risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 15/12/2024
Aceitação/Acceptance: 28/02/2025

Modelo predictivo para identificar hogares beneficiarios de programas de transferencias monetarias: una comparación de técnicas de machine learning

Jiang Wagner Mamani Lopez[1], Juliana Mery Bautista Lopez[1], Ignacio Aguaded[2]

jmamanilo@unsa.edu.pe; jbautistal@unsa.edu.pe; aguaded@uhu.es

[1] Universidad Nacional de San Agustín de Arequipa, 04001, Arequipa, Perú.

[2] Universidad de Huelva, 21071, Huelva, España.

DOI: 10.17013/risti.57.3–18

Resumen: Los programas de transferencias monetarias son una herramienta clave para reducir la pobreza y mejorar el bienestar de los hogares vulnerables en países en desarrollo. Sin embargo, la correcta selección de beneficiarios sigue siendo un desafío. Este estudio evalúa distintas técnicas de *machine learning* para predecir la participación del programa Juntos en Perú, empleando datos de la Encuesta Nacional de Hogares (ENAHO) del 2023. Se compararon modelos como regresión logística, árboles de decisión, máquina de soporte vectorial, *gradient boosting machine,* bosque aleatorio*, LightGBM, XGBoost* y *CatBoost*. Los resultados muestran que *XGBoost* presenta el mejor desempeño en la clasificación de beneficiarios. Estos hallazgos resaltan el potencial de las técnicas de *machine learning* para fortalecer la asignación de recursos en programas sociales. Su implementación impulsaría la modernización de la gestión pública, permitiendo una gestión de recursos económicos fundamentada en datos.

Palabras-clave: Transferencias Monetarias; Machine Learning; XGBoost; Optimización Bayesiana.

Predictive model for beneficiary households in cash transfer programs: a comparison of machine learning techniques

Abstract: Cash transfer programs are a key tool for reducing poverty and improving the well-being of vulnerable households in developing countries. However, the accurate selection of beneficiaries remains a challenge. This study evaluates different machine learning techniques to predict participation in the Juntos program in Peru, using data from the 2023 National Household Survey (ENAHO). Models such as logistic regression, decision trees, support vector machine, gradient boosting machine, random forest, LightGBM, XGBoost, and CatBoost were compared. The results show that XGBoost achieves the best performance in beneficiary classification. These findings highlight the potential of machine

learning techniques to enhance the allocation of resources in social programs. Their implementation would drive the modernization of public management, enabling data-driven economic resource management.

Keywords: Cash Transfers; Machine Learning; XGBoost; Bayesian Optimization.

1. Introducción

En la actualidad, los programas de transferencias monetarias se han consolidado como una herramienta esencial para combatir la pobreza y mejorar el bienestar de los hogares vulnerables, especialmente en países en desarrollo (CAF, 2018). Un ejemplo destacado es el programa Juntos implementado en Perú desde el año 2005, el cual otorga subsidios en efectivo condicionados al cumplimiento de requisitos como enviar a los niños a la escuela o asistir a controles de salud. La efectividad de estas intervenciones radica en su capacidad para llegar a los hogares con mayores carencias. Por lo tanto, contar con modelos predictivos que permitan identificar a los potenciales beneficiarios resulta crucial para optimizar la focalización y la distribución eficiente de los fondos públicos destinados a este programa.

Estas iniciativas han demostrado un impacto positivo en la reducción de la pobreza y en la mejora del bienestar en economías emergentes (Banerjee & Duflo, 2019). En este contexto, diversas investigaciones han implementado algoritmos avanzados para abordar el desafío de identificar a los beneficiarios de manera precisa, como árboles de decisión (Lee, Lessler, & Stuart, 2010; Linden & Yarnold, 2018), gradient boosting machine (Autenrieth et al., 2021; Maciel & Duarte, 2023; Tu, 2019) y bosque aleatorio (Goller et al., 2020; Nufus et al., 2024; Peñarreta & Armas, 2024; Watkins et al., 2013; Zhao et al., 2016). Estas técnicas, reconocidas por su capacidad para modelar relaciones complejas y no lineales, han encontrado aplicaciones relevantes en áreas como salud, economía y educación, donde es fundamental identificar con precisión a los beneficiarios.

El presente estudio tiene como objetivo comparar una serie de modelos predictivos para identificar a los hogares beneficiarios del programa de transferencias monetarias condicionadas Juntos, haciendo uso de técnicas de *machine learning* como: regresión logística, árboles de decisión, máquina de soporte vectorial, *gradient boosting machine,* bosque aleatorio*, LightGBM, XGBoost* y *CatBoost*. Se seleccionará el modelo con mayor desempeño que permita mejorar la focalización de este programa y fortalecer la gestión de recursos económicos por parte del gobierno peruano.

Este enfoque no solo busca perfeccionar la distribución de las ayudas económicas, sino que también pretende maximizar el impacto social al garantizar que los beneficios lleguen a quienes más los necesitan (Crespo, 2020). En este contexto, evaluar la efectividad de diferentes algoritmos de *machine learning* resulta fundamental para la toma de decisiones informadas y basadas en evidencia.

