

<https://doi.org/10.69639/arandu.v12i4.1835>

Un estudio de viabilidad y un marco metodológico que aplica modelos de aprendizaje automático en el nexo Agua – Energía – Alimentos en Galápagos: Oportunidades y desafíos ante la escasez de datos

A feasibility study and methodological framework that applies machine learning models to the Water – Energy – Food nexus in Galapagos: Opportunities and challenges in the face of data scarcity

María Elena Jiménez Copá
ejimenezc@est.unibe.edu.ec
<https://orcid.org/0009-0009-5376-9415>
Universidad Iberoamericana del Ecuador
Ecuador – Quito

Yasmany Fernández Fernández
yfernandezf@upec.edu.ec
Universidad Politécnica Estatal del Carchi
Ecuador – Tulcán

*Artículo recibido: 18 noviembre 2025 -Aceptado para publicación: 28 diciembre 2025
Conflictos de intereses: Ninguno que declarar.*

RESUMEN

En una época marcada por la creciente interconexión global y por los impactos evidentes del cambio climático, el enfoque del nexo WEF (agua-energía-alimentos) se posiciona como un instrumento fundamental para evaluar y anticipar escenarios orientados a la sostenibilidad. Este enfoque se articula directamente con la Agenda 2030 y sus Objetivos de Desarrollo Sostenible. El estudio analiza cómo el uso de técnicas de aprendizaje automático y modelos de *machine learning* puede contribuir a mejorar y gestionar de forma eficiente las relaciones entre estos sistemas y anticipar escenarios futuros a partir del análisis del nexo. La aplicación de KDD como metodología de implementación permitió la recopilación y limpieza de los datos, fases en las que se evidenciaron vacíos en las series. Para completar las series de datos de la producción y energía, se aplican interpolaciones y extrapolaciones lineales. Se validan 7 modelos de regresión supervisada (Regresión Lineal, Random Forest, SVR, Red Neuronal Artificial, Gradient Boosting, XGBoost, KNN). De los modelos evaluados, XGBoost presenta el mejor desempeño y bajo error $R^2 = 0.92$. No obstante, las estimaciones a diez años muestran que la producción agrícola tiende a mantenerse prácticamente sin variaciones, esto se atribuye a la escasez de datos y a su poca variabilidad, lo que evidenció 5 limitantes, las mismas que al ser tratadas de manera técnica le dotan al modelo una alta estabilidad ($R^2 = 0.965 \pm 0.025$), y una baja incertidumbre de

\pm 20.000 kg, reforzando su confiabilidad como herramienta de alerta temprana en el sistema agrícola de Galápagos.

Palabras clave: nexo, machine learning, datos

ABSTRACT

In an era marked by increasing global interconnectedness and the evident impacts of climate change, the WEF nexus approach (water-energy-food) is positioned as a fundamental tool for assessing and anticipating sustainability-oriented scenarios. This approach is directly aligned with the 2030 Agenda and its Sustainable Development Goals. The study analyzes how the use of machine learning techniques and models can contribute to improving and efficiently managing the relationships between these systems and anticipating future scenarios based on nexus analysis. The application of KDD as the implementation methodology allowed for data collection and cleaning, phases in which gaps in the data series became evident. To complete the production and energy data series, linear interpolations and extrapolations were applied. Seven supervised regression models were validated (Linear Regression, Random Forest, SVR, Artificial Neural Network, Gradient Boosting, XGBoost, KNN). Of the models evaluated, XGBoost presents the best performance and low error $R^2 = 0.92$. However, ten-year estimates show that agricultural production tends to remain practically unchanged. This is attributed to the scarcity of data and its low variability, which revealed 5 limitations. When these limitations are addressed technically, they give the model high stability ($R^2 = 0.965 \pm 0.025$) and a low uncertainty of $\pm 20,000$ kg, reinforcing its reliability as an early warning tool in the Galapagos agricultural system

Keywords: nexus, machine learning, data

Todo el contenido de la Revista Científica Internacional Arandu UTIC publicado en este sitio está disponible bajo licencia Creative Commons Atribution 4.0 International. 

INTRODUCCIÓN

En un escenario orientado a la sostenibilidad y marcado por crisis en los recursos hídricos, energéticos y alimentarios, la medición del nexo agua-energía-alimentos (WEF) en una región específica cobra especial relevancia debido a la vulnerabilidad de cada sector (Sánchez-Zarco et al., 2021). La interacción entre el nexo WEF y los servicios ecosistémicos urbanos (SEU) genera un impacto significativo, por lo que resulta esencial estudiar su correlación y efectos en la población (Ding et al., 2023).

