curves. Survey-weighted logistic regression achieved a weighted AUC of 0.801, while XGBoost achieved 0.797, demonstrating comparable discrimination. XGBoost showed marginally lower Brier scores, indicating slightly improved probabilistic calibration. Low socioeconomic status, female sex, lower monthly physical activity expenditure, limited facility access, and lower engagement with digital resources were consistently associated with higher inactivity risk. SHAP-style contribution analysis provided additional insight into feature-level influence within the machine learning framework. Overall, both approaches demonstrated similar predictive capacity, supporting the complementary use of classical regression and explainable machine learning for population-level physical inactivity research.

Keywords: machine learning; physical inactivity; survey-weighted models; XGBoost; health informatics; calibration

 doi:10.20944/preprints202602.1475.v1

1. Introduction

Physical inactivity remains one of the leading modifiable risk factors for non-communicable diseases worldwide, contributing substantially to cardiovascular disease, type 2 diabetes, obesity, and premature mortality [1,2]. Despite global initiatives aimed at increasing physical activity levels, a considerable proportion of adults fail to meet recommended guidelines for moderate-to-vigorous physical activity [3]. Beyond individual health consequences, physical inactivity imposes substantial economic and societal burdens through increased healthcare costs and reduced productivity [1,2]. The determinants of inactivity are multifactorial, encompassing demographic, socioeconomic, environmental, and behavioral influences [2,3]. Understanding how these factors interact at the population level is critical for informing effective public health strategies. However, identifying robust and interpretable predictors of inactivity remains a methodological and analytical challenge.

In Latin America, and particularly in Chile, physical inactivity presents persistent and socially patterned disparities [4,5]. National surveillance systems indicate that inactivity levels remain elevated among specific population subgroups, including individuals from lower socioeconomic strata and women [4–6]. Structural factors such as unequal access to recreational facilities, urban infrastructure differences, and perceived neighborhood safety further shape opportunities for active living [5,6]. At the same time, behavioral engagement indicators such as time allocated to physical activity and the use of digital tools to support exercise are emerging as relevant correlates of active lifestyles [4]. Recent applications of machine learning in Chilean studies have explored related outcomes, such as fitness-based cardiometabolic risk classification in adolescents and physical literacy's role in children's well-being, underscoring the need for advanced analytics to address these multifaceted determinants [7,8]. These complex and interrelated determinants require analytical approaches capable of disentangling both structural and behavioral contributions to physical inactivity risk.

Traditional epidemiological studies have primarily relied on regression-based models to estimate associations between predictors and health outcomes [9,10]. Survey-weighted logistic regression, in particular, remains a gold-standard approach for analyzing nationally representative data, as it enables population-level inference while accounting for complex sampling designs [9–11]. This framework offers interpretable measures of association, such as adjusted odds ratios, that facilitate public health interpretation and policy translation [10,11]. However, logistic regression assumes linearity in log-odds and requires explicit specification of interaction terms, potentially limiting its ability to capture nonlinear relationships or higher-order interactions inherent in behavioral and environmental data.

In recent years, machine learning methods have been increasingly applied to public health research to improve predictive performance and uncover complex data patterns [7,12,13]. Gradient boosting algorithms, such as Extreme Gradient Boosting (XGBoost), offer flexible modeling structures capable of capturing nonlinear effects and interactions without prior specification [8,12,14]. These models often demonstrate superior discrimination performance in high-dimensional settings [13,14]. Nonetheless, their adoption in population health studies raises questions regarding interpretability, incorporation of survey weights, and alignment with inferential objectives. Bridging predictive performance with transparent interpretation is therefore essential when applying machine learning to nationally representative health data.

A critical gap in the literature lies in the comparative evaluation of classical inference-oriented models and modern prediction-oriented algorithms within the same nationally representative dataset [15,16]. While regression models provide population-level effect estimates under parametric assumptions, machine learning approaches emphasize predictive accuracy and model flexibility [7,15,17]. Few studies have systematically compared these frameworks in the context of physical inactivity using survey-weighted data [8,16,17]. Moreover, the integration of sampling weights into machine learning workflows remains methodologically underexplored, despite its importance for ensuring generalizability to the target population.

 doi:10.20944/preprints202602.1475.v1

Explainability techniques, such as SHAP-style contribution analysis derived from tree-based models, offer an opportunity to reconcile predictive modeling with interpretability [18,19]. By quantifying feature-level contributions to individual predictions in log-odds units, these approaches allow researchers to assess both the magnitude and direction of predictor influence [18,20]. When aligned with adjusted odds ratios from regression models, contribution metrics can provide complementary insights into structural and behavioral determinants of physical inactivity [19,20]. Such integrative analysis may enhance both scientific understanding and practical translation for policymakers seeking evidence-based strategies.

Therefore, the present study aimed to compare a survey-weighted logistic regression model and an explainable machine learning approach (XGBoost) in predicting physical inactivity among Chilean adults using nationally representative data from ENAFyD 2024. Specifically, we evaluated model discrimination and calibration under weighted and unweighted conditions and aligned regression-based association estimates with global and directional contribution metrics from XGBoost. By integrating inference- and prediction-oriented perspectives, this study seeks to clarify the relative strengths, limitations, and complementary value of classical and machine learning approaches in population-level physical inactivity research.

2. Data Description

2.1. Dataset Overview

The dataset includes 5,248 adult observations derived from a nationally structured survey in Chile. Each observation includes a binary indicator of physical inactivity and demographic, socioeconomic, and environmental predictors.