El documento está estructurado de la siguiente manera: la primera sección introduce el contexto y los objetivos del estudio. La segunda sección detalla la revisión de literatura relevante. En la tercera sección presenta la metodología empleada. Posteriormente, se analizan los resultados, y finalmente, se exponen las conclusiones.

2. Revisión de la literatura

Una serie de estudios han resaltado el potencial de los modelos predictivos de *machine learning* en la selección de beneficiarios para programas sociales. Tal es el caso del aporte realizado por Chen (2018) sobre el programa Progresa en México, donde comparó algoritmos como bosque aleatorio y *AdaBoost* con modelos econométricos. Los resultados evidencian un rendimiento superior de las técnicas de *machine learning* en la predicción de asistencia escolar. No obstante, cuando los datos disponibles son limitados, los modelos econométricos estructurales pueden ofrecer una mayor precisión, subrayando la importancia de adaptar las metodologías a las características específicas del contexto.

En un trabajo más reciente, Dietrich et al. (2024), evidenciaron que métodos como *XGBoost* presentan ventajas significativas frente a enfoques tradicionales en la identificación de hogares en situación de pobreza. Sin embargo, los autores también advierten que la presencia de sesgos en los datos puede limitar el desempeño de estos modelos, lo que subraya la necesidad de enfoques más robustos.

En ese sentido, Aiken et al. (2023) demostraron que al combinar datos no tradicionales, como los registros de llamadas móviles, con los algoritmos de *machine learning*, es posible mejorar la detección de hogares elegibles para programas de transferencias monetarias en situaciones de emergencia. Este enfoque fue casi tan preciso como los métodos convencionales basados en activos y consumo, incluso superando a estos métodos cuando se añaden datos de encuestas, ofreciendo una alternativa más rentable, especialmente en contextos donde la recolección de información es difícil y costosa, como en zonas afectadas por conflictos o ante una emergencia sanitaria.

Asimismo, estrategias de validación cruzada y ensambles de modelos, logran minimizar errores en los procesos de clasificación de hogares vulnerables (McBride & Nichols, 2018). En esta misma dirección, el estudio de Tito et al. (2023) refuerza la utilidad de las métricas de evaluación (precisión, exhaustividad) y las técnicas de limpieza de datos para optimizar modelos predictivos en ámbitos educativos.

En consecuencia, el presente trabajo se enmarca en esta línea de investigación, evaluando el desempeño de algoritmos de *machine learning* que contribuyan a la mejora en el diseño e implementación de políticas sociales basadas en datos.

3. Metodología

En esta sección se detallan las etapas implementadas para seleccionar la técnica de *machine learning* más adecuada en la estimación de la probabilidad de hogares beneficiarios de las transferencias monetarias.

De acuerdo a la Figura 1, se inició recopilando la información estadística en los diferentes módulos disponibles en la Encuesta Nacional de hogares (ENAHO) del año 2023; luego se realizaron una serie de pasos correspondientes al preprocesamiento, como la detección de *outliers* y la imputación de registros faltantes. Durante la fase de ingeniería de características, las variables continuas fueron normalizadas y se codificaron las variables categóricas. Posteriormente, en la fase de entrenamiento se implementaron

siete algoritmos: regresión logística, árboles de decisión, máquina de soporte vectorial, *gradient boosting machine,* bosque aleatorio*, LightGBM, XGBoost* y *CatBoost*. La optimización de hiperparámetros se realizó mediante un enfoque de optimización bayesiana. Seguidamente, los modelos fueron evaluados utilizando métricas como *F1-score*, precisión y exhaustividad, seleccionando el que mostró mejor desempeño global.

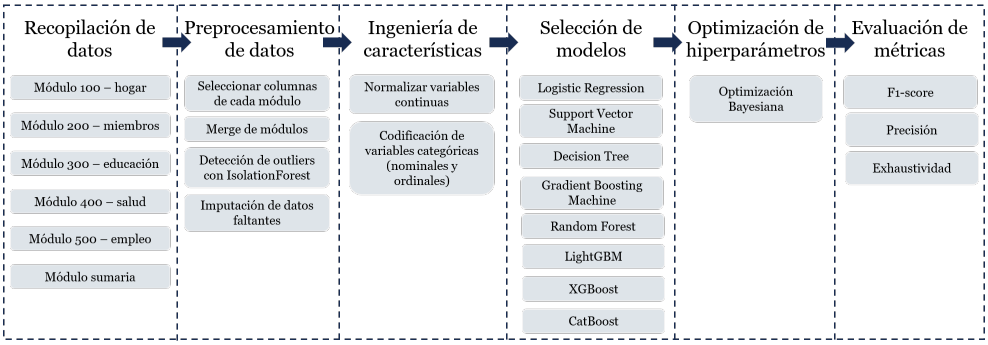

Figura 1 – Pasos para la estimación y selección del mejor modelo

3.1. Recopilación datos

La base de datos empleada proviene del Instituto Nacional de Estadística e Informática (INEI) que de forma anual lleva a cabo la ENAHO que se encuentra estructurada en 6 módulos. El módulo 100 presenta características físicas de los hogares, el módulo 200 expone información estadística sobre los miembros del hogar, el módulo 300 brinda información relacionada al nivel de educación alcanzado por los miembros del hogar, el módulo 400 incorpora datos relacionados a la salud de los entrevistados, el módulo 500 detalla la situación laboral de los miembros mayores a 14 años y finalmente el módulo sumaria, que calcula una serie de indicadores socioeconómicos a nivel hogar.

