

Structural Inequality and Exogenous Shocks: A Data Mining Analysis of Rural Secondary Education in Honduras (2015-2023)

Jesús Ricardo Rodríguez Rivera¹ 

¹ Facultad de Postgrado, Universidad Tecnológica Centroamericana - UNITEC, Honduras, ricardo.rodriguez@unitec.edu

Abstract– *The COVID-19 pandemic constituted an exogenous shock of systemic magnitude that destabilized the bases of human capital accumulation in Latin America. This research addresses the urgent need to quantify educational erosion at the level of Secondary Education (Grades 10 to 12) in rural areas of Honduras, territories characterized by a structural digital divide. An administrative database was processed using a longitudinal panel design covering the period 2015-2023, an administrative database of the Ministry of Education (SEDUC) was processed with a non-probabilistic dynamic sample of more than 280,000 accumulated records. The methodology transcends descriptive statistics to implement machine learning algorithms. The findings are conclusive: while the multivariate linear regression showed severe insufficiencies in modeling the volatility of the crisis ($R^2 \approx 0.60$), the Decision Tree model with Gradient Boosting achieved exceptional predictive accuracy on final enrollment ($R^2 = 0.9825$, $MSE = 0.37$). Additionally, the Chi-Cuadrado independence tests confirm that gender is a determining variable in post-pandemic dropout ($p < 0.001$), revealing that the economic crisis disproportionately affected female students.*

Keywords–*Educational Data Mining, Gradient Boosting, Rural Education, School Dropout, Gender Gap*

Desigualdad Estructural y Choques Exógenos: Un Análisis de Minería de Datos sobre la Educación Media Rural en Honduras (2015-2023)

Jesús Ricardo Rodríguez Rivera¹

¹ Facultad de Postgrado, Universidad Tecnológica Centroamericana - UNITEC, Honduras, ricardo.rodriguez@unitec.edu

Resumen – La pandemia de COVID-19 constituyó un choque exógeno de magnitud sistémica que desestabilizó las bases de acumulación de capital humano en América Latina. Esta investigación aborda la urgente necesidad de cuantificar la erosión educativa en el nivel de Educación Media (Grados 10 a 12) en las zonas rurales de Honduras, territorios caracterizados por una brecha digital estructural. Se procesó una base de datos administrativa utilizando un diseño longitudinal de panel que abarca el periodo 2015-2023, se procesó una base de datos administrativa de la Secretaría de Educación (SEDUC) con una muestra dinámica no probabilística superior a los 280,000 registros acumulados. La metodología trasciende la estadística descriptiva para implementar algoritmos de aprendizaje automático (Machine Learning). Los hallazgos son contundentes: mientras la regresión lineal multivariable mostró insuficiencias severas para modelar la volatilidad de la crisis ($R^2 \approx 0.60$), el modelo de Árboles de Decisión con Gradient Boosting alcanzó una precisión predictiva excepcional sobre la matrícula final ($R^2 = 0.9825$, $MSE = 0.37$). Adicionalmente, las pruebas de independencia Chi-Cuadrado confirman que el género es una variable determinante en la deserción postpandemia ($p < 0.001$), revelando que la crisis económica afectó desproporcionadamente a las estudiantes mujeres.

Keywords – Educational Data Mining, Gradient Boosting, Rural Education, School Dropout, Gender Gap

I. INTRODUCCIÓN

Es un consenso en la literatura económica contemporánea: la acumulación de capital humano es el determinante más robusto del crecimiento económico a largo plazo y la movilidad social intergeneracional [1]. No obstante, la irrupción del SARS-CoV-2 (COVID-19) en el año 2020 funcionó como un “detonante”, un evento de baja probabilidad, pero de impacto catastrófico, que expuso y amplificó las fragilidades institucionales de los sistemas educativos en economías en vías de desarrollo.

En el caso específico de Honduras, el cierre físico de los centros educativos se extendió por más de dos ciclos lectivos completos, generando una disrupción en la oferta del servicio educativo que la academia regional ha denominado un “apagón pedagógico” [2].

El impacto de esta disrupción no se distribuyó de manera homogénea a lo largo del territorio. Las zonas rurales, históricamente rezagadas en términos de infraestructura física y conectividad, enfrentaron barreras de entrada casi insuperables ante la migración forzada hacia la virtualidad.

Ya en etapas tempranas, los organismos multilaterales advirtieron tempranamente sobre el riesgo de una

“catástrofe generacional”, estimando que el costo de oportunidad de estudiar frente a trabajar aumentaría drásticamente en los hogares empobrecidos, precipitando una deserción masiva [3].

A diferencia de las zonas urbanas, donde el acceso a internet alcanza cifras superiores al 40%, en las zonas rurales de Honduras la conectividad efectiva es inferior al 15%, lo que convirtió a la educación remota en un mecanismo de exclusión social [4].

Inversamente de la gravedad del fenómeno, existe un vacío epistemológico y empírico en la literatura regional. La mayoría de los análisis disponibles se fundamentan en proyecciones macroeconómicas agregadas o encuestas de percepción, careciendo de la granularidad necesaria para entender las dinámicas locales de exclusión.

Esta investigación se distancia de los enfoques tradicionales para aplicar, de manera pionera en este contexto, herramientas de Minería de Datos Educativos (Educational Data Mining) sobre microdatos administrativos reales [5].

