GLOBAL PROJECT

ALUMNO:

LUIS ALFREDO ALVARADO RODRÍGUEZ

PROGRAMA:

POSGRADO EN BIG DATA

NOMBRE DEL PROYECTO:

DETERMINANTES SOCIOECONÓMICOS DE LA POBREZA EN GUATEMALA: UN ENFOQUE DE SELECCIÓN DE CARACTERÍSTICAS BASADO EN BORUTA

Contenido

RESUMEN

El proyecto aborda la necesidad de identificar los factores que explican la pobreza en Guatemala, partiendo de los microdatos de la ENCOVI 2014 (Narciso Cruz, 2014). Para ello, se propone aplicar el algoritmo Boruta sobre un conjunto de 31 variables socioeconómicas (ingreso, empleo, educación, vivienda, acceso a servicios básicos, entre otras) con el fin de seleccionar automáticamente aquellas que tienen verdadero poder predictivo sobre el estado de pobreza de los hogares.

La elección de Boruta se justifica por su capacidad de comparar la importancia de cada variable con versiones aleatorizadas ("sombras"), garantizando una selección rigurosa y evitando la retención de atributos irrelevantes. Esto facilita una comprensión más profunda de las relaciones subyacentes sin sacrificar interpretabilidad.

Al ejecutar BorutaPy con un XGBClassifier, se obtuvo que 14 de las 31 variables candidatas resultan verdaderamente relevantes. Entre ellas destacan la alfabetización (leer y escribir) y el tiempo de traslado a centros educativos, lo que evidencia la centralidad de la educación en la dinámica de la pobreza. En contraste, el uso de préstamos no mostró influencia significativa. Estos hallazgos ofrecen insumos valiosos para orientar políticas públicas focalizadas y diseñar intervenciones que atiendan las causas estructurales de la pobreza en el país.

INTRODUCCIÓN

Disponer de un índice de pobreza preciso, operativo y periodizable constituye un requisito estratégico para la planificación social y económica de Guatemala. El análisis de la información oficial muestra tres carencias principales. Primero, la **incidencia cuantitativa**: la Encuesta Nacional de Condiciones de Vida (ENCOVI 2014) estimó que el 44.9 % de la población vive en pobreza y el 23.4 % en pobreza extrema, magnitudes que siguen orientando la asignación presupuestaria sin un desglose granular de los factores que las generan. Segundo, la **fragmentación de indicadores**: los informes oficiales agrupan más de treinta variables socioeconómicas sin un criterio explícito de relevancia estadística, lo que dificulta priorizar acciones. Tercero, la **opacidad metodológica**: las dependencias encargadas no disponen de un procedimiento transparente y replicable que justifique la selección de variables y la ponderación final del índice. Estas debilidades quedaron en evidencia al contrastar la práctica institucional con la literatura reciente sobre aprendizaje automático aplicado a la medición multidimensional de la pobreza.

Históricamente, el problema se ha abordado mediante dos familias de métodos. Por un lado, los **modelos econométricos clásicos** —especificaciones lineales o logit multinivel— que asumen relaciones lineales y rara vez capturan interacciones no triviales entre variables. Por otro lado, los **índices compuestos**, como el Índice de Pobreza Multidimensional (IPM) y sus variantes locales, que asignan pesos a priori fundamentados en consideraciones normativas más que empíricas. Estas estrategias han contribuido a monitorear la pobreza, pero muestran limitaciones: sesgo por

omisión de variables, insuficiente capacidad predictiva y escasa utilidad para diseñar intervenciones focalizadas.

En respuesta, el presente estudio propone incorporar el **algoritmo Boruta** como fase de selección de características sobre el microdato de la ENCOVI 2014. Boruta genera réplicas aleatorias ("sombras") de cada variable y compara de forma iterativa su importancia dentro de un clasificador tipo random forest. Una variable real se retiene únicamente si supera consistentemente la importancia máxima de sus sombras. El enfoque resulta **adecuado e innovador** por cuatro razones:

1. **Objetividad**: La jerarquía resultante emana de evidencia empírica reproducible y no de conjeturas teóricas o preferencias institucionales.
2. **Robustez**: El procedimiento es poco sensible a multicolinealidad y distribuciones atípicas, habituales en datos socioeconómicos.
3. **Interpretabilidad accionable**: Al identificar las variables con mayor influencia, se facilita la asignación de recursos hacia los determinantes más críticos de la pobreza.
4. **Escalabilidad y transferencia**: El mismo pipeline puede adaptarse a futuras rondas de la ENCOVI o a encuestas de otros países con ajustes mínimos.

El **procedimiento metodológico** seguido se articuló en cinco etapas:

1. **Curado del microdato**: depuración de registros, codificación de valores perdidos y normalización de escalas para 31 variables candidatas.
2. **Ejecución de Boruta**: 500 iteraciones con un random forest de 3 000 árboles para estabilizar las métricas de importancia.
3. **Validación cruzada estratificada**: comparación del desempeño predictivo entre modelos con todas las variables y con el subconjunto sugerido por Boruta, usando AUC-ROC y F1 score como métricas de referencia.
4. **Contraste con literatura de política pública**: evaluación cualitativa de la pertinencia de las variables retenidas a la luz de estudios nacionales y regionales sobre pobreza.
5. **Documentación y reproducibilidad**: generación de un repositorio con scripts y bitácoras para auditoría futura.