3.2. Preprocesamiento de datos e ingeniería de características

Para garantizar la calidad de los datos, se verificó la correlación entre las variables seleccionadas y los criterios oficiales de elegibilidad del programa Juntos, asegurando que capturen dimensiones clave de vulnerabilidad reconocidas por el marco normativo.

Respecto al preprocesamiento, se consolidaron los registros por hogar, utilizando principalmente las características del jefe de hogar, como edad y años de educación. Asimismo, se incluyeron variables generales del hogar, como el total de horas trabajadas por sus miembros y condiciones de la vivienda. También, se tomó en cuenta las características sociodemográficas como la región y el estrato al que pertenecen.

La Tabla 1 resume las variables seleccionadas para identificar hogares beneficiarios del programa Juntos. Adicionalmente, se presenta el módulo de origen, el tipo de variable y una breve descripción.

Los hogares considerados *outliers* según el nivel de ingresos de sus miembros fueron eliminados utilizando el método *Isolation Forest* desarrollado por Liu, Ting, y Zhou (2008). Como resultado, se obtuvo un conjunto de datos compuesto por 11464 hogares. Cabe destacar que un bajo porcentaje de datos requirió imputación. Para estos casos, como estrategia, se emplearon la mediana para las variables numéricas y la moda para las categóricas.

En lo que respecta a la ingeniería de variables, se normalizaron las variables numéricas (edad, educación, horas de trabajo) mediante el método *Min-Max Scaler*, el cual ajusta los valores al rango de 0 a 1. En cuanto a las variables categóricas, como estrato y región, fueron transformadas mediante *one-hot encoding,* generando columnas binarias que indican la presencia o ausencia de cada categoría.

Variable	Fuente	Tipo	Descripción
Juntos	Módulo Sumaria	Binaria	1 si el hogar es beneficiario del programa Juntos, 0 caso contrario
Edad	Módulo Miembros	Numérica	Edad en años del jefe de hogar
Educación	Módulo Educación	Numérica	Años de educación del jefe de hogar
Horas de trabajo	Módulo Empleo	Numérica	Horas de trabajo de todos los miembros del hogar
Vivienda propia	Módulo Hogar	Binaria	1 si cuenta con una vivienda propia, 0 caso contrario
Pared de ladrillo	Módulo Hogar	Binaria	1 si las paredes del hogar son de ladrillo, 0 caso contrario
Agua potable	Módulo Hogar	Binaria	1 si el hogar cuenta con agua potable, 0 caso contrario
Cocinar a leña	Módulo Sumaria	Binaria	1 si el hogar solo cuenta con cocina a leña, 0 caso contrario
Internet	Módulo Hogar	Binaria	1 si el hogar cuenta con servicio de internet, 0 caso contrario
Concreto	Módulo Hogar	Binaria	1 si el piso del hogar es de concreto, 0 caso contrario
Región	Módulo Sumaria	Categórica	1 si el hogar se ubica en la costa, 2 en la sierra y 3 en la selva
Estrato	Módulo Miembros	Ordinal	0 si el hogar se encuentra en la zona rural, 1 si es del estrato E, 2 si es del D y 3 si es del estrato C

Tabla 1 – Descripción de variables

3.3. Técnicas de machine learning

Para la estimación de la probabilidad de que un hogar participe en el programa Juntos, se enfrenta un desbalance en la variable objetivo donde la muestra cuenta con el 30.14% de hogares beneficiarios frente a 69.86% de no beneficiarios. Por ello, se implementaron

siete técnicas de *machine learning*: regresión logística, árboles de decisión, máquina de soporte vectorial, *gradient boosting machine,* bosque aleatorio*, LightGBM, XGBoost* y *CatBoost*.

Las cuatro primeras se desarrollaron con *Scikit-learn* (Pedregosa et al., 2011), mientras que las tres últimas se implementaron siguiendo sus respectivas documentaciones oficiales (Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018). A continuación, se presenta una breve descripción de los modelos seleccionados:

3.3.1. Regresión logística

Es un modelo estadístico ampliamente utilizado que estima la probabilidad de ocurrencia de un evento mediante una función logística, siendo especialmente valorado por su interpretabilidad en diversas disciplinas como economía, finanzas y ciencias sociales (Hosmer, Lemeshow, & Sturdivant, 2013). Se implementó utilizando el *solver liblinear*, adecuado para bases de datos con pocas muestras. Los hiperparámetros optimizados incluyeron C, que controla la regularización para prevenir sobreajuste, y *max_iter*, que ajusta el número máximo de iteraciones requeridas para alcanzar la convergencia del modelo.