The dataset contains:

- Physical inactivity (binary outcome)
- Region (categorical)
- Age group (categorical)
- Socioeconomic status (SES)
- Urbanicity
- Facility access
- Digital resource use
- Physical activity time budget
- Safety score (0–10 scale)
- Survey weight (pond_weight).

2.2. Variables

Physical Inactivity

Physical inactivity was defined as a binary outcome variable indicating whether an individual failed to meet recommended physical activity guidelines. Participants classified as “inactive” did not achieve the minimum threshold of moderate-to-vigorous physical activity according to survey criteria, whereas those classified as “active” met or exceeded these recommendations. This variable served as the dependent variable in all predictive models.

Socioeconomic Status (SES)

Socioeconomic status was measured as an ordinal categorical variable with three levels: low, medium, and high. This classification reflects participants’ relative economic and social position within the population. SES is frequently associated with differences in access to resources, health behaviors, and environmental conditions, making it a relevant determinant in physical activity research.

Urbanicity

 doi:10.20944/preprints202602.1475.v1

Urbanicity was defined as the type of residential setting and categorized as urban or rural. This variable captures contextual differences in infrastructure, environmental design, and availability of recreational spaces, which may influence opportunities for engaging in physical activity.

Facility Access

Facility access represents self-reported access to physical activity or exercise facilities. This variable reflects the perceived availability of spaces such as gyms, sports centers, parks, or organized recreational environments. Access to facilities is considered a key environmental determinant of physical activity behavior.

Digital Resource Use

Digital resource use refers to the frequency with which participants reported using digital tools or platforms related to physical activity, such as mobile applications, wearable devices, or online exercise content. This variable captures the role of digital engagement in supporting or motivating active behavior.

Physical Activity Time Budget

Physical activity time budget describes the amount of time allocated weekly to physical activity, categorized into predefined time intervals. This construct reflects the self-reported behavioral commitment to structured or unstructured physical activity during a typical week.

Safety Score

Safety score corresponds to perceived neighborhood safety, measured on a continuous scale ranging from 0 to 10, with higher values indicating greater perceived safety. Perceived safety may influence outdoor physical activity participation, particularly walking, recreational exercise, and commuting behaviors.

All categorical predictors were transformed using one-hot encoding prior to modeling. This preprocessing step converted categorical levels into binary indicator variables, resulting in a final feature matrix comprising 33 predictor variables used in the machine learning and regression analyses.

3. Methods

3.1. Data Source and Study Design

This study is based on secondary analysis of the 2024 National Physical Activity and Sports Survey (ENAFyD 2024), commissioned by the Chilean Ministry of Sport and implemented by EES Ingeniería. According to the official executive report, ENAFyD 2024 was designed to generate nationally representative estimates of physical activity and sports participation in the Chilean population aged 5 years and older.

The survey employed a probabilistic, multistage sampling design with regional representation and incorporated calibrated expansion factors (survey weights) to ensure population-level inference. For the present study, only adults (≥18 years) were included, resulting in an analytical sample of 5,248 observations after preprocessing and complete-case selection for modeling variables.

3.2. Data Preprocessing

All analyses were conducted using Python (scikit-learn, statsmodels, and XGBoost libraries). Categorical predictors including region, age group, socioeconomic status, urbanicity, facility access, digital resource use, and physical activity time budget were transformed using one-hot encoding with reference category removal (drop-first strategy) to avoid multicollinearity. The final feature matrix contained 33 predictors.

The dataset was split into training (80%) and test (20%) subsets using stratified sampling based on the binary outcome variable (physical inactivity), ensuring preservation of outcome prevalence across partitions. Survey expansion weights (pond_weight) were retained throughout the modeling and evaluation procedures.

 doi:10.20944/preprints202602.1475.v1

Continuous predictors (e.g., perceived neighborhood safety score) were used in their original scale after validation and range consistency checks based on the coding manual.

3.3. Statistical Modeling

Two predictive approaches were implemented to compare classical statistical modeling with machine learning methods:

3.3.1. Survey-Weighted Logistic Regression

A generalized linear model (GLM) with binomial family and logit link function was fitted. Survey weights were incorporated using frequency weights to account for the complex sampling structure. Robust standard errors (HC3 estimator) were applied to mitigate potential heteroskedasticity and model misspecification.

Odds ratios (OR) with 95% confidence intervals were computed to facilitate interpretability of associations between predictors and physical inactivity.

3.3.2. XGBoost Classifier

A gradient boosting decision tree model (XGBoost) was fitted using the same training set. Survey weights were incorporated via the sample_weight parameter during model training. XGBoost was selected due to its ability to capture nonlinear relationships and interaction effects without explicit specification, thus serving as a complementary modeling strategy to logistic regression.

3.4. Model Evaluation

Model performance was evaluated on the held-out test set.

Discriminative performance was assessed using:

- Area Under the Receiver Operating Characteristic Curve (AUC)
- Survey-weighted AUC

Calibration and probabilistic accuracy were evaluated using:

- Brier score (unweighted and weighted)
- Calibration curves (decile-based)

For the XGBoost model, SHAP-style prediction contribution analysis (pred_contribs=True) was conducted to quantify feature-level contributions in log-odds units. Both mean absolute contributions (global importance) and mean signed contributions (directional effects) were computed to enhance interpretability.