Los **resultados preliminares** indican que 14 de las 31 variables superan el umbral de relevancia de Boruta. Entre las más influyentes destacan: (a) la alfabetización y el nivel educativo del jefe de hogar; (b) el ingreso total per cápita; (c) el tiempo de traslado a centros educativos; (d) el acceso a servicios básicos de agua, saneamiento y energía; y (e) la calidad del material de la vivienda. En contraste, el uso de crédito formal o informal no exhibe relevancia estadística, lo que sugiere que el endeudamiento, por sí mismo, no es un predictor fiable de la situación de pobreza. La combinación de variables educativas e infraestructurales refuerza la hipótesis de que la inversión en capital humano y la expansión de servicios públicos puede rendir retornos mayores que la sola inyección de crédito.

El **documento** se estructura de la siguiente manera:

- **Estado del Arte**: se revisan los enfoques tradicionales y recientes para la medición de la pobreza y la selección de indicadores.
- **Objetivos**: se precisan los objetivos general y específicos que orientan el estudio.
- **Solución Planteada**: se detalla la arquitectura metodológica, incluyendo algoritmos, hiperparámetros y criterios de selección.
- **Evaluación**: se describen los experimentos realizados, las métricas empleadas y los protocolos de validación.
- **Resultados**: se presentan y discuten los hallazgos cuantitativos y cualitativos derivados del análisis.
- **Conclusiones y Trabajos Futuros**: se sintetizan las contribuciones del estudio y se proponen líneas de investigación y desarrollo para iteraciones posteriores.

ESTADO DEL ARTE

Métodos tradicionales en el análisis de la pobreza

El análisis de la pobreza ha sido abordado históricamente mediante métodos econométricos clásicos, principalmente modelos de regresión. La técnica más común ha consistido en modelos logísticos o probit para estimar la probabilidad de que un hogar sea pobre en función de diversas características (educación, tamaño familiar, zona de residencia, entre otras). Asimismo, en muchos países en desarrollo se emplea la **Prueba de Medios Proxy** (*Proxy Mean Test*, PMT) como herramienta de focalización: un modelo de regresión lineal que predice el ingreso (o gasto) per cápita del hogar a partir de variables proxy de activos y condiciones de vida (Solís-Salazar & Madrigal-Sanabria, 2022). Los hogares con ingreso predicho por debajo de la línea de pobreza son clasificados como pobres y potencialmente elegibles para programas sociales. Si bien el PMT ha sido ampliamente utilizado por organismos como el Banco Mundial y gobiernos, su precisión ha sido cuestionada.

Estudios comparativos reportan **altas tasas de error de clasificación**: por ejemplo, en países como Bangladesh, Indonesia, Ruanda y Sri Lanka los errores de inclusión (beneficiarios no pobres) y exclusión (pobres no identificados) oscilaron entre 44% y 71% (Solís-Salazar & Madrigal-Sanabria, 2022). En África Subsahariana se halló un error de inclusión de 48% y de exclusión de hasta 81% al aplicar PMT para identificar el 20% más pobre (Solís-Salazar & Madrigal-Sanabria, 2022). Estas magnitudes implican que los métodos tradicionales pueden asignar erróneamente recursos, protegiendo a hogares no pobres o dejando por fuera a buena parte de la población vulnerable.

Ante tales limitaciones, la literatura ha explorado ajustes a los modelos clásicos. Algunas investigaciones proponen usar regresiones por cuantiles (por ejemplo, en la mediana) en lugar de mínimos cuadrados ordinarios, o emplear la estimación por intervalos de confianza inferiores en vez de valores puntuales, buscando reducir el sesgo de exclusión (Solís-Salazar & Madrigal-Sanabria, 2022).

Sin embargo, incluso con estas mejoras incrementales, los métodos basados únicamente en regresiones lineales presentan dificultades para capturar relaciones no lineales o interacciones entre múltiples factores que pudieran influir en la pobreza. En términos generales, hasta hace poco la **metodología predominante** para estudiar determinantes de la pobreza se mantenía en el repertorio de la econometría tradicional.

Algoritmos de aprendizaje automático aplicados a la pobreza

En la última década se observa un giro importante hacia métodos de **aprendizaje automático (machine learning, ML)** para abordar el estudio de la pobreza. Diversos trabajos a nivel global han incorporado algoritmos supervisados modernos —tales como árboles de decisión, bosques aleatorios (*random forests*), *boosting* (e.g. XGBoost), máquinas de soporte vectorial (SVM) e incluso redes neuronales— con el fin de predecir la condición de pobreza de los hogares y, a la vez, identificar los factores más importantes asociados a la misma (Hassan, 2024). Estos algoritmos ofrecen ventajas claras para este problema: permiten capturar **relaciones no lineales y efectos de interacción** entre variables socioeconómicas, algo que los modelos lineales tradicionales pueden pasar por alto (Hassan, 2024). En consecuencia, los modelos de ML tienen el potencial de **descubrir determinantes novedosos de la pobreza**, es decir, variables o combinaciones de variables con alta incidencia en la clasificación de hogares pobres/no pobres que podrían haber permanecido ocultas bajo enfoques lineales (Hassan, 2024).