3.3.2. Árboles de Decisión

Los árboles de decisión son modelos no paramétricos que representan procesos de decisión en forma jerárquica, permitiendo capturar relaciones no lineales. Aunque efectivos, su principal desafío radica en evitar el sobreajuste (Breiman et al., 1984). Para optimizar el rendimiento, se ajustaron los hiperparámetros: *max_depth*, que controla la profundidad del árbol; *min_samples_split*, que establece el número mínimo de observaciones para dividir un nodo; *min_samples_leaf*, que define el tamaño mínimo de hojas; y *max_features*, que selecciona la proporción de variables evaluadas en cada división.

3.3.3. Máquina de Soporte Vectorial

Es un modelo de clasificación supervisada que busca encontrar el hiperplano que mejor separa las clases, maximizando el margen entre ellas (Cortes & Vapnik, 1995). En esta investigación se utilizó *kernel RBF*, optimizando los hiperparámetros *C* y *gamma*. El parámetro *C* regula la penalización por errores en el entrenamiento, controlando el equilibrio entre sobreajuste y generalización. Por su parte, γ determina la influencia de cada muestra en la función de decisión, afectando la forma y complejidad de la frontera de separación.

3.3.4. Gradient Boosting Machine

Es un método de ensamble basado en árboles de decisión débiles entrenados secuencialmente, donde cada iteración corrige errores previos minimizando una función de pérdida mediante descenso de gradiente (Friedman, 2001). Este enfoque destaca por su capacidad para manejar datos desbalanceados y su alto rendimiento predictivo. Se optimizaron los hiperparámetros: *max_depth* para controlar la complejidad de los

árboles, *min_samples_split* para definir las divisiones mínimas y *max_features* para garantizar diversidad en las divisiones.

3.3.5. Bosque Aleatorio

Es un modelo de ensamble basado en la construcción de varios árboles de decisión. Cada árbol se entrena con una muestra aleatoria del conjunto de datos y un subconjunto aleatorio de las variables. La predicción se realiza mediante el voto mayoritario entre todos los árboles, lo que reduce la varianza y mejora la capacidad de generalización del modelo (Breiman, 2001). Se consideraron los hiperparámetros: *max_depth*, que regula la profundidad de los árboles; *min_samples_split*, para dividir nodos con un número mínimo de observaciones y *max_features* que selecciona un subconjunto de variables en cada división.

3.3.6. LightGBM

Es un algoritmo de *boosting* basado en árboles de decisión que optimiza la velocidad y eficiencia del entrenamiento mediante histogramas y reducción del tamaño de los datos, sin sacrificar precisión. Su capacidad para manejar grandes conjuntos de datos y características categóricas lo hace especialmente útil en aplicaciones de *machine learning* modernas (Ke et al., 2017).

Se optimizaron los siguientes hiperparámetros: *num_leaves*, que controla la complejidad de los árboles, *max_depth* que limita la profundidad máxima, *feature_fraction* que selecciona una fracción de características para cada iteración, *bagging_fraction* que define la proporción de datos usada en cada paso, *min_split_gain* que regula la ganancia mínima para dividir nodos y *min_child_weight*, que restringe el tamaño mínimo de hojas.

3.3.7. XGBoost

Es también un algoritmo de *boosting* que combina precisión, velocidad y eficiencia computacional mediante optimizaciones como el uso de memoria fuera del núcleo y la paralelización. Su versatilidad lo ha consolidado como una herramienta líder en competencias de *machine learning* (Chen & Guestrin, 2016). En este estudio, se ajustaron los hiperparámetros: *max_depth* que limita la profundidad de los árboles y controla el sobreajuste; *gamma* que establece la ganancia mínima requerida para dividir un nodo; *colsample_bytree* que selecciona una fracción de características para construir cada árbol y *subsample* que define la proporción de datos utilizada para cada árbol, favoreciendo la robustez y diversidad del modelo.

3.3.8. CatBoost

Es un algoritmo de *boosting* que se destaca por su manejo eficiente de variables categóricas sin necesidad de preprocesamiento, lo que minimiza el riesgo de sobreajuste y mejora el rendimiento en tareas supervisadas con técnicas avanzadas de regularización (Prokhorenkova et al., 2018).

Se optimizaron los hiperparámetros: *depth* que define la profundidad máxima de los árboles; *l2_leaf_reg* que ajusta la fuerza de regularización, *bagging_temperature* que controla la aleatoriedad en el muestreo de datos y *border_count* que determina la cantidad de divisiones usadas para variables continuas, mejorando su precisión en conjuntos de datos complejos.

3.4. Optimización de hiperparámetros

Como parte del proceso de entrenamiento, el conjunto de datos fue dividido en un 80% para el entrenamiento del modelo y un 20% para *test*, donde se evaluaron las métricas que permitieron seleccionar al mejor modelo. Dado que la variable a predecir se encontró desbalanceada, se tomó el *F1-score* como métrica principal ya que combina de forma equilibrada la precisión y la exhaustividad, lo que permitió evaluar el rendimiento del modelo considerando tanto los falsos positivos como los falsos negativos (Abhishek & Abdelaziz, 2023).

La optimización de hiperparámetros se llevó a cabo mediante validación cruzada con *10-folds*, utilizando la función *BayesianOptimization* desarrollado por Gardner et al. (2014). Este enfoque permitió identificar configuraciones óptimas al predecir el rendimiento del modelo y ajustar iterativamente los hiperparámetros. La Tabla 2 muestra los hiperparámetros optimizados y sus rangos para cada algoritmo.