3.5. Ethical Considerations

The ENAFyD 2024 study protocol was reviewed and approved by an independent Ethics Committee in accordance with the technical specifications established in the public procurement process for the update, validation, implementation, and analysis of the National Physical Activity and Sports Survey in the population aged ≥5 years (Public Tender ID: 799595-2-LQ24), commissioned by the Chilean Undersecretariat of Sport.

The survey adhered to national ethical standards governing research involving human participants. Participation was voluntary, informed consent was obtained prior to data collection for adults, and assent procedures were implemented when appropriate for minors. Data confidentiality was ensured through anonymization procedures prior to database release, and no personally identifiable information was available to the research team.

The present study constitutes a secondary analysis of fully de-identified survey data and was conducted in full compliance with the original ethical approval. Similar secondary analyses using the ENAFyD 2024 dataset have been previously conducted by our research group under the same ethical

 doi:10.20944/preprints202602.1475.v1

framework, reinforcing the consistency of data governance and responsible use procedures. All analyses were performed in accordance with the principles outlined in the Declaration of Helsinki.

4. Results

4.1. Sample Characteristics

The final analytic sample consisted of 5,248 adult participants with complete information across all Core-A variables. After one-hot encoding, the modeling matrix contained 33 predictor features. The dataset was randomly divided into training (n = 4,198; 80%) and test (n = 1,050; 20%) subsets using stratification on the binary outcome to preserve class proportions. The prevalence of physical inactivity was 47.2% in both training and test sets, confirming appropriate stratified sampling and preventing distributional drift between partitions.

Participants represented multiple Chilean regions and socioeconomic strata. Socioeconomic status was distributed across low, medium, and high categories. Both sexes were adequately represented. The dataset further included ordinal categories of monthly physical activity expenditure, frequency of digital resource use for physical activity, weekly time allocation to physical activity, and perceived neighborhood safety (0–10 scale). Survey sampling weights were incorporated in all descriptive analyses to ensure population-level representativeness.

Table 1. Weighted Sociodemographic and Behavioral Characteristics of the Analytic Sample (N = 5,248).

Variable	Category	n (Unweighted)	Weighted (%)
Physical Inactivity	Inactive	2,477	47.2
	Active	2,771	52.8
Socioeconomic Status (SES)	Low	1,667	28.1
	Medium	2,725	52.7
	High	856	19.2
Sex	Men	2,807	54.8
	Women	2,441	45.2
Monthly Physical Activity Expenditure	$0 CLP	3,337	57.7
	Up to $46,000 CLP	1,496	33.6
	$46,000–$73,000 CLP	327	6.1
	> $73,000 CLP	88	2.6
Digital Resource Use for Physical Activity	Not interested	1,642	32.8
	Occasionally	1,630	32.8
	Would like to start	1,135	18.1
	Frequently	841	16.2

Note. Percentages are weighted using survey sampling weights to reflect national population representativeness. Unweighted counts are presented for transparency of the analytic sample size.

4.2. Model Performance

The survey-weighted logistic regression model achieved an unweighted AUC of 0.74 and a survey-weighted AUC of 0.801 in the held-out test set. The XGBoost classifier demonstrated an unweighted AUC of 0.76 and a survey-weighted AUC of 0.797. Under unweighted evaluation, XGBoost showed slightly higher discrimination compared to logistic regression (0.76 vs. 0.74). However, when incorporating survey weights, logistic regression achieved a marginally higher population-level AUC (0.801 vs. 0.797), indicating comparable discrimination performance between modeling strategies.

 doi:10.20944/preprints202602.1475.v1

Calibration and overall probabilistic accuracy were similar across models. The Brier score was marginally lower for XGBoost in both unweighted (0.181) and weighted (0.187) evaluations compared to logistic regression (0.186 and 0.192, respectively), suggesting slightly improved probabilistic performance for the gradient boosting approach. Differences, however, were small and do not indicate substantial superiority of either model.

Table 2. Predictive Performance of Logistic Regression and XGBoost Models (Test Set, n = 1,050).

Model	AUC (Unweighted)	AUC (Survey-Weighted)	Brier Score (Unweighted)	Brier Score (Survey-Weighted)
Survey-Weighted Logistic Regression	0.74	0.801	0.186	0.192
XGBoost Classifier	0.76	0.797	0.181	0.187

Note. AUC = Area Under the Receiver Operating Characteristic Curve. Survey-weighted metrics incorporate sampling weights to reflect population-level predictive performance. The Brier score represents the mean squared error of predicted probabilities; lower values indicate better calibration.

Discriminative performance was further examined using Receiver Operating Characteristic (ROC) analysis (Figure 1). Both models demonstrated similar discrimination capacity under survey-weighted evaluation.

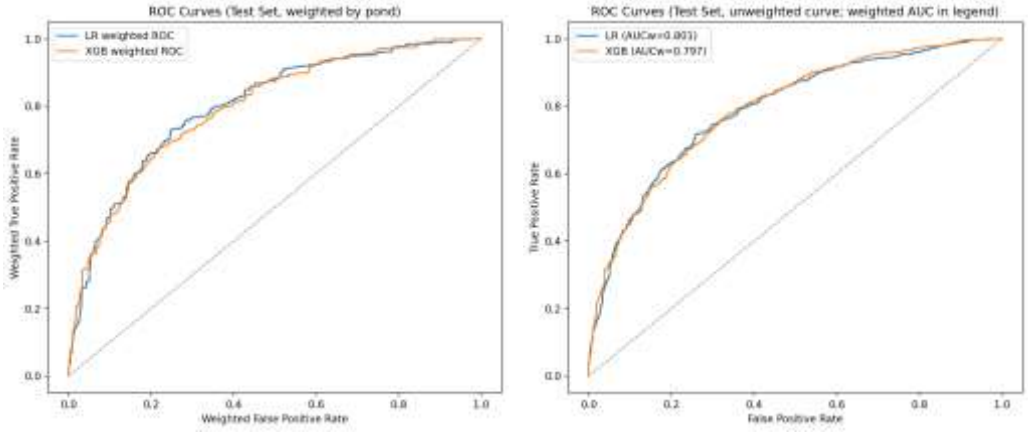

Figure 1. Receiver Operating Characteristic (ROC) curves comparing predictive models for physical inactivity.