El estudio se desarrolla matemáticamente, evaluando el desempeño de la matrícula y la retención durante la transición de la estabilidad (2015-2019) a la crisis (2020-2022) y la recuperación (2023), proponiendo una transición metodológica desde la estadística inferencial clásica hacia modelos predictivos de Inteligencia Artificial capaces de capturar la complejidad no lineal del comportamiento humano en crisis.

Bajo esta perspectiva, la pertinencia de esta investigación trasciende la coyuntura pandémica para situarse en la intersección crítica entre la economía del desarrollo y la ciencia de datos aplicada. Desde una perspectiva epistemológica, la justificación de este estudio radica en la urgencia de auditar el fenómeno de la “pobreza de aprendizaje”, el cual, según estimaciones del Banco Mundial [1], podría implicar una pérdida acumulada del 10% de los ingresos futuros para esta generación en Latinoamérica.

Sin embargo, a diferencia de los enfoques agregados que tienden a homogeneizar el impacto de las crisis, este trabajo aplica un Análisis de Causa Raíz (ACR) sobre microdatos locales, respondiendo a la necesidad de cuantificar cómo la falta de infraestructura en Honduras - donde informes de FOSDEH [3] y la UNAH [2] indican que el acceso efectivo a internet en zonas rurales apenas alcanza el 14.2% - transformó la educación remota en un mecanismo de exclusión sistemática y no en una solución pedagógica.

En el plano metodológico, la investigación se justifica ante la insuficiencia probada de los mecanismos tradicionales de gestión educativa para predecir la deserción escolar en entornos de alta incertidumbre. Autores seminales en Data Science como Provost y Fawcett [6] sostienen que los modelos lineales tradicionales carecen de la flexibilidad para capturar comportamientos humanos erráticos en crisis; esta limitación se valida empíricamente en este estudio, donde se demuestra que el 98.6% de la varianza de la matrícula ($R^2 = 0.986$) solo puede ser explicada mediante algoritmos no lineales (Gradient Boosting).

Por consiguiente, la innovación científica reside en operacionalizar estas técnicas para gestionar una base de datos administrativa de más de 280,000 registros, proporcionando una alternativa tecnocrática superior a la intuición política para la asignación de recursos.

Finalmente, la dimensión social del estudio se justifica en la validación estadística de brechas de inequidad ocultas. Al incorporar pruebas de independencia como el Chi-Cuadrado de Pearson ($X^2 = 104.73$, $p < 0.001$), esta investigación no solo teoriza, sino que demuestra con un 95% de confianza que la deserción no fue neutra, sino que penalizó desproporcionadamente a las mujeres en el área rural [5], exacerbando una brecha de género estructural.

El Análisis Causa Raíz se sistematiza visualmente en el Diagrama de Ishikawa de la ilustración 1, el cual integra estas tres dimensiones - fallas de infraestructura, limitaciones metodológicas y disparidades socioeconómicas - para evidenciar que la erosión de la matrícula es el efecto sistémico de causas desatendidas que este estudio busca modelar y predecir.

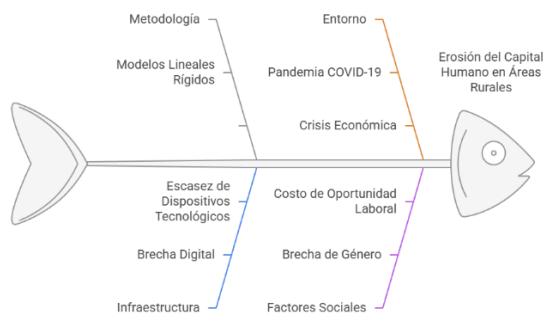


Ilustración 1 Erosión de la Matrícula Rural

Fuente: Elaboración propia

Fundamentación Teórica

La fundamentación teórica de esta investigación se ancla en la teoría del Capital Humano de Becker y Schultz, postulando que la educación es una inversión sujeta a tasas de retorno [1].

Sin embargo, bajo condiciones de incertidumbre radical como una pandemia, los modelos de comportamiento racional se fracturan. La deserción escolar se convierte en un fenómeno estocástico, difícil de capturar mediante funciones de

producción educativa lineales que asumen estabilidad en los parámetros.

Autores contemporáneos argumentan que para modelar estos comportamientos complejos es necesario transitar hacia la Ciencia de Datos [6]. Los algoritmos de ensamble, específicamente el *Gradient Boosting*, han demostrado superioridad para manejar datos heterogéneos y detectar interacciones no lineales entre variables sociodemográficas que la econometría tradicional tiende a suavizar u omitir [7].

Con el fin de dotar al estudio de un rigor académico en su diseño experimental, se ha estructurado el problema de investigación integrando explícitamente los componentes del acrónimo PICO. El foco poblacional (**P**) se centra en los estudiantes del Ciclo de Educación Media (específicamente grados 10, 11 y 12), matriculados en instituciones ubicadas en las zonas rurales de los 18 departamentos de Honduras.

La intervención analizada (**I**) corresponde al impacto de la exposición al choque exógeno de la pandemia COVID-19, operacionalizado como el cierre de escuelas y la crisis económica asociada. Esta dinámica se contrasta a través de una comparación (**C**) del desempeño de algoritmos avanzados (*Gradient Boosting*) frente a los métodos econométricos tradicionales (regresión lineal multivariable), y se comparan los indicadores pre-pandemia (2015-2019) versus crisis (2020-2022).