Modelo	Parámetro	Rango
Regresión Logística	C	(0.0001, 10)
	max_iter	(50, 300)
Árboles *de decisión*	max_depth	(3, 30)
	min_samples_split	(20, 150)
	min_samples_leaf	(3, 2)
	max_feature	(0.1, 0.999)
Máquina de Soporte Vectorial	C	(0.01, 10)
	gamma	(0.001, 1)
Gradient boosting machine	max_depth	(3, 30)
	min_samples_split	(3, 30)
	max_features	(0.1, 0.999)
Bosque Aleatorio	max_depht	(3, 30)
	max_features	(0.1, 0.999)
	min_samples_split	(3, 25)
LightGBM	num_leaves	(20, 100)
	max_depth	(3, 15)
	feature_fraction	(0.7, 0.9)
	bagging_fraction	(0.7, 0.9)
	min_split_gain	(0.001, 0.5)
	min_child_weight	(5, 50)

Modelo	Parámetro	Rango
XGBoost	max_depth	(3, 15)
	gamma	(0, 3)
	colsamples_bytree	(0.5, 1)
	subsample	(0.2, 0.8)
CatBoost	depth	(4, 10)
	l2_leaf_reg	(1, 10)
	bagging_temperature	(0, 1)
	border_count	(32, 255)

Tabla 2 – Rango de hiperparámetros a optimizar

3.5. Métricas para la evaluación de algoritmos de machine learning

La evaluación del desempeño de los modelos de clasificación binaria requiere métricas específicas que capturen la capacidad predictiva del modelo. Entre las métricas que se evaluaron en este trabajo se encuentran:

Formato de una lista con marcas:

- *F1-score*: es la media armónica entre precisión y exhaustividad, especialmente útil en conjuntos de datos desbalanceados. Brinda una métrica que balancea la capacidad del modelo para identificar correctamente los casos positivos y evitar falsos positivos (Sokolova & Lapalme, 2009).
- *Precisión*: mide la proporción de casos positivos correctamente identificados por el modelo respecto al total de predicciones positivas realizadas. Es fundamental en contextos donde los falsos positivos tienen un mayor costo, como asignar recursos económicos limitados a hogares no elegibles.
- *Exhaustividad*: también conocida como la tasa de verdaderos positivos, mide la proporción de casos positivos correctamente identificados por el modelo respecto al total de casos positivos reales. Esta métrica es importante en contextos donde los falsos negativos tienen consecuencias sociales graves, como excluir hogares elegibles del programa Juntos.
- *Exactitud*: es una métrica ampliamente utiliza ya que mide la proporción de predicciones correctas (tanto positivas y como negativas) realizadas por el modelo respecto al total de casos evaluados.

4. Resultados

En la evaluación del desempeño de los modelos, se utilizó validación cruzada con 10 particiones para abordar el problema de desbalance de clases, donde la clase minoritaria corresponde a los hogares beneficiarios del programa Juntos. Este diseño permitió estimar métricas robustas y confiables en términos de rendimiento de clasificación.

La Figura 2 ilustra el *F1-score* obtenido para cada modelo en los *10 k-folds*. Dado el desequilibrio en los datos, esta métrica fue priorizada debido a su capacidad para

balancear precisión y exhaustividad. En términos generales, los modelos *XGBoost, LightGBM* y *CatBoost* mostraron un desempeño más consistente, destacándose sobre el resto.

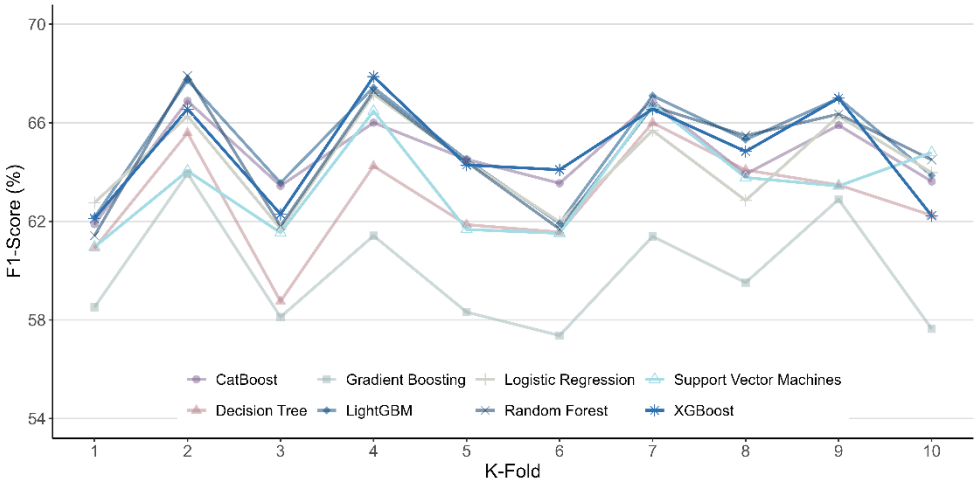

Figura 2 – Métrica F1-score en los 10 k-folds

De forma complementaria, las Figuras 3 y 4 presentan los resultados para las métricas de precisión y exhaustividad, respectivamente. El modelo *Gradient Boosting Machine* obtuvo la mayor precisión promedio, mientras que Regresión Logística destacó por alcanzar la mayor exhaustividad (81.33%). Sin embargo, estos modelos no lograron un equilibrio ideal entre ambas métricas, lo que impactó negativamente en su *F1-score*.