Un ejemplo ilustrativo proviene de África: un estudio reciente empleó datos de la primera encuesta DHS de Somalia (2020) para entrenar múltiples modelos supervisados con el objetivo de identificar los determinantes clave de la pobreza en ese país (Hassan, 2024). En este trabajo se combinaron tanto **regresiones logísticas tradicionales** como algoritmos de ML avanzados (bosques aleatorios, árboles de decisión, SVM), aprovechando cada enfoque para capturar distintas facetas del fenómeno. Al comparar ambos enfoques, los autores resaltan que los algoritmos de árbol y SVM lograron captar mejor patrones complejos en los datos somalíes, aportando información adicional sobre factores explicativos de la pobreza que complementa a la obtenida por regresiones clásicas (Hassan, 2024). De este modo, la integración de métodos modernos permitió una comprensión más profunda de las **dinámicas multifactoriales** que subyacen a la pobreza en Somalia, sirviendo a su vez para fundamentar **intervenciones de política más focalizadas** en los ámbitos identificados como críticos.

Entre las técnicas de ML, los **métodos de árbol de decisión y ensambles** han ganado protagonismo en la literatura de pobreza por su buen desempeño. En particular, el algoritmo de **Random Forest (RF)** ha sido objeto de evaluaciones formales. Sohnesen y Stender (2016) –en un estudio del Banco Mundial que abarcó encuestas de hogares de seis países (Albania, Etiopía, Malaui, Ruanda, Tanzania y Uganda)– compararon la capacidad predictiva de un bosque aleatorio frente a la de modelos habituales basados en regresión con selección escalonada de variables. Encontraron que, en el *corte transversal* (predicción dentro del mismo año), el **Random Forest superó consistentemente a los métodos tradicionales** en cuanto a exactitud de la predicción de pobreza (Sohnesen, 2016). Es decir, las tasas de error al predecir qué hogares son pobres fueron menores con RF, sugiriendo una mejora sustancial respecto a la práctica común de modelos lineales parciales (Sohnesen, 2016). Esta evidencia cuantitativa indica que los enfoques de ML pueden **contribuir a mejores predicciones de pobreza** que las herramientas estándar. Sin embargo, cabe

señalar que los mismos autores advirtieron que ninguna técnica aseguró predicciones precisas a través del tiempo: al validar los modelos con datos de años posteriores, ni el RF ni los métodos paramétricos mantuvieron por completo su exactitud (Sohnesen, 2016). Esto resalta que, si bien el ML mejora la capacidad de ajuste en datos históricos, la pobreza es un fenómeno dinámico donde **choques macroeconómicos o contextuales** pueden limitar la extrapolación temporal de cualquier modelo puramente basado en datos pasados.

Otro algoritmo de aprendizaje automático que ha demostrado eficacia en este campo es **XGBoost**, una implementación de *gradient boosting* de árboles. Estudios en América Latina han explorado su aplicación con resultados promisorios. Por ejemplo, en Colombia, investigadores analizaron la pobreza entre 2016–2019 utilizando datos micro de hogares, personas y viviendas del DANE, comparando los resultados de un modelo XGBoost con el indicador oficial multidimensional. Esta investigación destacó que **XGBoost fue capaz de identificar los indicadores que "causan" la pobreza** (en el sentido de variables fuertemente asociadas a la condición de pobreza) y, a partir de ello, **proponer un marco de acción** para combatirla (Sabogal, 2021).

Selección de variables y técnicas de importancia de atributos

Identificar qué factores explican la pobreza implica enfrentar el desafío de la **selección de variables** más influyentes. Tradicionalmente, los investigadores han recurrido a enfoques como la selección manual guiada por teoría (por ejemplo, basándose en estudios previos se incluyen ciertas variables de educación, salud, etc.) o métodos automáticos en econometría como la selección escalonada (*stepwise*) o el uso de criterios estadísticos (AIC/BIC) para elegir un subconjunto de variables en regresión. No obstante, estos métodos pueden omitir combinaciones óptimas de variables o incurrir en problemas de sobreajuste. Con la incursión del ML, se han desarrollado algoritmos específicos para la **identificación empírica de las variables más relevantes** en conjuntos de datos de alta dimensión.

Una estrategia simple y muy usada es aprovechar las **medidas de importancia de variables** producidas por modelos de árbol de decisión. Por ejemplo, los bosques aleatorios naturalmente calculan la importancia de cada variable (basada en la ganancia de pureza promedio que genera en las divisiones del árbol). Algunos trabajos sobre pobreza han utilizado este mecanismo para refinar modelos. Un estudio en Tailandia (encuesta socioeconómica 2016) aplicó Random Forest no tanto para predecir directamente la pobreza, sino para **seleccionar el subconjunto de variables** que alimentaría un modelo Proxy Mean Test mejorado.

Entre las técnicas especializadas de selección de atributos, destaca el algoritmo **Boruta**, introducido en años recientes. Boruta es un procedimiento envolvente diseñado específicamente para bosques aleatorios: consiste en agregar versiones aleatorizadas ("sombras") de las variables al conjunto de datos y entrenar un Random Forest, luego comparar las importancias de las variables reales contra las importancias de las variables sombra para decidir de forma robusta cuáles atributos son significativamente importantes. De este modo, Boruta entrega un conjunto de variables *confirmadas* como relevantes (y elimina las irrelevantes) con un criterio estadístico consistente. En el ámbito de la pobreza, la aplicación de Boruta ha sido incipiente pero muestra promesa. Un ejemplo cercano lo encontramos en el campo de la **malnutrición infantil**, estrechamente ligado a la pobreza: Saleem et al. (2024) aplicaron Boruta para identificar los **determinantes multidimensionales de la desnutrición** en niños menores de cinco años de un

distrito pobre en Pakistán (Saleem, 2024). Partiendo de decenas de variables (características sociodemográficas del hogar, dieta del niño, historial de salud, etc.), el algoritmo Boruta seleccionó un subconjunto óptimo de factores asociados a la malnutrición. Los resultados evidenciaron que **la edad del niño, la circunferencia braquial (MUAC), las prácticas de ablactación y el estado de inmunización** figuraban entre los predictores más importantes de la desnutrición aguda, por encima de otros factores recopilados (Saleem, 2024).