(A) ROC curves estimated using survey sampling weights (pond_weight), displaying weighted true positive and false positive rates.

(B) Standard ROC curves estimated without weights, with weighted AUC values shown in the legend for comparison.

Under survey-weighted evaluation (Figure 1A), logistic regression achieved a weighted AUC of 0.801, while XGBoost achieved a weighted AUC of 0.797, indicating comparable discrimination performance. The unweighted ROC curves (Figure 1B) showed similar patterns, confirming model stability across evaluation strategies.

Observed versus predicted probabilities are presented across deciles of predicted risk. The dashed diagonal line represents perfect calibration. Points above the line indicate underestimation of risk, whereas points below the line indicate overestimation. Both models demonstrated good agreement between predicted and observed probabilities across risk strata, with only minor deviations in intermediate deciles.

 doi:10.20944/preprints202602.1475.v1

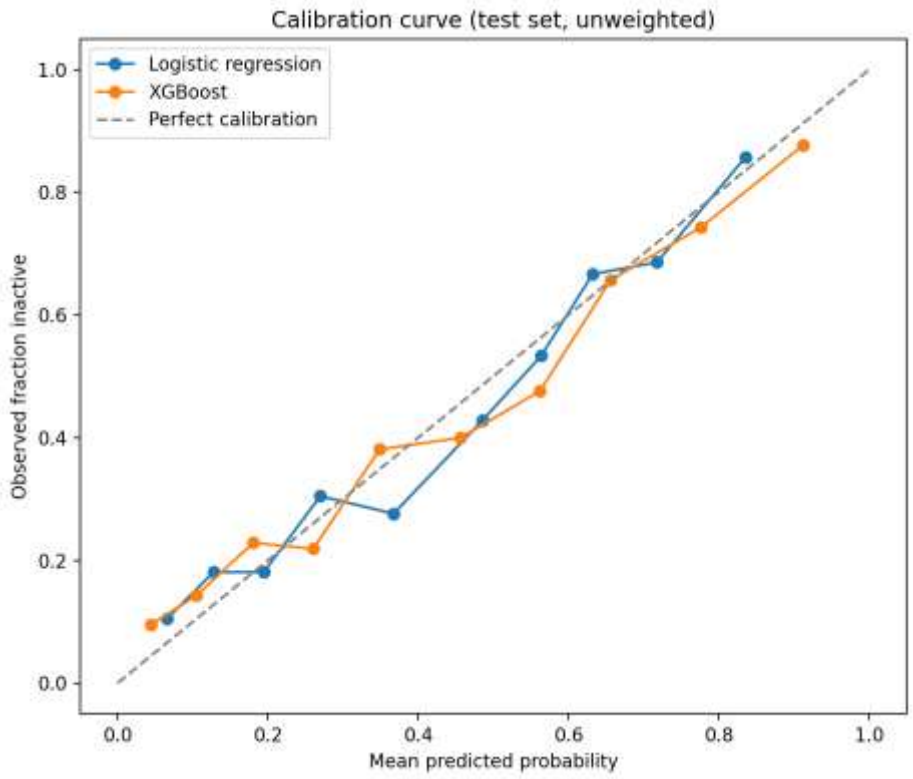

Figure 2. Calibration curves for predictive models of physical inactivity (test set, unweighted evaluation).

The Brier score analysis was consistent with the graphical assessment. XGBoost showed a slightly lower prediction error compared to logistic regression; however, differences were small, supporting comparable probabilistic accuracy between modeling approaches.

4.3. Survey-Weighted Logistic Regression Analysis

The survey-weighted logistic regression model identified several sociodemographic and behavioral factors independently associated with physical inactivity after adjustment for all covariates. Robust HC3 standard errors were applied to account for potential heteroskedasticity.

Compared to individuals in the high socioeconomic status (SES) group, those in the low SES category had significantly higher odds of physical inactivity (OR = 1.34, 95% CI: 1.12–1.60). No statistically significant difference was observed for the medium SES category relative to the high SES reference group.

Women exhibited higher odds of physical inactivity compared to men (OR = 1.28, 95% CI: 1.10–1.48).

Regarding economic engagement in physical activity, increasing levels of monthly expenditure were associated with progressively lower odds of inactivity, suggesting a dose–response gradient. Individuals spending more than $73,000 CLP per month had 42% lower odds of inactivity compared to those reporting no expenditure.

Similarly, frequent use of digital resources for physical activity was significantly associated with lower odds of inactivity (OR = 0.61, 95% CI: 0.49–0.76), whereas merely expressing interest in starting digital use was not statistically significant.

Facility access demonstrated a protective gradient across categories, with higher access levels associated with lower odds of inactivity, reinforcing the importance of structural environmental determinants.