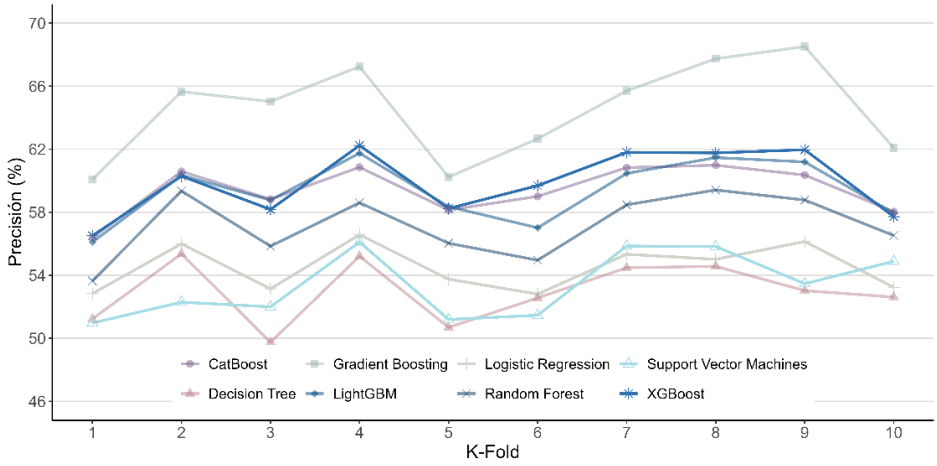

Figura 3 – Métrica Precisión en los 10 k-folds

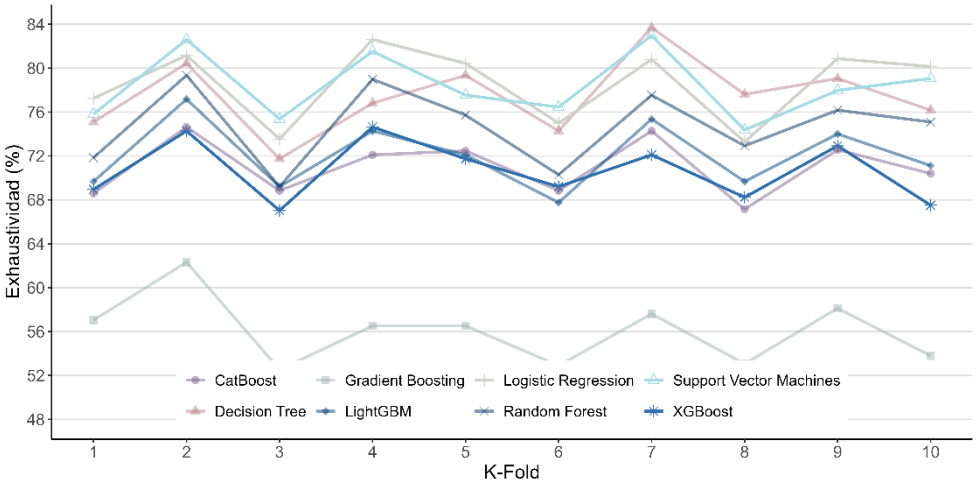

Figura 4 – Métrica Exhaustividad en los 10 k-folds

Para seleccionar el modelo óptimo, los hiperparámetros ajustados durante la etapa de validación cruzada fueron evaluados en el conjunto de datos de *test*. La Tabla 3 resume el desempeño de los modelos implementados, considerando métricas como *F1-score*, precisión, exhaustividad y exactitud.

Modelo	**F1-score (%)**	**Precisión (%)**	**Exhaustividad (%)**	**Exactitud (%)**
Regresión logística	65.42	54.72	**81.33**	74.10
Árboles de decisión	63.85	54.25	77.57	73.53
Máquina de Soporte Vectorial	63.99	52.94	80.90	72.57
Gradient boosting machine	62.84	**64.72**	61.07	78.24
Bosque Aleatorio	66.21	58.57	76.12	76.58
LightGBM	67.44	60.62	75.98	77.89
XGBoost	**67.54**	61.55	74.82	**78.33**
CatBoost	67.23	60.65	75.40	77.85

Tabla 3 – Comparación de modelos sobre el dataset de test

El modelo *XGBoost* fue identificado como el más adecuado, alcanzando un F1-score de 67.54 %, lo que refleja un balance óptimo entre precisión (61.55 %) y exhaustividad (74.82 %). Aunque la Regresión Logística obtuvo la mayor exhaustividad (81.33 %), su baja precisión resultó en un menor *F1-score*. De manera similar, *Gradient Boosting Machine* destacó en precisión (64.72 %), pero a expensas de una exhaustividad considerablemente menor.