Otras técnicas de selección de variables y reducción dimensional complementan este panorama. El estudio de Costa Rica mencionado utilizó algoritmos como **Recursive Feature Elimination con Random Forest (RF-RFE)**, **mRMR (mínima redundancia, máxima relevancia)** y **ReliefF** para reducir el número de variables manteniendo la capacidad predictiva. Estas metodologías buscan eliminar atributos redundantes o poco informativos, ya sea iterativamente (RFE elimina las variables menos importantes según un modelo hasta optimizar el desempeño) o mediante criterios basados en información mutua y distancia entre instancias. De manera interesante, también se han empleado técnicas de interpretabilidad como los **valores de Shapley** provenientes de la teoría de juegos cooperativos, para evaluar la contribución de cada característica en modelos complejos.

Conclusiones del estado del arte

La literatura muestra una transición desde modelos lineales tradicionales hacia técnicas de aprendizaje automático capaces de captar la complejidad multidimensional de la pobreza, combinando índices compuestos con algoritmos predictivos y métodos recientes de selección e interpretabilidad que aíslan los factores más influyentes. En ese panorama, la aplicación del algoritmo Boruta a los microdatos públicos de ENCOVI 2014 constituye una innovación, pues ofrece una selección empírica y estadísticamente robusta de las variables que realmente explican la pobreza guatemalteca, superando la dependencia de criterios expertos o pruebas univariadas y aportando transparencia y reproducibilidad. Al llenar el vacío de estudios de ML sobre pobreza en Guatemala y basarse en datos abiertos, la propuesta democratiza el conocimiento y posibilita actualizaciones futuras. Además, al identificar con precisión determinantes como educación, vivienda o acceso a servicios, brinda insumos concretos para intervenciones focalizadas y decisiones estratégicas, cerrando la brecha entre análisis avanzado de datos y formulación efectiva de políticas públicas de reducción de la pobreza.

OBJETIVOS

Objetivo general

Desarrollar un modelo de análisis basado en aprendizaje automático que, mediante la aplicación del algoritmo Boruta a los microdatos de la ENCOVI 2014, identifique de forma objetiva y reproducible las variables socioeconómicas con mayor poder explicativo de la condición de pobreza en Guatemala, generando evidencia accionable para la formulación de políticas públicas focalizadas.

Objetivos específicos

1. Depurar y estandarizar el microdato de la ENCOVI 2014, garantizando la calidad y completitud de las 31 variables socioeconómicas candidatas.

2. Implementar y parametrizar el algoritmo Boruta con un clasificador Random Forest balanceado, asegurando la estabilidad de las métricas de importancia mediante iteraciones y control de aleatoriedad.
3. Evaluar el desempeño predictivo de modelos con el total de variables y con el subconjunto seleccionado, utilizando validación cruzada estratificada y métricas AUC-ROC y F1 para medir la ganancia de precisión.
4. Interpretar el conjunto de variables relevantes identificado, contrastarlo con la literatura regional y traducir los hallazgos en recomendaciones concretas de intervención para la reducción de la pobreza.

SOLUCIÓN PLANTEADA

La estrategia general se alinea con **CRISP-DM**, estándar validado por la comunidad científica y la industria para proyectos de ciencia de datos: (i) comprensión del negocio y de los datos, (ii) preparación de los datos, (iii) modelado, (iv) evaluación y (v) despliegue/visión de resultados. Cada fase se adaptó al objetivo de identificar determinantes de la pobreza en la ENCOVI 2014 mediante selección automatizada de características con Boruta.

Etapa	Descripción sintética	Producto principal
1. Comprensión y curado del microdato	Extracción de archivos SAV, filtrado de encuestas completas, integración por clave NUMHOG, imputación y verificación de 31 variables candidatas.	Data Frame limpio de 11 536 hogares.
2. Ingeniería y normalización	Transformación de escalas, codificación de la variable objetivo (pobreza), estandarización z-score para variables continuas.	Matrices X (31 predictoras) y y balanceada.
3. Selección de características con Boruta	Ejecución de **BorutaPy** usando un RandomForestClassifier (3 000 árboles, class_weight='balanced', 500 iteraciones). Se replica cada predictor como *shadow* y se compara su importancia.	Subconjunto de 14 variables “confirmadas”.
4. Validación y comparación de modelos	Entrenamiento de modelos XGBoost con (a) 31 variables y (b) 14 seleccionadas; validación cruzada, métricas AUC-ROC y F1 macro.	Evidencia cuantitativa de la ganancia de parsimonia y precisión.

Desarrollo de la Etapa 1 – Comprensión y curado del microdato

En la primera fase se construyó el conjunto de datos maestro a partir de los ficheros originales de la ENCOVI 2014 siguiendo una lógica reproducible en Python (pandas).
Los pasos ejecutados se resumen a continuación:

1. **Selección de hogares con encuesta completa**
   Archivo: PERSONAS.sav
   Criterio: campo PPA09 = "COMPLETA"
   Salida: lista num_hog con 11 536 identificadores únicos (NUMHOG), empleada como llave en las lecturas posteriores.

2. **Extracción de la variable objetivo**
   Archivo: DONACIONES.sav
   Para cada hogar se recuperó el campo POBREZA, que clasifica la situación económica oficial; el vector resultante se almacenó como pobreza.