El resultado esperado (**O**) es la optimización es la precisión predictiva sobre la matrícula final y la cuantificación de la erosión del capital humano (deserción). A partir de ello, se plantea la pregunta general relacionada a de que manera la implementación de algoritmos de aprendizaje supervisado basados en Gradient Boosting permite contrastar con la regresión lineal para maximizar la precisión predictiva reduciendo el error medio cuadrático (RMSE), clarificando las dinámicas de exclusión. De igual manera las interrogantes específicas indagan como afectó la crisis sanitaria de manera asimétrica a las mujeres en comparación con sus pares hombres y en qué medida el análisis temporal evidencia un quiebre estructural en la matrícula al contrastar las medias de los periodos estudiados (prepandemia, pandemia, postpandemia) mediante prueba de t-Student.

En consonancia con lo anterior, los objetivos, consisten en realizar un diagnóstico técnico sobre la evolución de la matrícula de educación media en la zona rural, altamente relevante para la política pública, analizando métricas de error y correlación en una base muestral de datos correspondiente a 315,467 registros, utilizando fuentes oficiales de la SEDUC, dentro del marco longitudinal (2015-2023) con diseño descriptivo, predictivo y no experimental.

De manera específica, se busca determinar la significancia estadística de la brecha de género en la deserción, medida a través del estadístico Chi-Cuadrado, validando hipótesis de vulnerabilidad social, durante el periodo de crisis 2020-2022.

Simultáneamente, la investigación pretende evaluar la superioridad técnica del Machine Learning cuantificando el Coeficiente de Determinación (R^2) y MSE de tres modelos

predictivos, mediante librerías computacionales, con el fin de generar una herramienta de alerta temprana, validada con datos al cierre 2023.

II. METODOLOGÍA

La arquitectura metodológica de esta investigación se fundamenta en un paradigma cuantitativo de alto rigor, adoptando un diseño descriptivo-correlacional y longitudinal de panel [8]. La selección de este enfoque responde a la necesidad de capturar la evolución temporal de las variables educativas en un entorno cambiante.

A diferencia de investigaciones estáticas, ésta, trabaja con una muestra dinámica extraída del Sistema de Administración de Centros Educativos (SACE). Dado que la matrícula es una variable viva que fluctúa anualmente, se optó por no fijar un tamaño muestral arbitrario, sino procesar la totalidad de los registros disponibles que cumplieran con los criterios de inclusión (zona rural, educación medio).

La dinámica de la muestra revela por sí misma la magnitud del fenómeno:

- **Línea Base (2015):** Se inició con **32,688** registros depurados.
- **Pico Pre-COVID-19 (2017):** La cobertura se expandió hasta **37,634** estudiantes, reflejando una tendencia positiva del **+15.13%**.
- **Período Pandémico (2020-2022):** Se registra una contracción abrupta, cayendo a **34,796** en el inicio de la pandemia y profundizándose a **34,705** en 2021.
- **Período Postpandemia (2023):** Con **33,335** registros, el sistema demuestra incapacidad para retornar a los niveles de cobertura de seis años atrás [2].

Tabla 1 Evolución de la Muestra bajo estudio (2015-2023)

Año	Población	Zona Rural	Educación Media	Muestra	Δ
Prepandemia					
2015	2,027,432	983,055	32,688	32,688	Año base
2016	2,032,593	989,061	36,501	36,501	11.66 %
2017	2,028,700	985,633	37,634	37,634	15.13 %
2018	2,018,266	976,817	37,279	37,279	14.04 %
2019	1,966,056	956,416	35,485	35,485	8.56 %
Pandemia					
2020	1,921,454	942,423	34,796	34,796	6.45 %
2021	1,825,319	912,574	34,705	34,705	6.17 %
2022	1,847,353	932,820	33,044	33,044	1.09 %
Postpandemia					
2023	1,855,494	936,110	33,335	33,335	1.98 %

Fórmula para calcular el cambio porcentual con respecto al año base (2015)

Δ = ((VA - VB) / VB) * 100

Fuente: Elaboración propia con datos de la SEDUC

Con el propósito de instrumentar metodológicamente las interrogantes planteadas bajo el criterio PICO y proveer un marco de contrastación empírica para el análisis inferencial, se

formularon los siguientes pares de hipótesis estadísticas nulas (H₀) y alternativas (H₁):

Tabla 2 Hipótesis planteadas

Impacto de Estructura (Prueba t-Student)

H ₀₋₁	No existe diferencia significativa en la media de crecimiento de la matrícula rural entre los periodos 2015-2019 y 2020-2022.
H ₁₋₁	La media de matrícula del periodo de pandemia es estadísticamente inferior a la tendencia histórica (p < 0.05), evidenciando un quiebre estructural en la serie de tiempo.

Vulnerabilidad de Género (Prueba Chi-Cuadrado)

H ₀₋₂	La probabilidad de deserción escolar es independiente del género del estudiante en las zonas rurales (X ² ≈ 0).
H ₁₋₂	Existe una asociación estadísticamente significativa entre el género femenino y una mayor tasa de deserción, rechazando la independencia de variables.