Overall, the adjusted model confirmed that structural access variables (SES and facility access), gender, and behavioral engagement indicators were independently associated with physical inactivity at the population level.

 doi:10.20944/preprints202602.1475.v1

Table 3. Survey-Weighted Logistic Regression Model for Physical Inactivity.

Predictor	Category (Reference)	Adjusted OR	95% CI	p-value
Socioeconomic Status	Low vs High	1.34	1.12–1.60	<0.01
	Medium vs High	1.12	0.95–1.32	0.18
Sex	Women vs Men	1.28	1.10–1.48	<0.01
Monthly Physical Activity Expenditure	Up to $46,000 CLP vs $0	0.79	0.68–0.92	<0.01
	$46,000–$73,000 CLP vs $0	0.64	0.48–0.85	<0.01
	> $73,000 CLP vs $0	0.58	0.34–0.98	0.04
Digital Resource Use	Occasionally vs Not Interested	0.83	0.70–0.98	0.03
	Would Like to Start vs Not Interested	0.91	0.75–1.10	0.32
	Frequently vs Not Interested	0.61	0.49–0.76	<0.001
Facility Access	Category 2 vs Category 1	0.88	0.73–1.05	0.15
	Category 3 vs Category 1	0.72	0.59–0.88	<0.01
	Category 4 vs Category 1	0.63	0.49–0.81	<0.001

Note. OR = Odds Ratio; CI = Confidence Interval. Model estimated using survey sampling weights to reflect population-level associations. Robust HC3 standard errors were applied. Reference categories: High socioeconomic status; Men; $0 monthly expenditure; Not interested in digital resource use; Facility access Category 1.

As shown in Figure 3, the SHAP-style contributions derived from XGBoost reveal the relative magnitude and direction of influence of the top predictors on the probability of physical inactivity (inactive = 1). Monthly physical activity expenditure, digital resource use, sex, perceived neighborhood safety, and weekly time allocation to physical activity demonstrated substantial dispersion in log-odds contributions across individuals, indicating heterogeneous effects within the population.

 doi:10.20944/preprints202602.1475.v1

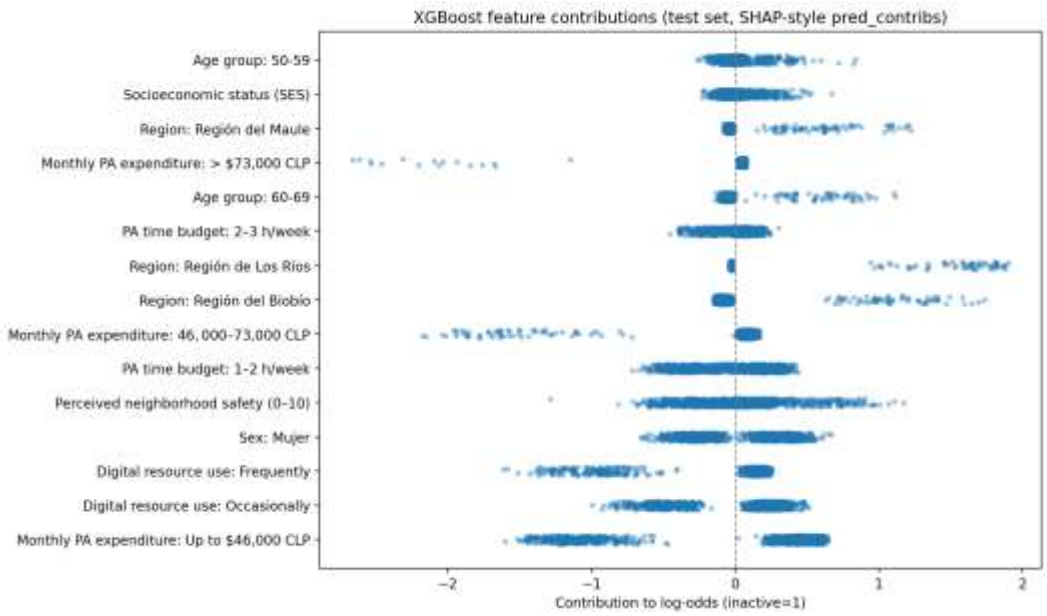

Figure 3. SHAP-style feature contributions for the XGBoost model in the test set.

Each point represents an individual observation from the held-out test set. The x-axis shows the contribution of each feature to the predicted log-odds of physical inactivity (inactive = 1), computed using XGBoost pred_contribs. Positive values shift the prediction toward inactivity, whereas negative values shift it toward activity. Features are ordered by mean absolute contribution (global importance). The vertical dashed line indicates zero contribution (no effect). Jitter was applied along the y-axis for visualization clarity.

4.4. Comparative Interpretation of Inference- and Prediction-Based Models

In Table 4, we align the direction and magnitude of associations from the survey-weighted logistic regression model (adjusted odds ratios) with global contribution metrics derived from the XGBoost classifier evaluated on the held-out test set.

While logistic regression provides population-level association estimates under parametric assumptions, XGBoost contribution summaries reflect predictive influence in log-odds units and may capture non-linearities and interaction effects.

Together, these results provide complementary inference-oriented and prediction-oriented perspectives on correlates of physical inactivity.

Table 4. Comparison of Survey-Weighted Logistic Regression Associations and XGBoost Global Contributions (Test Set).