El enfoque del nexo agua–energía–alimentos (WEF) aborda la forma en que estos tres sistemas se relacionan e influyen entre sí de manera dinámica. Su interacción no solo es inseparable, sino que además varía de acuerdo con el espacio y el tiempo, lo que evidencia que la administración sectorial tradicional resulta limitada para enfrentar los problemas globales actuales. Entre estos desafíos se incluyen el cambio climático, la reducción de recursos disponibles, el deterioro de los ecosistemas y la creciente preocupación por la seguridad alimentaria. (Simpson & Jewitt, 2019).

En América Latina y el Caribe, una proporción significativa de las actividades económicas se sustenta en la explotación y uso intensivo de los recursos naturales. En este marco, el sector agrícola depende del recurso hídrico para garantizar el riego y el adecuado desarrollo de los cultivos. De manera paralela, el sector energético requiere agua tanto para la generación hidroeléctrica como para los procesos de refrigeración asociados a diversas tecnologías de producción. Asimismo, la energía constituye un insumo esencial para la captación, transporte y distribución del agua destinada al consumo humano y a la operación de múltiples actividades productivas, entre ellas la agricultura, la minería y la extracción de hidrocarburos. (Naranjo & Willaarts, 2020), Dada la importancia del nexo en América Latina y el Caribe, se considera prioritario el establecimiento de Políticas, inversiones público-privadas que mejoren la gobernanza de los recursos naturales. Durante el foro Virtual “Desafíos para la Planificación y el Monitoreo de la Agenda 2030 en América Latina y el Caribe”, Alba Llavona representante de la Unidad de Agua y Energía de la División de Recursos Naturales de la CEPAL, presentó el estudio “Implementación de políticas con enfoque nexo en ALC: indicadores y políticas basadas en evidencia para el caso de Bolivia”. En su intervención subrayó la importancia del enfoque Nexo para avanzar en la Agenda 2030 y en los Objetivos de Desarrollo Sostenible (Naciones Unidas, 2018), señalando que esta perspectiva favorece el cumplimiento de los ODS de manera más eficiente y con menores costos, gracias a la generación de sinergias positivas. (Gil & Llavona, 2020), Asimismo, destacó que adoptar este enfoque disminuye la posibilidad de que las acciones orientadas a distintos ODS se contrapongan entre sí, contribuyendo a una gestión responsable de los recursos (Weitz, Nilsson & Davis, 2014). De acuerdo con la CEPAL (2016), avanzar hacia la Agenda 2030 requiere impulsar transformaciones ambientales a través de inversiones estratégicas.

que minimicen los impactos ecológicos. (CEPAL, 2016). Entre los ODS más directamente vinculados al enfoque Nexo se encuentran el Objetivo 2 (Hambre Cero), el Objetivo 6 (Agua Limpia y Saneamiento) y el Objetivo 7 (Energía Asequible y No Contaminante). (Naranjo & Willaarts, 2020), (Gil & Llavona, 2020)

En 2018, el Estado ecuatoriano incorporó oficialmente la Agenda 2030 dentro de sus políticas públicas, marcando un paso decisivo para avanzar en el cumplimiento de los Objetivos de Desarrollo Sostenible. Desde entonces, la Secretaría Nacional de Planificación ha integrado los ODS como un instrumento central de orientación estratégica, utilizándolos como guía para promover un modelo de desarrollo equilibrado, inclusivo y sostenible. (Secretaría Nacional de Planificación, 2021)

En este marco, la implementación del enfoque Nexo adquiere un papel fundamental, ya que permite una gestión más coherente y eficiente de los recursos naturales, además de facilitar el cumplimiento de los compromisos establecidos tanto en la normativa nacional como en las agendas internacionales adoptadas por los países de América Latina y el Caribe.

Ecuador cuenta con regiones ecológicamente sensibles como las Islas Galápagos, donde el desarrollo sostenible es particularmente difícil debido a su aislamiento geográfico, fragilidad ecológica, exposición al cambio climático y a la globalización (Douglas, 2006).

La condición geográfica y geológica de las Islas Galápagos impide el desarrollo de un sistema hídrico continuo o extenso, por lo cual su disponibilidad es crítica, ya que la recarga del sistema hidrológico depende principalmente de la lluvia y la garúa (niebla en las tierras altas), las cuales se agotan por la evaporación (Echeverría et al., 2024). La producción energética depende en gran medida de infraestructuras vulnerables y de la importación de combustibles fósiles; esto hace que el sistema energético sea frágil e insostenible a largo plazo (Llerena Pizarro et al., 2019). En el archipiélago, la producción alimentaria enfrenta grandes retos, ya que es necesario desarrollar sistemas de producción local sostenibles para reducir los alimentos importados (International Partnership for the Satoyama Initiative (IPSI), 2025).