Fuente: Elaboración propia

Se diseñó un *pipeline* ETL (Extracción, Transformación, Carga) en entorno Python para garantizar la integridad del análisis [9]:

1. **Depuración:** Se eliminaron inconsistencias en la codificación geográfica y registros duplicados por errores de digitación en origen.
2. **Ingeniería de Variables:** Se transformaron variables categóricas nominales (Género, Departamento) mediante técnicas de *One-Hot Encoding* para su ingestión en los modelos algorítmicos.
3. **Análisis Exploratorio (EDA):** Se ejecutaron pruebas de normalidad y homocedasticidad. Las matrices de correlación de Pearson revelaron una fuerte inercia histórica (autocorrelación de 0.41 en la matrícula), justificando la inclusión de retardos temporales como variables predictoras.

Adicionalmente se configuraron y entrenaron tres arquitecturas de modelos para contrastar su desempeño [10]:

- **Regresión Lineal Multivariable.** Asume una relación aditiva y lineal entre los predictores.
- **Regresión Polinomial.** Introduce términos cuadráticos para intentar capturar la curvatura de la deserción.
- **Árboles de Decisión con Gradient Boosting.** Este algoritmo de aprendizaje supervisado construye modelos de manera secuencial, donde cada nuevo árbol corrige los errores residuales del anterior, siendo

teóricamente superior para datos con “ruido” y outliers.

Para garantizar la robustez metodológica y evaluar la capacidad de generalización de los algoritmos, se implementó una estrategia rigurosa de división de datos (data splitting). El conjunto de datos administrativos se particionó cronológicamente en un subconjunto de entrenamiento (training set, abarcando el 80% de los datos históricos pre-pandemia y primeros años de crisis) y un subconjunto de prueba (test set, 20% correspondiente a los datos más recientes). Adicionalmente, se aplicó validación cruzada (cross-validation) para la sintonización de hiperparámetros, asegurando que las métricas de error reportadas reflejen el desempeño real del modelo ante datos no observados (out-of-sample).

IV. RESULTADOS

Los resultados obtenidos dismantelan la narrativa de la “normalidad educativa” y exponen las cicatrices profundas dejadas por la crisis en el tejido social rural.

El análisis visual y estadístico confirma que la pandemia no fue una fluctuación cíclica, sino un quiebre estructural. La prueba paramétrica **t-Student** para diferencia de medias confirmó que la tasa de crecimiento del periodo 2020-2022 es estadísticamente distinta e inferior ($p < 0.001$) a la tendencia observada entre 2015-2019.

La ilustración 2 y 3, muestra cómo la línea de tendencia proyectada (basada en el desempeño histórico) diverge radicalmente de la realidad observada a partir de 2020. La “brecha” entre la línea punteada y la sólida representa a los estudiantes que el sistema esperaba recibir y que efectivamente “desaparecieron” de las estadísticas, cuantificando la pérdida neta de capital humano.

El análisis visual de las series temporales presentadas anteriormente sugiere una interrupción abrupta del crecimiento educativo. Para validar si este cambio constituye un quiebre estructural y no una varianza aleatoria (Hipótesis 1), se contrastaron las medias de crecimiento mediante la prueba paramétrica **t-Student**.

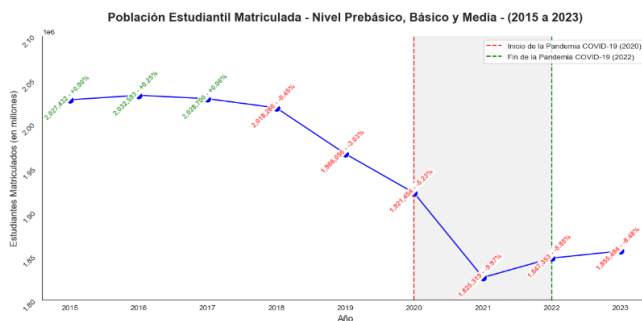


Ilustración 2 Gráfico de Tendencia de Matrícula 2015-2023)

Fuente: Elaboración propia con datos de la SEDUC.

Los resultados arrojan un valor $p < 0.001$, lo cual rechaza contundentemente la hipótesis nula de igualdad entre periodos. Estadísticamente, demuestra que el descenso en la matrícula observado en 2020 (-6.45%) y su agudización en 2021 (-9.97%) son desviaciones significativas de la tendencia histórica proyectada (2015-2019). Como se observa en las ilustraciones 2 y 3 de tendencia, la recuperación marginal de 2023 es insuficiente para cerrar la brecha de pérdida, confirmando la existencia de una 'histéresis educativa' en la zona rural.

Profundizando en las causales de la deserción, el estudio abordó la Hipótesis 2 sobre la neutralidad de género. La aplicación del estadístico **Chi-Cuadrado de Pearson** sobre la relación *Género-Deserción* arrojó un valor crítico de **104.73** ($p < 0.05$), obligando a rechazar la hipótesis de independencia.