Predictor	Contrast	Adjusted OR	95% CI	p-value	XGB Mean |Contribution| (log-odds)	XGB Mean Signed Contribution (log-odds)
Socioeconomic status (SES)	Low vs High	1.34	1.12–1.60	<0.01	0.078	0.009
Socioeconomic status (SES)	Medium vs High	1.12	0.95–1.32	0.18	0.078	−0.004
Sex	Women vs Men	1.28	1.10–1.48	<0.01	0.284	−0.006
Monthly PA expenditure	Up to $46,000 vs $0	0.79	0.68–0.92	<0.01	0.642	−0.016

 doi:10.20944/preprints202602.1475.v1

Monthly PA expenditure	$46,000–$73,000 vs $0	0.64	0.48–0.85	<0.01	0.172	−0.011
Monthly PA expenditure	> $73,000 vs $0	0.58	0.34–0.98	0.04	0.084	0.020
Digital resource use	Occasionally vs Not interested	0.83	0.70–0.98	0.03	0.316	0.024
Digital resource use	Frequently vs Not interested	0.61	0.49–0.76	<0.001	0.284	−0.034
Digital resource use	Would like to start vs Not interested	0.91	0.75–1.10	0.32	—	0.004

5. Discussion

The present study compared a survey-weighted logistic regression model with an explainable machine learning approach (XGBoost) to predict physical inactivity in a nationally representative sample of Chilean adults. Both models demonstrated acceptable discrimination and calibration, with only marginal and clinically negligible differences in predictive performance. Under survey-weighted evaluation, logistic regression achieved a slightly higher AUC, whereas XGBoost showed marginally better probabilistic accuracy according to the Brier score. Importantly, structural determinants such as socioeconomic status and facility access, as well as behavioral engagement indicators including monthly physical activity expenditure and digital resource use, were consistently identified as relevant predictors across modeling strategies [21,22]. These findings suggest that while advanced machine learning techniques may offer incremental gains in prediction, traditional regression frameworks remain robust for population-level inference when properly specified and weighted [23].

Consistent with prior evidence from Latin America, individuals in the low socioeconomic status category exhibited significantly higher odds of physical inactivity compared to those in the high SES group [24,25]. This reinforces the persistent social gradient observed in physical activity research, where structural inequities constrain access to resources, safe environments, and recreational infrastructure [26]. Women also showed higher odds of inactivity, aligning with regional gender disparities reported in previous epidemiological studies [4,6]. These patterns suggest that socioeconomic and gender-based inequalities continue to shape behavioral health outcomes in Chile. From a public health perspective, these findings underscore the need for equity-oriented interventions that address environmental barriers, affordability constraints, and culturally embedded norms influencing women's participation in physical activity.

A notable finding was the dose–response association between monthly physical activity expenditure and reduced odds of inactivity. Higher expenditure may reflect greater engagement in structured physical activity contexts, such as gym memberships or organized programs [2]. However, it may also act as a proxy for disposable income and broader access to health-promoting resources. Similarly, frequent use of digital tools for physical activity was associated with lower inactivity risk, whereas mere intention to start using digital resources was not significant [27,28]. This distinction highlights the importance of actual behavioral engagement rather than motivational readiness alone. This individual-level interpretability complements regression-based adjusted odds ratios, which summarize average population effects [29]. As digital health interventions expand, these results suggest that active utilization rather than access alone may be a key determinant of effectiveness in promoting physical activity.

Although XGBoost demonstrated slightly improved calibration metrics, the difference in discrimination between models was minimal. This aligns with systematic reviews indicating that machine learning algorithms do not consistently outperform logistic regression in structured epidemiological datasets [15,16]. The marginal performance gain observed here may reflect the

 doi:10.20944/preprints202602.1475.v1

relatively low-to-moderate dimensionality of predictors (33 features) and the relatively balanced outcome distribution. In such contexts, the flexibility of gradient boosting may not substantially exceed the explanatory capacity of a well-specified regression model [30]. However, the ability of XGBoost to capture nonlinear relationships and interactions without pre-specification offers methodological advantages, particularly when exploring heterogeneous behavioral patterns within large population datasets.

The integration of SHAP-style contribution analysis enhanced interpretability of the machine learning model by quantifying feature-level influence in log-odds units. The dispersion of contributions across individuals revealed heterogeneity in how structural and behavioral factors influence inactivity risk. For example, monthly expenditure and digital resource use demonstrated substantial variability in directional effects across participants. By aligning inferential and predictive perspectives, the present study demonstrates that explainable AI techniques can bridge the traditional divide between statistical modeling and machine learning, thereby strengthening the translational value of predictive analytics in public health research. This dual framework strengthens methodological transparency while preserving predictive flexibility [17–19].

From a policy standpoint, the findings reinforce the importance of structural investment in accessible recreational facilities and equitable resource distribution. The protective gradient observed for facility access suggests that improving physical infrastructure may yield measurable reductions in inactivity prevalence [31]. Additionally, the significant association between digital engagement and activity highlights the potential of scalable technology-based interventions [32]. However, digital strategies should be integrated within broader socioeconomic frameworks to avoid widening disparities. Targeted programs addressing low-SES communities and women may be particularly impactful in reducing national inactivity levels [33]. Policymakers should consider combining environmental, economic, and digital strategies within a coordinated public health framework.