Estos resultados evidenciaron que el modelo *XGBoost* logra el mejor rendimiento global al equilibrar métricas clave, lo que lo hace particularmente adecuado para predecir la probabilidad de ser beneficiario del programa Juntos en un contexto de datos desbalanceados.

Asimismo, el análisis de interpretabilidad del modelo mediante valores *SHAP (SHapley Additive exPlanations)* propuesto por Lundberg y Lee (2017), permite identificar las variables con mayor impacto sobre la probabilidad de pertenecer al programa Juntos. Como se muestra en la Figura 5, características como residir en una zona rural, tener un bajo nivel educativo y utilizar cocina a leña se asocian con un aumento significativo en la probabilidad estimada por el modelo, lo que sugiere que estas condiciones son altamente predictivas de la elegibilidad al programa. En contraste, características asociadas a mejores condiciones de vida, como el acceso a agua potable o vivir en viviendas con paredes de ladrillo, contribuyen negativamente a dicha probabilidad. Además, variables como edad y región de residencia mostraron un menor impacto relativo. Estos hallazgos no solo respaldan la capacidad del modelo para capturar patrones relevantes en poblaciones vulnerables, sino que también ofrecen evidencia empírica sobre los determinantes sociales que podrían estar influyendo en los criterios de focalización del programa.

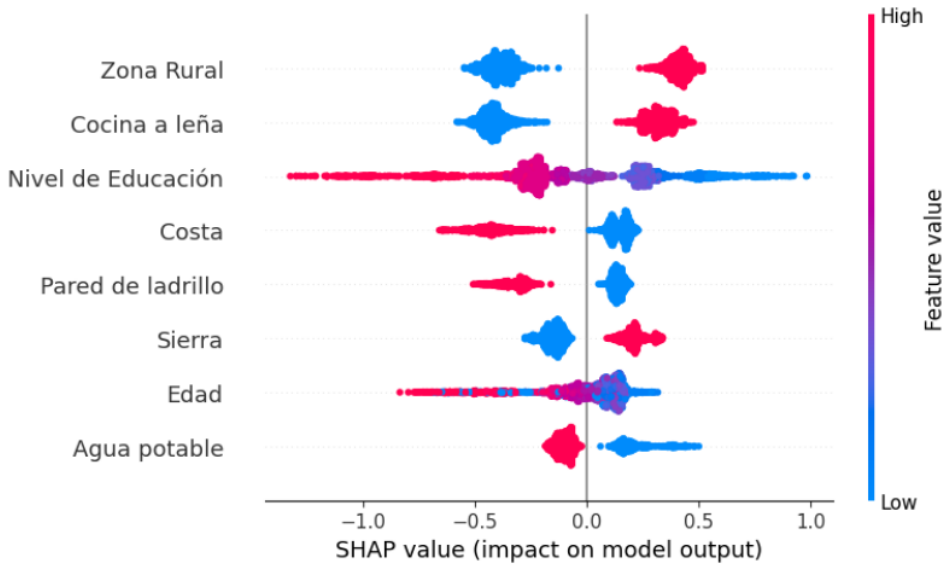

Figura 5 – Shap values del modelo XGBoost

5. Conclusiones

El presente estudio evaluó distintas técnicas de *machine learning* para predecir la probabilidad de ser beneficiario del programa Juntos en un contexto de datos desbalanceados. Para abordar este desafío, se aplicaron métodos como validación cruzada

y optimización bayesiana para encontrar la mejor combinación de hiperparámetros, priorizando el *F1-score* como métrica clave al equilibrar precisión y exhaustividad. Este enfoque es especialmente relevante en la gestión de programas sociales, donde la clasificación errónea de hogares elegibles puede generar costos sociales significativos.

Los resultados indicaron que *XGBoost* presentó el mejor desempeño, lo que resalta su capacidad para identificar beneficiarios en programas sociales. Su integración en procesos de clasificación automatizada podría contribuir a la modernización de la gestión pública al mejorar la distribución de transferencias monetarias con criterios basados en datos.

No obstante, la implementación de estos modelos requiere considerar desafíos éticos fundamentales: la posible amplificación de sesgos históricos presentes en los datos y la necesidad de mecanismos de auditoría que garanticen equidad en las decisiones automatizadas, particularmente para poblaciones vulnerables

Este estudio aporta evidencia sobre la utilidad de las técnicas de *machine learning* en la selección de beneficiarios, proporcionando un enfoque que puede fortalecer la toma de decisiones en políticas públicas. No obstante, su capacidad predictiva podría potenciarse con la integración de información macroeconómica regional y datos geoespaciales, lo que permitiría capturar dinámicas socioeconómicas más precisas y adaptar mejor la identificación de beneficiarios a distintos contextos territoriales.

Como futuras líneas de investigación, se recomienda explorar arquitecturas de *deep learning* y modelos híbridos que permitan capturar patrones más complejos en los datos. Esto permitiría desarrollar sistemas más adaptativos y precisos en la identificación de beneficiarios, especialmente en escenarios de alta incertidumbre.

Agradecimientos

Al Grupo de Investigación Ágora (HUM-648) de la Universidad d Huelva (España) por su asesoramiento científico-metodológico y su cooperación institucional.