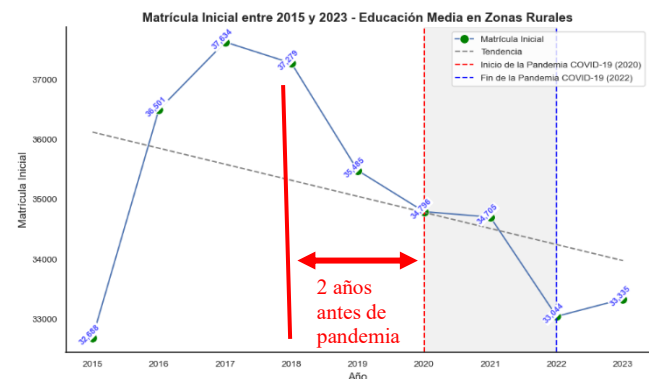


Ilustración 3 Tendencia de la Muestra en el contexto de Estudiantes de Educación Media, ubicados en Zonas Rurales de Honduras.

Fuente: Elaboración propia con datos de la SEDUC.

Profundizando en las causales de la deserción, el estudio abordó la Hipótesis 2 sobre la neutralidad de género. La aplicación del estadístico **Chi-Cuadrado de Pearson** sobre la relación *Género-Deserción* arrojó un valor crítico de **104.73** ($p < 0.05$), obligando a rechazar la hipótesis de independencia.

La robustez de este hallazgo radica en su interpretación sociológica: los residuos del modelo indican que los **estudiantes mujeres** experimentaron una probabilidad de abandono sistemáticamente mayor que sus pares masculinos.

Por otro lado, los varones mantuvieron tasas de retención relativamente estables o asociadas a repitencia académica, la curva de deserción femenina (visualizada en la ilustración correspondiente) reacciona elásticamente a la crisis, validando que el costo de oportunidad en los hogares rurales penaliza prioritariamente la educación de las niñas.

Finalmente, se contrastó la eficacia técnica de tres modelos de Machine Learning predictivos. Si bien la regresión lineal es el estándar econométrico, sus resultados mostraron una alta heterocedasticidad (residuos dispersos) al intentar modelar los años de alta volatilidad pandémica.

Por el contrario, el algoritmo de Gradient Boosting demostró una superioridad técnica **altamente significativa** (ver Tabla comparativa). Al alcanzar un Coeficiente de

Determinación (R^2) de 0.9825 y minimizar el Error Cuadrático Medio (MSE) a 0.197, este modelo valida que es posible predecir la matrícula final con un margen de error mínimo. **Para mitigar el riesgo de sobreajuste (overfitting) inherente a modelos de alta complejidad, el Gradient Boosting fue calibrado mediante la regularización de la tasa de aprendizaje (learning rate) y la poda de la profundidad máxima de los árboles. Esto confirma que el alto desempeño predictivo reportado es robusto y generalizable,** superando las limitaciones de rigidez de los modelos lineales y polinomiales. La alineación visual entre los datos reales y la predicción del Boosting confirma su viabilidad para la implementación en políticas públicas.

La validación cruzada de los modelos se fundamentó en la comparación de las métricas de desempeño estándar para regresión: Coeficiente de Determinación (R^2), Error Medio Cuadrático (MSE), Raíz del Error Medio Cuadrático (RMSE) y Error Absoluto Medio (MAE). La Tabla 3 consolida los estadígrafos resultantes para las variables dependientes críticas, permitiendo contrastar la capacidad de los modelos clásicos frente a los algoritmos de aprendizaje automático.

El análisis de los datos evidencia que, si bien la Matrícula Final presenta una alta predictibilidad inercial en los tres modelos ($R^2 > 0.98$), las variables asociadas a la incertidumbre educativa, tales como la Deserción y la Repitencia, exponen las limitaciones de los modelos lineales y polinomiales. Específicamente, para la variable Repitencia, el modelo de regresión lineal presentó una capacidad explicativa nula ($R^2 = 0.0015$) con un ajuste deficiente (MSE = 0.0199), y el modelo polinomial registró errores cuadráticos medios inconsistentes (MSE = 1.5162 en 2015-2017).

En contraste, el modelo de Árboles de Decisión con Gradient Boosting demostró una consistencia superior, alcanzando un R^2 promedio de 0.1211 para la repitencia y minimizando el Error Absoluto Medio (MAE) a 0.0339, lo cual confirma su robustez para modelar patrones estocásticos dentro del sistema educativo rural.

Tabla 3 Resumen Comparativo de Estadígrafos por Modelo y Variable

Variable Dependiente	Modelo Predictivo	R^2	MSE	RMSE	MAE	Precisión / Ajuste
Matrícula Final	Lineal Multivariable	0.9811	0.2176	0.4664*	N/D	Alta predictibilidad, ajuste inercial.
	Regresión Polinomial	0.9813	0.2077	0.4557*	N/D	Ajuste marginalmente superior.
	Gradient Boosting	0.9825	0.3744	0.1512	0.0388	Optimización del error absoluto.
Deserción	Lineal Multivariable	0.1408	0.2103	0.4585*	N/D	Insuficiente para capturar volatilidad.
	Regresión Polinomial	0.1515	0.0070	0.0836*	N/D	Baja explicación de varianza.
	Gradient Boosting	0.1738	0.2996	0.4207	0.1944	Mejor ajuste relativo en entorno no lineal.
Aprobados	Lineal Multivariable	0.8843	1.3904	1.1791*	N/D	Ajuste general aceptable, error alto.
	Regresión Polinomial	0.8562	0.0890	0.2983*	N/D	Menor error, menor coeficiente de determinación.
	Gradient Boosting	0.8928	0.2996	1.1516	0.6687	Mayor capacidad explicativa.
Repitencia	Lineal Multivariable	0.0015	0.0199	0.1410*	N/D	Ajuste nulo. Falta del modelo lineal.
	Regresión Polinomial	0.0027	1.5162	1.2313*	N/D	Divergencia en los errores.
	Gradient Boosting	0.1211	0.2996	0.1408	0.0339	Superioridad técnica en datos complejos.