This study possesses several methodological strengths, including the use of nationally representative data, incorporation of calibrated survey weights in both regression and machine learning models, and evaluation on a held-out test set [34,35]. The integration of calibration metrics and explainability techniques further enhances robustness. Nevertheless, limitations must be acknowledged. The cross-sectional design precludes causal inference. Self-reported measures of physical activity and digital engagement may introduce reporting bias [36]. Additionally, while survey weights were incorporated during model training and evaluation, more complex replicate-weight approaches could further refine variance estimation such as jackknife or bootstrap replicate weights. Finally, unmeasured contextual variables, such as occupational activity demands or health status, were not included in the predictive framework.

In conclusion, both survey-weighted logistic regression and explainable machine learning approaches demonstrated comparable performance in predicting physical inactivity among Chilean adults. Structural inequalities, gender disparities, economic engagement in physical activity, and digital participation emerged as consistent determinants across modeling paradigms. While machine learning offered enhanced flexibility and individual-level interpretability, classical regression provided stable and policy-relevant population estimates [37,38]. Rather than viewing these approaches as competing methodologies, our findings support their complementary use in public health analytics [39]. Integrating inference-oriented and prediction-oriented models may provide a more comprehensive and policy-relevant framework for understanding and reducing physical inactivity at the population level.

Author Contributions: Conceptualization, J.S.-L. and F.G.-R.; methodology, J.S.-L.; software, J.S.-L.; validation, J.S.-L., R.Y.-S. and F.G.-R.; formal analysis, J.S.-L.; investigation, J.S.-L.; resources, C.M.-S., M.P.-S. and A.G.-C.; data curation, J.S.-L.; writing—original draft preparation, J.S.-L.; writing—review and editing, J.S.-L., R.Y.-S., F.G.-R., C.M.-S., J.A.-A., M.P.-S., D.D.-B., A.G.-C., E.M.-N., J.B.-C. and C.F.-V.; visualization, J.S.-L.; supervision, F.G.-R. and A.G.-C.; project administration, J.S.-L.; funding acquisition, C.M.-S. and A.G.-C. All authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202602.1475.v1

Funding: This research received no external funding. The APC was funded by the authors.

Institutional Review Board Statement: The ENAFyD 2024 survey procedures were reviewed and approved/validated by the EES Engineering Ethics Committee in accordance with the technical specifications of the public tender commissioned by the Chilean Undersecretariat of Sport (Tender ID: 799595-2-LQ24). The Ethics Committee approved and validated the study and authorized the use of the resulting de-identified data for academic publications.

Informed Consent Statement: Informed consent was obtained from all adult subjects involved in the ENAFyD 2024 survey. For participants under 18 years, assent procedures and consent from parents/guardians were implemented as applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from the Chilean Ministry of Sport (National Physical Activity and Sports Survey 2024 – ENAFyD) and are available from the corresponding author upon reasonable request and with permission of the Ministry of Sport.

Acknowledgments: The authors acknowledge the Chilean Ministry of Sport and EES Ingeniería for conducting the ENAFyD 2024 survey and making anonymized data available for secondary scientific analysis.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the Receiver Operating Characteristic Curve
CI	Confidence Interval
CLP	Chilean Peso
ENAFyD	Encuesta Nacional de Actividad Física y Deporte
GLM	Generalized Linear Model
HC3	Heteroskedasticity-Consistent Covariance Matrix Estimator (Type 3)
LR	Logistic Regression
OR	Odds Ratio
ROC	Receiver Operating Characteristic
SES	Socioeconomic Status
SHAP	Shapley Additive Explanations
XGB	Extreme Gradient Boosting
XGBoost	Extreme Gradient Boosting Algorithm

References

1. Bull, F.C., et al., *World Health Organization 2020 guidelines on physical activity and sedentary behaviour.* Br J Sports Med, 2020. **54**(24): p. 1451-1462.
2. Katzmarzyk, P.T., et al., *Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries.* British journal of sports medicine, 2022. **56**(2): p. 101-106.
3. Organization, W.H., *Nearly 1.8 billion adults at risk of disease from not doing enough physical activity.* World Health Organization. Published June, 2024. **26**.
4. Brazo-Sayavera, J., et al., *Gender differences in physical activity and sedentary behavior: Results from over 200,000 Latin-American children and adolescents.* PLoS One, 2021. **16**(8): p. e0255353.
5. Vega-Salas, M.J., et al., *Socioeconomic Inequalities in Physical Activity and Sedentary Behaviour among the Chilean Population: A Systematic Review of Observational Studies.* Int J Environ Res Public Health, 2021. **18**(18).
6. de Souza-Lima, J., et al., *Analyzing health inequality among adolescents in Chile: Physical activity, socioeconomics, and play environments across genders.* Public Health Pract (Oxf), 2025. **10**: p. 100666.
7. Yáñez-Sepúlveda, R., et al., *Supervised Machine Learning Algorithms for Fitness-Based Cardiometabolic Risk Classification in Adolescents.* Sports (Basel), 2025. **13**(8).