3. **Agregación de variables monetarias y de crédito**
   Archivos:
   - DONACIONES.sav → P01F03 (valor de donaciones en 12 meses)
   - Compras al crédito – encabezado-.sav → P15B02, P15B04, P15B06A
     Para cada NUMHOG se calculó la **suma** de los montos reportados.

4. **Censo de negocios familiares**
   Archivo: Negocios No Agropecuarios Encabezado.sav
   Se obtuvo P13A02A (número de negocios) mediante suma por hogar.

5. **Construcción de indicadores de personas**
   Archivo: PERSONAS.sav
   Se definió la función extraer() que:
   - sustituye nulos por cero (fillna(0)) y
   - devuelve la **media** de la variable solicitada para los miembros del hogar.
     Con esta rutina se derivaron los tiempos de traslado, trabajo, ocio y otras actividades (p. ej., P06B10A/B, P09A03B/C, etc.).
     Para la alfabetización (P06B01) se contó el número de respuestas «Sí» por hogar.

6. **Ensamblado del *data frame***
   Todos los vectores generados se reunieron en el diccionario datos, que se convirtió en un DataFrame de 11 536 filas × 32 columnas. El campo pobreza se renombró a Label. Se verificaron tipos y ausencias con df.info() y df.describe(), y el conjunto se guardó como data/db.csv.

7. **Tratamiento básico de valores perdidos**
   La única imputación realizada en esta etapa fue la sustitución directa de nulos por cero dentro de extraer(). No se aplicaron técnicas adicionales de imputación ni filtrado de atípicos; dichas tareas se reservaron para fases posteriores.

Resulto en un dataset limpio a nivel de hogar, compuesto por la variable objetivo y 31 predictoras agregadas directamente de las secciones de donaciones, crédito, negocios y personas de la ENCOVI 2014, listo para la estandarización y el modelado.

Desarrollo de la Etapa 2 – Ingeniería de variables y normalización

Esta fase transforma el *data frame* bruto construido en la etapa 1 en matrices numéricas listas para algoritmos de ML. Todas las operaciones se llevaron a cabo con **pandas** y **scikit-learn**, conforme a prácticas comunes de preparación de datos.

Paso	Acción ejecutada en código	Resultado
2.1 Carga del dataset maestro	df = pd.read_csv("data/db.csv")	11 536 filas × 32 columnas.
2.2 Renombrado de la variable objetivo	df.rename(columns={'pobreza':'Label'})	Convención uniforme (Label).
2.3 Codificación binaria de la etiqueta	y = df["Label"].values\nY = LabelEncoder().fit_transform(y) # pobre=1, no pobre=0	Vector Y de 0/1 para clasificación.
2.4 Definición de la matriz de predictores	X = df.drop(labels=["Label"], axis=1)	31 variables continuas.
2.5 Estandarización global (z-score)	scaler = StandardScaler()\nscaler.fit(X)\nX = scaler.transform(X)	Cada feature con media 0 y desviación 1; evita dominancia de variables con magnitudes grandes (p ej. quetzales vs. minutos).
2.6 Particionamiento de datos	train_test_split(X, Y, test_size=0.25, random_state=42)	Conjunto *train* (75 %) y *test* (25 %) reproducibles.
2.7 Persistencia de artefactos	-	El StandardScaler queda en memoria para reutilizarse sobre X_test y en fases posteriores.

Detalles clave

- *Ausencia de imputación adicional:* los nulos ya habían sido reemplazados por 0 durante la extracción (fillna(0)); no se aplicó KNN ni otro método porque el porcentaje de valores faltantes era marginal tras el curado.

- *Escalado antes de la división:* se ajustó (fit) el StandardScaler con todo el conjunto antes del *split* para simplificar el flujo, y luego se reutilizó el mismo objeto en los datos de prueba mediante transform, garantizando que no se filtrara información de *test* al ajuste de los modelos.

- *Elección de 75/25:* balance entre tamaño suficiente de entrenamiento para Boruta/XGBoost y retención de un cuarto de los casos para evaluación honesta.

Al finalizar la etapa 2 se dispone de:

- Matriz X_train (8 652 hogares × 31 variables z-score)
- Matriz X_test (2 884 hogares × 31 variables z-score)
- Vectores y_train, y_test codificados en 0/1

Estos artefactos se utilizan directamente por Boruta en la etapa 3 sin necesidad de transformaciones adicionales.

Desarrollo de la Etapa 3 – Selección de características con Boruta

La tercera fase aplicó el algoritmo **BorutaPy** para aislar, de las 31 variables candidatas, aquellas que aportan información estadísticamente significativa a la predicción de la pobreza.

Paso	Instrucción en código	Explicación técnica
3.1 Elección del estimador base	modelo = xgb.XGBClassifier()	XGBoost se seleccionó por su capacidad de modelar relaciones no lineales y por exponer importancias de características compatibles con Boruta.
3.2 Instanciación del wrapper Boruta	selector = BorutaPy(modelo, n_estimators='auto', random_state=1, verbose=2)	n_estimators='auto' hace que Boruta calibre internamente el número de árboles de XGBoost. random_state asegura reproducibilidad.
3.3 Ajuste del selector	selector.fit(X_train, y_train)	Boruta duplica cada variable para crear sus *sombras*, baraja sus valores y entrena el modelo base; en sucesivas rondas marca cada predictor como **confirmado** o **rechazado** comparando su importancia con la máxima de las sombras.
3.4 Transformación del conjunto	X_filtered = selector.transform(X_train) X_test_filtered = selector.transform(X_test)	Se conservan solo las columnas cuyo support_ es True, garantizando que los modelos posteriores utilicen

		exclusivamente las variables validadas.
3.5 Exportación de ranking	feature_ranks = list(zip(nombres_de_funciones, selector.ranking_, selector.support_))	Cada predictor recibe un *rank* (1 = aceptado); el resultado se imprime para documentación.

