PLANTILLA

GLOBAL PROJECT

**ALUMNO: Luis Alfredo Alvarado Rodríguez**

**PROGRAMA:**

**POSGRADO EN BIG DATA**

**NOMBRE DEL PROYECTO:**

Contenido

[**RESUMEN** 3](#_Toc196133619)

[**INTRODUCCIÓN** 3](#_Toc196133620)

[**ESTADO DEL ARTE** 5](#_Toc196133621)

[Métodos tradicionales en el análisis de la pobreza 5](#_Toc196133622)

[Algoritmos de aprendizaje automático aplicados a la pobreza 6](#_Toc196133623)

[Selección de variables y técnicas de importancia de atributos 7](#_Toc196133624)

[Conclusiones del estado del arte 8](#_Toc196133625)

[**OBJETIVOS** 8](#_Toc196133626)

[Objetivo general 8](#_Toc196133627)

[Objetivos específicos 8](#_Toc196133628)

[**SOLUCIÓN PLANTEADA** 9](#_Toc196133629)

[**EVALUACIÓN** 9](#_Toc196133630)

[**RESULTADOS** 10](#_Toc196133631)

[**CONCLUSIONES Y TRABAJOS FUTUROS** 10](#_Toc196133632)

[**REFERENCIAS** 11](#_Toc196133633)

[Bibliografía 11](#_Toc196133634)

# **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 RandomForestClassifier balanceado, 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**

En esta sección el estudiante debe describir la solución planteada, comenzando por la metodología (pasos que siguió) de desarrollo. Posteriormente la descripción del desarrollo de cada etapa seguida.

La metodología que debe utilizar el estudiante de master debe estar validada por la comunidad científica y el estudiante debe justificarlo.

1. METODOLOGÍA
   * Etapa 1
   * Etapa 2
   * Etapa 3
   * Etapa 4
   * Etapa 5
2. DESARROLLO DE CADA ETAPA

Extensión máxima: 15 páginas.

Se recomienda utilizar anexos tras el final de la memoria, únicamente si se estima estrictamente necesario incluir una gran cantidad de material gráfico o tabular.

*Escribe aquí…*

# **EVALUACIÓN**

El estudiante en esta sección debe describir cómo realizó la evaluación de la solución planteada.

La evaluación de un Global Project debe basarse en una metodología validada y bien establecida que el estudiante deberá justificar.

Extensión máxima: 6 páginas.

Se recomienda utilizar anexos tras el final de la memoria, únicamente si se estima estrictamente necesario incluir una gran cantidad de material gráfico o tabular.

*Escribe aquí…*

# **RESULTADOS**

El estudiante en esta sección debe describir los resultados obtenidos en el proceso de evaluación de la solución planteada.

Debe usar notación estándar para presentación de resultados de carácter científico.

Extensión máxima: 6 páginas.

Se recomienda utilizar anexos tras el final de la memoria, únicamente si se estima estrictamente necesario incluir una gran cantidad de material gráfico o tabular.

*Escribe aquí…*

# **CONCLUSIONES Y TRABAJOS FUTUROS**

El estudiante plantea las conclusiones de su trabajo, y cómo considera que puede seguir avanzando en el mejoramiento de la solución planteada.

El apartado de conclusiones es una parte fundamental del GP, ya que permite presentar los resultados y sacar conclusiones sobre lo aprendido. En este apartado, es importante alinear los objetivos del trabajo con los resultados obtenidos. Esto significa que se debe analizar cómo los resultados se relacionan con los objetivos establecidos al principio del GP.

Para que tus lectores comprendan bien las conclusiones, se recomiendan seguir los siguientes pasos:

1. Resumen de los resultados: Comienza tu apartado de conclusiones con un breve resumen de los resultados más importantes que has obtenido. Esto permitirá a tus lectores tener una idea general de lo que has logrado en tu TFM.
2. Relación con los objetivos: A continuación, es importante que analices cómo tus resultados se relacionan con los objetivos que estableciste al inicio de tu TFM. Es importante que destaques cómo has logrado alcanzar cada uno de los objetivos planteados y cómo los resultados obtenidos los respaldan.
3. Interpretación de los resultados: En este punto, debes interpretar tus resultados y explicar su relevancia. Debes analizar los resultados desde diferentes perspectivas y explicar qué implicaciones tienen en tu campo de estudio.
4. Limitaciones y futuras líneas de investigación: Por último, es importante que hables sobre las limitaciones de tu trabajo y las posibles líneas de investigación futura que se podrían desarrollar a partir de tu TFM.

Nota: las conclusiones deben ser claras y concisas, y estar bien estructuradas. De esta forma, los lectores podrán entender fácilmente los resultados y la relevancia del trabajo.

Extensión máxima: 2 páginas

*Escribe aquí…*

# **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.