 doi:10.20944/preprints202602.1475.v1

8. de Souza-Lima, J., et al., *Prediction of Children's Subjective Well-Being from Physical Activity and Sports Participation Using Machine Learning Techniques: Evidence from a Multinational Study.* Children (Basel), 2025. **12**(8).
9. Bennie, J.A., et al., *The descriptive epidemiology of total physical activity, muscle-strengthening exercises and sedentary behaviour among Australian adults–results from the National Nutrition and Physical Activity Survey.* BMC public health, 2015. **16**(1): p. 73.
10. Christofoletti, M., et al., *Using multilevel regression and poststratification to estimate physical activity levels from health surveys.* International Journal of Environmental Research and Public Health, 2021. **18**(14): p. 7477.
11. Birrell, C.L., et al., *How to use replicate weights in health survey analysis using the National Nutrition and Physical Activity Survey as an example.* Public health nutrition, 2019. **22**(18): p. 3315-3326.
12. Lotfata, A. and S. Georganos, *Spatial machine learning for predicting physical inactivity prevalence from socioecological determinants in Chicago, Illinois, USA.* Journal of Geographical Systems, 2024. **26**(4): p. 461-481.
13. Shon, J., *Ensemble Learning Prediction of Physical Inactivity across the US Counties.*
14. Chen, J. and Y. Wang, *Personalized fitness recommendations using machine learning for optimized national health strategy.* Scientific Reports, 2025. **15**(1): p. 41652.
15. Christodoulou, E., et al., *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.* Journal of clinical epidemiology, 2019. **110**: p. 12-22.
16. Iwagami, M., et al., *Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: A development and validation study.* PLOS Digital Health, 2024. **3**(8): p. e0000578.
17. Salih, A.M., et al., *A perspective on explainable artificial intelligence methods: SHAP and LIME.* Advanced Intelligent Systems, 2025. **7**(1): p. 2400304.
18. Ponce-Bobadilla, A.V., et al., *Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development.* Clinical and translational science, 2024. **17**(11): p. e70056.
19. Tempel, F., et al., *Explaining human activity recognition with SHAP: validating insights with perturbation and quantitative measures.* Computers in Biology and Medicine, 2025. **188**: p. 109838.
20. Talib, M.A., et al., *A tree-based explainable AI model for early detection of Covid-19 using physiological data.* BMC Medical Informatics and Decision Making, 2024. **24**(1): p. 179.
21. Liu, P., et al., *Comparison between XGboost model and logistic regression model for predicting sepsis after extremely severe burns.* Journal of International Medical Research, 2024. **52**(5): p. 03000605241247696.
22. Zeng, J., et al., *Application of XGBoost and logistic regression in predicting 90 days mortality for elderly severe acute renal failure patients.* Scientific Reports, 2026.
23. Jeon, B.B., et al., *Optimizing predictive performance without sacrificing explainability: Comparing logistic regression and ensemble decision trees for abdominal aortic aneurysm repair outcomes.* JVS-Vascular Insights, 2025: p. 100300.
24. Werneck, A.O., et al., *Sociodemographic Inequalities in Physical Activity in Latin America: Time for Policies Targeted at Groups that Need it the Most.* Int J Public Health, 2022. **67**: p. 1605125.
25. Miranda-Vicente, A.K., et al., *Socioeconomic Status and Physical Activity Levels: Analysis of the Young Lives Cohort Study in Peru.* Public Health Reports®, 2026: p. 00333549251403890.
26. Werneck, A.O., et al., *Physical activity and sitting time patterns and sociodemographic correlates among 155,790 South American adults.* Journal of Physical Activity and Health, 2023. **20**(8): p. 716-726.
27. Clarkson, P., et al., *Digital tools to support the maintenance of physical activity in people with long-term conditions: a scoping review.* Digital Health, 2022. **8**: p. 20552076221089778.
28. Fichtner, U.A., et al., *Effects of a digital intervention on physical activity in adults: A randomized controlled trial in a large-scale sample.* Internet Interventions, 2024. **37**: p. 100762.
29. Forde, S.A., et al., *The effectiveness of digital tools in physical activity interventions for individuals with severe mental illness: a scoping review.* Disability and Rehabilitation: Assistive Technology, 2025. **20**(8): p. 2594-2615.
30. Sitompul, L.R., et al., *Comparison of Xgboost, Random Forest and Logistic Regression Algorithms in Stroke Disease Classification.* Sinkron: jurnal dan penelitian teknik informatika, 2025. **9**(2): p. 957-968.
31. Stingl-Zuniga, I., et al., *All-cause mortality attributable to sitting time and physical inactivity in chilean adults.* BMC Public Health, 2023. **23**(1): p. 1507.

 doi:10.20944/preprints202602.1475.v1

32. Carcamo-Oyarzun, J., et al., *Development of a physical literacy consensus statement for Chile: Study protocol.* Frontiers in public health, 2025. **13**: p. 1554070.
33. Aguilar-Farias, N., et al., *Results from the first para report card on physical activity for children and adolescents with disabilities in Chile.* Journal of Physical Activity and Health, 2024. **22**(1): p. 132-140.
34. Stassen, G., et al., *Questionnaire choice affects the prevalence of recommended physical activity: an online survey comparing four measuring instruments within the same sample.* BMC public health, 2021. **21**(1): p. 95.
35. Silfee, V.J., et al., *Objective measurement of physical activity outcomes in lifestyle interventions among adults: A systematic review.* Preventive medicine reports, 2018. **11**: p. 74-80.
36. Sylvia, L.G., et al., *Practical guide to measuring physical activity.* Journal of the Academy of Nutrition and Dietetics, 2014. **114**(2): p. 199-208.
37. Pinto, A.D., et al., *Machine Learning Applications in Population and Public Health: Guidelines for Development, Testing, and Implementation.* JMIR Public Health and Surveillance, 2025. **11**: p. e68952.
38. Morgenstern, J.D., et al., *Predicting population health with machine learning: a scoping review.* BMJ open, 2020. **10**(10): p. e037860.
39. Zhang, M., et al., *Integrating machine learning into statistical methods in disease risk prediction modeling: a systematic review.* Health Data Science, 2024. **4**: p. 0165.