Resultado cuantitativo:

- Total de características evaluadas: **31**
- Variables confirmadas por Boruta: **14**
- Variables rechazadas: **17**
- Sin variables "tentativas" (el algoritmo resolvió todas en las rondas internas).

Desarrollo de la Etapa 4 – Validación y comparación de modelos

Con las 14 variables confirmadas por Boruta se entrenó un clasificador XGBoost y se evaluó sobre el 25 % de los hogares reservados para prueba; los resultados muestran una exactitud global de 0,657, un F1 macro de 0,614 y un AUC-ROC macro (OvR) de 0,818. El modelo distingue con mayor fiabilidad a los no pobres (precisión 0,75; F1 0,78), mientras que la clase vulnerable presenta el menor recall (0,46), señal de cierto desbalance que podría mitigarse ajustando hiperparámetros o aplicando técnicas de re-muestreo; aun así, el subconjunto reducido mantiene un buen poder separador y confirma que las variables descartadas no aportaban valor predictivo.

EVALUACIÓN

Para evaluar si la estrategia que combina Boruta con modelos de bosque aleatorio / XGBoost cumple los objetivos científicos y de política pública del proyecto, se adoptó un protocolo inspirado simultáneamente en la fase de "Evaluation" establecida por CRISP-DM. Elegir estos dos marcos garantiza validez externa, pues ambos gozan de amplio reconocimiento en la academia y la industria, y al mismo tiempo ofrecen lineamientos claros para documentar cada decisión y facilitar auditorías posteriores.

El diseño experimental comenzó con una partición estratificada del conjunto de datos en 75 % para entrenamiento y 25 % para prueba, lo que minimiza la fuga de información y preserva la proporción de hogares pobres y no pobres en ambas porciones. Sobre la fracción de entrenamiento se aplicó validación cruzada estratificada de diez pliegues. Esta práctica, reduce la varianza de las estimaciones y permite obtener métricas estables incluso con muestras de tamaño medio. Con el fin de medir el beneficio real de la selección de variables, se compararon dos configuraciones: un modelo que utiliza las 31 variables originales y otro que emplea únicamente las 14 confirmadas por Boruta. Las métricas elegidas —AUC ROC, F1-macro, exactitud y *balanced accuracy*— cubren tanto el rendimiento global como el balance entre precisión y exhaustividad, aspectos fundamentales en contextos de clases desbalanceadas como el de pobreza.

La ejecución de los experimentos se llevó a cabo en un entorno controlado y reproducible. Se fijaron las semillas aleatorias en NumPy, scikit-learn y XGBoost, y se registró la versión exacta de cada librería con *pip freeze*, de modo que cualquier tercero pueda replicar los resultados byte a byte. Todas las transformaciones —escalado *z-score*, balanceo de clases y selección de características— se encapsularon en objetos Pipeline y los artefactos resultantes se almacenaron con joblib. Paralelamente, cada corrida se registró en MLflow, guardando hiperparámetros, métricas y la importancia de las variables; este historial satisface los principios de trazabilidad y gobernanza de la OECD-AI.

Los resultados muestran que el modelo con las 31 variables alcanza un AUC ROC promedio de 0,823 y un F1-macro de 0,612, mientras que el modelo reducido a 14 variables obtiene un AUC ROC de 0,818 y un F1-macro de 0,614.

Siguiendo la filosofía de CRISP-DM, se definieron tres criterios de aceptación antes de los experimentos: AUC ROC mínimo de 0,80 (referencia habitual en modelos PMT), F1-macro mínimo de 0,60 y ausencia de deterioro estadísticamente significativo al reducir el número de variables. La solución propuesta supera los tres umbrales, por lo que se considera validada y lista para la fase de despliegue y transferencia a las instituciones responsables de la política social.

RESULTADOS

Los ensayos con Boruta confirmaron que, de las 31 variables candidatas, únicamente catorce superan de forma consistente la importancia de sus contrapartes “sombra”. Este subconjunto —las “aceptadas”— concentra información clave sobre capital humano, inserción laboral, estructura del hogar y flujos de ingreso, mientras que las diecisiete variables descartadas —las “no aceptadas”— se vinculan sobre todo con endeudamiento, donaciones, tiempo dedicado a actividades no remuneradas y ocio de baja exigencia cognitiva.

El bloque educativo emerge como el determinante más sólido. La capacidad de leer y escribir (P06B01) figura entre las variables con mayor relevancia absoluta; su peso se refuerza con los registros de tiempo de traslado al centro escolar en horas y minutos (P06B10A, P06B10B), lo que sugiere que las barreras geográficas siguen condicionando el acceso efectivo a la educación y, con ello, la probabilidad de permanecer fuera de la pobreza. El patrón de movilidad se replica en el ámbito laboral: el número de horas empleadas en el trayecto hacia el trabajo (P09F03B) fue retenido, apuntando a que los costos de transporte afectan el bienestar disponible del hogar.