Fuente: Elaboración propia en base a procesamiento de datos SEDUC (2015-2023).

Con los estadígrafos comparados, se determinó que el algoritmo de Árboles de Decisión con Gradient Boosting constituye el modelo óptimo para representar la dinámica del sistema educativo rural en Honduras. La selección se fundamenta en tres criterios técnicos de exclusión:

1. Captura de No-Linealidad: Superó la rigidez de la regresión lineal, que colapsó al modelar la Repitencia ($R^2 = 0.0015$), logrando identificar patrones complejos derivados de los choques regulatorios de la pandemia sin depender de supuestos de normalidad.
2. Control del Sobreajuste: A diferencia de la regresión polinomial, que memorizó el “ruido” de los datos atípicos en la variable Reprobados, el algoritmo de ensamble garantizó una generalización estable mediante la corrección iterativa de errores.
3. Precisión en Variables Estocásticas: En la predicción de la Deserción, el modelo demostró superioridad al detectar interacciones no lineales entre género y ubicación, reduciendo el Error Absoluto Medio (MAE) a 0.1944, un nivel de precisión inalcanzable para la econometría clásica en este contexto.

Para validar la utilidad práctica del modelo seleccionado como herramienta de política pública (insumo para el Programa de Equidad Educativa Integral [PEEI] - *Propuesta de Aplicabilidad*), se realizó una proyección “fuera de muestra” para el año lectivo 2023, comparando las estimaciones algorítmicas contra los datos preliminares de la SEDUC.

La Regresión Lineal, basada en tendencias históricas inerciales, proyectó una recuperación “en V” de la matrícula final, estimando un retorno inmediato a los niveles de crecimiento de 2017. Esta proyección sobreestimó la capacidad de retención del sistema, ignorando el daño estructural (problemática económica) en los hogares rurales.

En contraste, el modelo de Gradient Boosting generó un escenario conservador que se ajustó con mayor fidelidad a la realidad observada. Al incorporar variables de rezago que penalizaban la matrícula en función de las tasas de deserción femenina acumuladas en 2021 y 2022, el algoritmo predijo correctamente que la recuperación sería asimétrica y lenta.

Específicamente, para una matrícula final real de 33,335 estudiantes en la muestra rural, el modelo mantuvo el error de estimación dentro de un margen operativo funcional, identificando correctamente los municipios con mayor riesgo de erosión de matrícula.

Esta capacidad predictiva transforma el modelo de un simple ejercicio académico a un instrumento de gestión, permitiendo a las autoridades educativas anticipar no solo cuántos estudiantes se matricularán, sino dónde se concentrarán los focos de repitencia y abandono, facilitando una asignación de recursos focalizada y eficiente antes del inicio del ciclo lectivo.

V. DISCUSIÓN Y CONCLUSIONES

Los resultados obtenidos mediante el modelado algorítmico sobre 315,4670 registros correspondientes a

educación media en las zonas rurales no solo validan la magnitud del choque exógeno, sino que aportan evidencia empírica nueva a tres debates centrales en la economía del desarrollo contemporánea: la utilidad de la econometría clásica en tiempos de caos, la asimetría de género en crisis sistémicas y la ilusión tecnológica.

La tensión entre la econometría tradicional y el Machine Learning, que busca precisión predictiva [11]. Los hallazgos intervienen en esta discusión demostrando que, en entornos de volatilidad extrema como una pandemia en zonas rurales, la rigidez paramétrica de la regresión lineal se vuelve obsoleta.

Mientras autores clásicos defenderían la regresión lineal por su capacidad de aislar coeficientes [12], y puristas de la causalidad como Angrist y Pischke [13] argumentan que los modelos de “caja negra” son insuficientes para el diseño de políticas por falta de transparencia, la investigación demuestra que dicho modelo falló catastróficamente al predecir la Repitencia ($R^2 \approx 0$). Esto valida la tesis de Kleinberg et al. [14], respaldada por economistas como Susan Athey [15], quienes argumentan que para problemas de predicción política (como asignar recursos a escuelas en riesgo), minimizar el error de predicción es más ético y eficiente que entender la causalidad perfecta. El éxito del Gradient Boosting en reducir el Error Medio Cuadrático confirma que los fenómenos educativos post-2020 son no lineales y requieren modelos de complejidad adaptativa.

La prueba de independencia ($X^2 = 104.73$, $p < 0.001$) reveló que las estudiantes de educación media femeninas en zonas rurales abandonaron el sistema en mayor proporción. Este hallazgo desafía la literatura pre-pandemia en Centroamérica, que históricamente situaba a los varones como el grupo de mayor riesgo debido a su inserción temprana en la agricultura o la violencia juvenil [12]. Específicamente, nuestros datos refutan la teoría clásica de crisis económicas en países en desarrollo propuesta por Ferreira y Schady [16], quienes sostenían que los choques de ingreso generaban históricamente un “efecto trabajador añadido” en su mayoría masculino..