Referencias

Abhishek, K., & Abdelaziz, M. (2023). *Machine Learning for Imbalanced Data: Tackle imbalanced datasets using machine learning and deep learning techniques*. Packt Publishing.

Aiken, E. L., Bedoya, G., Blumenstock, J. E., & Coville, A. (2023). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *Journal of Development Economics*, 161. https://doi.org/10.1016/j.jdeveco.2022.103016

Autenrieth, M., Levine, R. A., Fan, J., & Guarcello, M. A. (2021). Stacked ensemble learning for propensity score methods in observational studies. *Journal of Educational Data Mining*, 13(1). https://doi.org/10.5281/zenodo.5048425

Banerjee, A., & Duflo, E. (2019). *Good Economics for Hard Times: Better Answers to Our Biggest Problems* (1st Ed.). Public Affairs.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). Introduction To Tree Classification. En J. Kimmel & A. Cava (Eds.). *Classification and Regression Trees*, (pp. 27–72). Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 13-17). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chen, T. S. (2018). *Evaluating Conditional Cash Transfer Policies with Machine Learning Methods*. https://arxiv.org/abs/1803.06401

Crespo, C. (2020). Two become one: Improving the targeting of conditional cash transfers with a predictive model of school dropout. *Economia*, 21(1). https://doi.org/10.1353/eco.2020.0011

Corporación Andina de Fomento. (2018). Programas de transferencias monetarias condicionadas de dinero en efectivo: ¿solución mágica para mejorar la salud y la educación de las personas?

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. https://doi.org/10.1007/BF00994018

Dietrich, S., Malerba, D., & Gassmann, F. (2024). Predicting social assistance beneficiaries: On the social welfare damage of data biases. *Data & Policy*, 6(e3). https://doi.org/10.1017/dap.2023.38

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189 – 1232. https://doi.org/10.1214/aos/1013203451

Gardner, J. R., Kusner, M. J., Xu, Z. E., Weinberger, K. Q., & Cunningham, J. P. (2014). Bayesian optimization with inequality constraints. In *Proceedings of the 31st International Conference on Machine Learning*, (pp. 937–945). JMLR

Goller, D., Lechner, M., Moczall, A., & Wolff, J. (2020). Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. *Labour Economics*, 65. https://doi.org/10.1016/j.labeco.2020.101855

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Interpretation of the Fitted Logistic Regression Model. In D. Balding, N. A. C. Cressie, G. M. Fitzmaurice, H. Goldstein, I. M. Johnstone, G. Molenberghs, D. W. Scott, A. F. M. Smith, R. S. Tsay & S. Weisberg (ed.). *Applied Logistic Regression* (pp. 49–88). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118548387.ch3

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. 3149-3157). Curran Associates Inc.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3), 337–346. https://doi.org/https://doi.org/10.1002/sim.3782

Linden, A., & Yarnold, P. R. (2018). Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice*, 24(2). https://doi.org/10.1111/jep.12859

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, (pp. 413–422). https://doi.org/10.1109/ICDM.2008.17

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Maciel, F. A., & Duarte, D. (2023). The impact of cash transfer participation on unhealthy consumption in Brazil. *Health Policy OPEN*, 4. https://doi.org/10.1016/j.hpopen.2022.100087

McBride, L., & Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *World Bank Economic Review*, 32(3). https://doi.org/10.1093/wber/lhw056

Nufus, R. D. S., Susetyo, B., & Sartono, B. E.. (2024). Exploring the Potential Impact of Ginger Consumption on the Duration of COVID-19 Recovery: A Propensity Score Matching using Random Forest. In *IOP Conference Series: Earth and Environmental Science*, 1359(1). https://doi.org/10.1088/1755-1315/1359/1/012139

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel V. Thirion, B., Grisel, O., Blondel, M., Prettenhofer P. Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Peñarreta, M., & Armas, R. (2024). Educación financiera en jóvenes universitarios ecuatorianos: Una aplicación de Machine Learning. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (E71), 528 – 542.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 2018-December.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. In *Information Processing and Management*, (pp. 427–437). https://doi.org/10.1016/j.ipm.2009.03.002

Tito, A. E. A., Condori, B. O. H., & Vera, Y. P. (2023). Comparative analysis of Machine Learning Techniques for the prediction of cases of university dropout. RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, (e51), 84-98. https://doi.org/10.17013/risti.51.84-98

Tu, C. (2019). Comparison of various machine learning algorithms for estimating generalized propensity score. *Journal of Statistical Computation and Simulation*, 89(4). https://doi.org/10.1080/00949655.2019.1571059

Watkins, S., Jonsson-Funk, M., Brookhart, M. A., Rosenberg, S. A., O'Shea, T. M., & Daniels, J. (2013). An empirical comparison of tree-based methods for propensity score estimation. *Health Services Research*, 48(5). https://doi.org/10.1111/1475-6773.12068

Zhao, P., Su, X., Ge, T., & Fan, J. (2016). Propensity score and proximity matching using random forest. *Contemporary Clinical Trials*, 47. https://doi.org/10.1016/j.cct.2015.12.012