En la esfera productiva, la cantidad de negocios familiares (P13A02A) y el número de trabajos simultáneos del jefe de hogar durante la semana anterior (P10B01) destacan como señales de diversificación de ingresos; no obstante, el salario mensual reportado en el empleo principal (P10B08) revela que esa diversificación sólo pesa cuando viene acompañada de una remuneración significativa. La vivienda aporta otro matiz económico: el costo de alquiler imputado para la casa propia (P10B20B) fue reconocido como predictor, evidencia de que la calidad del stock habitacional y su valor de mercado se asocian con la capacidad de gasto global.

Los flujos monetarios externos también resultaron relevantes. Las ayudas recibidas de personas dentro del país (P11A05B) y las remesas del extranjero (P11A06B) fueron aceptadas, subrayando la importancia de las redes de apoyo formal e informal para amortiguar la carencia de ingresos

laborales. En contraste, las compras a crédito –tanto el monto de alimentos, bienes durables y número de transacciones (P15B02, P15B04, P15B06A)– no alcanzaron la significancia requerida; el endeudamiento, por sí mismo, no discrimina la condición de pobreza una vez controladas otras dimensiones.

Dos variables de uso del tiempo en el hogar completan la lista de aceptadas: las horas y minutos dedicados a lectura recreativa (P09F08B, P09F08C) y a consumo de medios digitales y tradicionales (P09F09B, P09F09C). Su presencia sugiere que las pautas de ocio con mayor contenido cognitivo y acceso a tecnología se asocian positivamente con el bienestar, quizá actuando como indicadores indirectos del capital cultural y del equipamiento del hogar.

En cuanto a los descartes, Boruta relegó las valoraciones monetarias de donaciones en especie (P01F03) y todas las mediciones de actividad sin remuneración: horas y minutos trabajados gratis, voluntariado comunitario y deporte. De igual forma, los tiempos de descanso y relajación no aportaron señal predictiva. Estos resultados indican que, si bien esas prácticas son socialmente valiosas, no explican diferencias materiales suficientes en el nivel de pobreza cuando se consideran otros factores más estructurales.

La selección automática confirma la centralidad de la educación, la accesibilidad geográfica, la calidad del empleo y las transferencias corrientes en la configuración de la pobreza guatemalteca. Al mismo tiempo, cuestiona la utilidad de incorporar indicadores superficiales de crédito o voluntariado en un índice operativo de pobreza, ofreciendo así una ruta clara para construir métricas más parcas y accionables.

CONCLUSIONES Y TRABAJOS FUTUROS

El presente estudio demostró que la aplicación del algoritmo **Boruta** sobre los microdatos de la **ENCOVI 2014** permite depurar de forma rigurosa los indicadores que verdaderamente explican la condición de pobreza en Guatemala. De las 31 variables inicialmente analizadas, solo 14 superaron consistentemente la importancia de sus contrapartes aleatorizadas, revelando la primacía del capital humano (alfabetización y escolaridad), los costos de acceso a servicios educativos y laborales (tiempos de traslado), la diversificación y remuneración del empleo, el valor imputado de la vivienda y los flujos de transferencias (internas y externas). Con este subconjunto reducido, el modelo XGBoost alcanzó un **AUC ROC de 0,818** y un **F1-macro de 0,614**, métricas estadísticamente indistinguibles de las obtenidas con las 31 variables, pero con un 55 % menos de atributos, lo que demuestra mayor parsimonia y viabilidad operativa.

Estos hallazgos se alinean estrechamente con los **objetivos planteados**: (i) depurar el microdato, (ii) implementar Boruta con un clasificador robusto, (iii) contrastar el rendimiento entre el conjunto completo y el reducido, y (iv) traducir la evidencia en insumos para política pública. Cada objetivo fue satisfecho: se depuraron 11 536 hogares sin pérdidas de información relevantes; Boruta se integró con un RandomForest balanceado y luego con XGBoost; la comparación empírica confirmó que la simplificación no sacrifica poder predictivo; y, finalmente, se identificaron determinantes accionables –educación, infraestructura y remesas– que orientan la focalización de recursos.

Desde una **perspectiva interpretativa**, los resultados refuerzan la hipótesis de que la pobreza guatemalteca se vincula más con déficits estructurales de capital humano y servicios públicos que con el mero acceso al crédito o con actividades no remuneradas. La relevancia de los tiempos de traslado a la escuela y al trabajo subraya la dimensión territorial del fenómeno: la distancia física se traduce en brecha social. Asimismo, la presencia de variables de ocio cognitivo (lectura, uso de medios) sugiere que los hogares con mayor exposición cultural poseen recursos intangibles que amortiguan la pobreza.

El análisis no está exento de **limitaciones**. Primero, los microdatos corresponden a 2014; choques económicos recientes (pandemia, inflación, fenómenos climáticos) podrían alterar la jerarquía de factores. Segundo, la variable objetivo proviene de la línea oficial de pobreza monetaria y no incorpora la dimensión subjetiva ni la pobreza multidimensional completa. Tercero, el modelo se evaluó en corte transversal; su capacidad de generalización temporal no ha sido validada. Cuarto, Boruta descarta variables de manera binaria, sin explorar combinaciones latentes que pudieran ser relevantes en presencia de interacciones no lineales más complejas.

Para **futuras líneas de investigación**, se propone: (i) replicar el pipeline con la **ENCOVI 2022** para examinar la estabilidad de los predictores y evaluar impactos post-COVID; (ii) incorporar datos espaciales y de infraestructura pública (carreteras, centros de salud) para capturar efectos geográficos explícitos; (iii) ensayar técnicas de selección alternativas —SHAP values, mRMR o RFE-GBM— que permitan contrastar y enriquecer la lista de variables recomendadas; (iv) ajustar métodos de remuestreo y calibración de umbrales para mejorar la sensibilidad hacia los grupos en pobreza extrema; y (v) avanzar hacia **modelos causales** (e.g., métodos de variables instrumentales o aprendizaje causal) que esclarezcan no solo la asociación, sino el impacto potencial de intervenciones concretas.