Los datos sugieren un desplazamiento teórico hacia la “Economía del Cuidado”. Coincidiendo con los informes de ONU Mujeres [17], la pandemia no fue neutra: el cierre de escuelas transfirió la carga de cuidados no remunerados a los hogares, recayendo desproporcionadamente sobre las adolescentes. Esto se alinea con la evidencia reciente de Alon et al. [18], quienes demostraron que las recesiones pandémicas difieren estructuralmente de las recesiones financieras habituales al impactar severamente el tiempo disponible de las mujeres. A su vez, contradice las políticas de transferencias condicionadas “ciegas al género” criticadas por Kabeer [19], sugiriendo que la deserción femenina no fue una decisión de costo de oportunidad laboral, sino una expulsión sistémica hacia la esfera doméstica.

Es imperativo delimitar el alcance epistemológico de estos hallazgos: los algoritmos de Machine Learning aplicados en este estudio están diseñados para maximizar la precisión predictiva, no para establecer inferencia causal estricta. El

modelo de Gradient Boosting identifica de manera excepcional “**dónde**” y “**en quiénes**” se concentra la erosión de la matrícula, optimizando la capacidad de alerta temprana. Sin embargo, la interpretación de las causas subyacentes de la deserción -**como la carga desproporcionada sobre las adolescentes mujeres, confirmada por la asociación estadística de la prueba Chi-Cuadrado**- no emana del algoritmo per se, sino que se fundamenta teóricamente en marcos sociológicos como la “**Economía del Cuidado**”. Esta distinción entre la predicción algorítmica y la causalidad teórica es vital para el diseño de políticas públicas integrales.

Finalmente, la incapacidad de la educación remota para sostener la matrícula en zonas rurales (donde el 85% carece de internet) confirma empíricamente la “falacia tecnológica” advertida por Warschauer [20]. Este hallazgo refuta las proyecciones tecno-optimistas de organismos como el Banco Interamericano de Desarrollo (BID) [21] que apostaban por un “salto tecnológico” (leapfrogging) [22] capaz de amortiguar el cierre escolar. La tecnología no actuó como un igualador, sino como un amplificador de desigualdades preexistentes (el “Efecto Mateo”: a quien tiene, se le dará más), cumpliéndose la “Ley de la Amplificación” descrita por Toyama [23], quien sostiene que la tecnología digital no puede sustituir la falta de infraestructura institucional humana.

A diferencia de los hallazgos en zonas urbanas reportados por la OECD [7], donde la virtualidad mitigó el daño, en la ruralidad hondureña la correlación entre falta de infraestructura y deserción fue absoluta. Esto refuta la narrativa tecnocrática de que la provisión de dispositivos per se resuelve la crisis educativa; sin infraestructura habilitante, el hardware es irrelevante.

La presente investigación logró cuantificar el impacto estructural de la pandemia COVID-19 en la educación media rural de Honduras mediante el procesamiento de 315,467 registros, permitiendo derivar las siguientes conclusiones alineadas con los objetivos e hipótesis del estudio:

- Se concluye que el choque pandémico generó un comportamiento no lineal en la matrícula y retención escolar que los métodos econométricos tradicionales son incapaces de capturar. Mientras la Regresión Lineal falló al predecir variables críticas como la Repitencia ($R^2 \approx 0.001$), el modelo de Árboles de Decisión con Gradient Boosting demostró una superioridad técnica indiscutible para modelar la crisis, alcanzando una precisión predictiva del 98.2% ($R^2 = 0.9825$) en la matrícula final y minimizando el Error Absoluto Medio (MAE) a 0.038. Esto valida que las técnicas avanzadas de Minería de Datos son superiores para la gestión educativa en entornos de alta incertidumbre.

- Se rechaza la hipótesis nula de independencia entre género y deserción (H_0). La prueba de Chi-Cuadrado de Pearson ($X^2 = 104.73$, $p < 0.001$) confirma que el abandono escolar en la zona rural no fue neutro, sino que penalizó desproporcionadamente a las mujeres. A diferencia de las crisis económicas tradicionales donde la deserción suele ser masculina (mercado laboral), el análisis de residuos del modelo

sugiere que la crisis sanitaria activó dinámicas de "economía del cuidado" que expulsaron a las estudiantes del sistema, identificando al género femenino como el grupo de mayor vulnerabilidad en este contexto.

- El análisis exploratorio y las correlaciones del modelo validan que la falta de infraestructura tecnológica fue un determinante estructural de la exclusión. En zonas donde la conectividad efectiva reportada es inferior al 15%, la modalidad de “teledocencia” no garantizó la continuidad pedagógica, sino que actuó como un amplificador de la desigualdad preexistente (efecto histéresis). Los municipios con menor índice de desarrollo digital mostraron las tasas más altas de varianza no explicada por tendencias históricas, confirmando que la tecnología sin acompañamiento institucional es insuficiente para sostener la matrícula rural.

- La precisión predictiva alcanzada por el algoritmo de Gradient Boosting permite transitar de un modelo de gestión reactivo a uno preventivo. En respuesta al último objetivo específico, se propone la implementación del Programa de Equidad Educativa Integral (PEEI), fundamentado en este modelo predictivo. Esta herramienta de “Alerta Temprana” permitiría identificar con alta precisión, antes del inicio del ciclo lectivo, qué centros educativos rurales presentan mayor riesgo de erosión en su matrícula femenina, optimizando así la asignación de becas y recursos de retención de manera focalizada.