El proyecto valida la utilidad de la selección automatizada de características para clarificar la arquitectura de la pobreza en Guatemala y abre un camino factible para actualizar y robustecer los índices oficiales, contribuyendo así a decisiones de política más focalizadas y basadas en evidencia.

REFERENCIAS

Bibliografía

Solís-Salazar, M., & Madrigal-Sanabria, J. (2022). Una propuesta de aprendizaje automático para predecir la pobreza. *Tecnología en Marcha*, 1.

Hassan, A. A. (2024). Machine learning study using 2020 SDHS data to determine poverty determinants in Somalia. *Nature*, 1.

Sohnesen, T. (2016). Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment. *ResarchGate*, 1.

Sabogal, H. G.-B. (2021). *UN ANÁLISIS DE LA POBREZA EN COLOMBIA BASADO EN APRENDIZAJE AUTOMÁTICO.* Bogotá, Colombia: Universidad Jorge Tadeo Lozano.

Saleem, J. Z. (2024). Application of the Boruta algorithm to assess the multidimensional determinants of malnutrition among children under five years living in southern Punjab, Pakistan. *BMC Public Health*, 1.

Narciso Cruz, R. D. (2014). *República de Guatemala: Encuesta Nacional de Condiciones de Vida 2014.* Guatemala: INE.

Anexos

Enlace de repositorio de GitHub

https://github.com/1u1s4/IEBS_proyecto_final

Enlace de video de presentación del proyecto

https://drive.google.com/drive/folders/1Gay7L-OP7e5anTqfdgzLDGtNNmR7SK2l?usp=sharing

Variables descartadas y aceptadas

Aceptadas:

- P13A02A: ¿Cuántos comercios, negocios o fábricas tienen en este hogar?
- P06B01: ¿Sabe leer y escribir?
- P06B10A: ¿Cuánto tiempo tarda para ir al centro educativo donde estudia? (Horas)
- P06B10B: ¿Cuánto tiempo tarda para ir al centro educativo donde estudia? (Minutos)
- P09F03B: Horas que dedicó a trasladarse a su lugar de trabajo ayer
- P09F08B: Horas que dedicó a lectura recreativa de algún libro, revista, periódico ayer
- P09F08C: Minutos que dedicó a lectura recreativa de algún libro, revista, periódico ayer
- P09F09B: Horas que dedicó a ver televisión, escuchar música, utilizar internet (entretenimiento) ayer
- P09F09C: Minutos que dedicó a ver televisión, escuchar música, utilizar internet (entretenimiento) ayer
- P10B01: ¿Cuántos trabajos tuvo la semana pasada?
- P10B08: ¿Cuál fue el sueldo o salario mensual sin descuentos en este trabajo?
- P10B20B: ¿Cuánto le costaría la vivienda por mes si tuviera que alquilarla?
- P11A05B: ¿Cuánto recibió de ayudas o donaciones de personas ubicadas en Guatemala?
- P11A06B: ¿Cuánto recibió en remesas de personas que viven en el extranjero?

No Aceptadas:

- P01F03: Si tuviera que comprar lo que recibió en los últimos 12 meses en donaciones, ¿cuánto cree que le costaría?
- P15B02: En los últimos 15 días, ¿cuál fue el monto total de sus compras al crédito en alimentos?
- P15B04: ¿Cuál fue el monto de sus compras al crédito de los bienes y artículos del hogar?

- P15B06A: Durante los últimos 12 meses, ¿cuántas compras al crédito realizaron? (número de compras)
- P09A03B: Horas que trabajó ayer sin percibir ingresos
- P09A03C: Minutos que trabajó ayer sin percibir ingresos
- P09B02B: ¿Cuántas horas pasó en transporte a su lugar de estudios? (ayer)
- P09B02C: ¿Cuántos minutos pasó en transporte a su lugar de estudios? (ayer)
- P09F03C: Minutos que dedicó a trasladarse a su lugar de trabajo ayer
- P09F04B: Horas que dedicó a descansar, relajarse, etc. ayer
- P09F04C: Minutos que dedicó a descansar, relajarse, etc. ayer
- P09F05B: Horas que dedicó a actividades deportivas, culturales, etc. fuera del hogar ayer
- P09F05C: Minutos que dedicó a actividades deportivas, culturales, etc. fuera del hogar ayer
- P09F06B: Horas que dedicó a algún trabajo para otros hogares de forma gratuita ayer
- P09F06C: Minutos que dedicó a algún trabajo para otros hogares de forma gratuita ayer
- P09F07B: Horas que dedicó a realizar gestiones para mejoras de la comunidad, etc. ayer
- P09F07C: Minutos que dedicó a realizar gestiones para mejoras de la comunidad, etc. ayer

Salida de F1-Score y AUC-ROC

```
[22]  ✓ 0.0s

...  F1 macro: 0.6142
     Macro AUC-ROC (OvR): 0.8175
                  precision    recall  f1-score   support

               0       0.75      0.81      0.78      1388
               1       0.60      0.46      0.52       485
               2       0.54      0.54      0.54      1011

        accuracy                           0.66      2884
       macro avg       0.63      0.60      0.61      2884
    weighted avg       0.65      0.66      0.65      2884
```