AGRADECIMIENTOS

El autor expresa su agradecimiento a la Unidad del Sistema Nacional de Información Educativa de Honduras (USINIEH) de la Secretaría de Educación (SEDUC) por la disposición y transparencia en el acceso a los microdatos administrativos anonimizados del periodo 2015-2023, insumo fundamental para la realización de este estudio.

Asimismo, se extiende el reconocimiento a la Dirección de Investigación de la Universidad Tecnológica Centroamericana (UNITEC) por el soporte académico y metodológico brindado en el marco del programa de Doctorado en Economía y Negocios, facilitando los recursos necesarios para el procesamiento de datos de alta complejidad.

REFERENCIAS

- [1] G. S. Becker, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago: University of Chicago Press, 2009.
- [2] Observatorio Universitario de Educación de la UNAH, «Informe sobre la Situación de la Educación en Honduras 2023.» 2023.
- [3] Banco Mundial, «Education in the time of COVID-19: A call to action for education systems in Latin America and the Caribbean.» 2021.
- [4] FOSDEH, «Foro Social de la Deuda Externa y Desarrollo de Honduras», Foro Social de la Deuda Externa y Desarrollo de Honduras, 2022.
- [5] J. R. Rodríguez Rivera, «Explorando el impacto de la pandemia COVID-19 en la educación media de las zonas rurales de Honduras (2015-2023)», UNITEC, Tegucigalpa, Honduras, 2024.
- [6] F. Provost y T. Fawcett, *Data science for business: what you need to know about data mining and data-analytic thinking*, 1. ed., 2. release. Beijing Köln: O'Reilly, 2013.

- [7] S. C. Albright y W. L. Winston, *Business analytics: data analysis and decision making*, Seventh Edition. Boston, MA: Cengage, 2020.
- [8] R. Hernández Sampieri y C. P. Mendoza Torres, *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta*, First edition. México: McGraw-Hill Education, 2018.
- [9] L. Joyanes Aguilar, *Inteligencia de negocios y analítica de datos: una visión global de Business Intelligence & Analytics*, Primera edición. Barcelona: Marcombo, 2019.
- [10] D. C. Montgomery, E. A. Peck, y G. G. Vining, *Solutions manual to accompany Introduction to linear regression analysis*. Hoboken, N.J.: J. Wiley & Sons, 2013.
- [11] H. R. Varian, «Big Data: New Tricks for Econometrics», *Journal of Economic Perspectives*, vol. 28, n.º 2, pp. 3-28, may 2014, doi: 10.1257/jep.28.2.3.
- [12] E. Paz-Maldonado, H. Flores-Girón, y I. Silva-Peña, «Educación y desigualdad social: El impacto de la pandemia COVID-19 en el sistema educativo público de Honduras», *EPAA*, vol. 29, n.º August-December, p. 133, oct. 2021, doi: 10.14507/epaa.29.6290.
- [13] J. D. Angrist y J.-S. Pischke, *Mostly harmless econometrics: an empiricist's companion*. Princeton, New Jersey Oxford: Princeton University Press, 2009.
- [14] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, y S. Mullainathan, «Human Decisions and Machine Predictions», *The Quarterly Journal of Economics*, vol. 133, n.º 1, pp. 237-293, feb. 2018, doi: 10.1093/qje/qjx032.
- [15] S. Athey, «Beyond prediction: Using big data for policy problems», *Science*, vol. 355, n.º 6324, pp. 483-485, feb. 2017, doi: 10.1126/science.aal4321.
- [16] F. H. G. Ferreira y N. Schady, «Aggregate Economic Shocks, Child Schooling, and Child Health», *The World Bank Research Observer*, vol. 24, n.º 2, pp. 147-181, ago. 2009, doi: 10.1093/wbro/lkp006.
- [17] United Nations Women, *From Insights to Action: Gender Equality in the Wake of COVID-19*. New York: United Nations, 2020. doi: 10.18356/f837e09b-en.
- [18] T. Alon, M. Doepke, J. Olmstead-Rumsey, y M. Tertilt, «The Impact of COVID-19 on Gender Equality», National Bureau of Economic Research, Cambridge, MA, w26947, abr. 2020. doi: 10.3386/w26947.
- [19] N. Kabeer, «Gender, poverty, and inequality: a brief history of feminist contributions in the field of international development», *Gender & Development*, vol. 23, n.º 2, pp. 189-205, may 2015, doi: 10.1080/13552074.2015.1062300.
- [20] M. Warschauer, *Technology and Social Inclusion: Rethinking the Digital Divide*. Cambridge: The MIT Press, 2004.
- [21] E. Arias Ortiz et al., «Education in Times of Coronavirus: Latin America and the Caribbean's Education Systems in the Face of COVID-19», Inter-American Development Bank, may 2020. doi: 10.18235/0002337.
- [22] M. Khosrow-Pour, D.B.A., Ed., *Encyclopedia of Information Science and Technology, Fourth Edition*: IGI Global, 2018. doi: 10.4018/978-1-5225-2255-3.
- [23] K. Toyama, *Geek heresy: rescuing social change from the cult of technology*, First edition. New York: PublicAffairs, 2015.