Nuestro estudio pretende demostrar cómo las tecnologías emergentes, como el aprendizaje automático, pueden ayudar a analizar la recuperación de estos sistemas ante diversas perturbaciones ambientales, económicas y sociales.

La inteligencia artificial abarca un conjunto de métodos y algoritmos capaces de permitir que los sistemas computacionales aprendan de manera autónoma a partir de la información disponible. Estas herramientas son altamente valoradas y se aplican ampliamente en la solución de problemáticas vinculadas con los sistemas ambientales. Mediante el aprendizaje automático, se extraen conocimientos y patrones de los datos, que se utilizan para resolver problemas y encontrar las mejores condiciones con el propósito de optimizar el uso de los recursos y promover alternativas ambientalmente responsables (Shen et al., 2022). Los modelos a evaluar en el

presente estudio son: Regresión Lineal, Random Forest, SVR, Red Neuronal, Gradient Boosting, XGBoost y KNN.

La aplicabilidad de este tipo de modelos se ha hecho presente en varias investigaciones para predecir demandas de agua, energía, cadenas de suministro alimentaria y evaluar impactos ecosistémicos, sin embargo, aún existen vacíos en lo que respecta al análisis y desarrollo de modelos que apoyen en la predicción y optimización del NEXUS específicamente en entornos insulares y biodiversos como Galápagos.

Por consiguiente, el presente estudio propone un modelo de predicción de la producción agrícola local en los próximos 10 años, a partir de su relación con la energía, con la finalidad de evaluar el efecto en conjunto sobre los alimentos y de esta forma establecer un modelo de ML que responda satisfactoriamente a los desafíos particulares de esta región.

MATERIALES Y MÉTODOS

La estrategia metodológica adoptada es KDD; esta metodología busca usar el conocimiento descubierto para apoyar la toma de decisiones, para lo cual usa datos numéricos, aplica métodos estadísticos y modelos de Machine Learning. Ha sido aplicada en algunas investigaciones relacionadas con la agricultura, ya que permite extraer información que se encuentra oculta en variables típicas de este entorno, como son: el agua, la energía, el clima (Bagal et al., 2020). Al permitir la gestión de datos incompletos o ruidosos mediante su etapa de limpieza o preprocesamiento de datos, presenta una mejora significativa en los resultados predictivos (Danubianu, 2014). Al integrar una fase para la minería de datos, se permite la integración y aplicación de algoritmos como Random Forest, haciendo de esta metodología una herramienta óptima para alcanzar los objetivos planteados.

Figura 1

Metodología KDD



Elaboración Propia

Recopilación de datos

Por las exigencias propias de esta investigación, se requirió un esfuerzo significativo para reunir datos provenientes de fuentes acreditadas. En consecuencia, se efectuó una evaluación rigurosa de las diversas fuentes identificadas para garantizar la validez y consistencia de los datos,

por lo cual se aplica un criterio de exhaustividad y completitud para elegir las fuentes que presentan datos más completos, unidades estandarizadas y, en algunos casos, fuentes oficiales.

A continuación, el detalle de cada variable empleada en la evaluación del nexo:

- La población, la población anual de Galápagos y su tasa de crecimiento se toman de la página de la revista “Ecuador Galápagos Info”. La misma que presenta la población anual únicamente de ciertos años (2001, 2010, 2015 y 2022), por lo cual fue necesario completar la serie de años intermedios y extrapolar hacia los años 2023 y 2024. Lo anterior se logró utilizando una interpolación de tipo exponencial fundamentada en el cálculo del crecimiento anual compuesto entre los censos sucesivos, proceso que será explicado en la sección correspondiente a la limpieza de datos.
- Precipitación (mm), los registros fueron obtenidos del portal de datos abiertos del INAMHI, entidad que proporciona estadísticas de precipitación mensual generadas por estaciones meteorológicas en todo el país. En el análisis preliminar se detectaron meses sin datos disponibles, por lo cual, en la fase de limpieza se aplica la técnica estadística “Interpolación temporal lineal”, aplicada para completar esta información.
- Consumo energético (kwh), los datos se obtienen de la empresa eléctrica de Galápagos, del “Plan de Transición Energética de las Islas Galápagos”, en este plan se presentan los datos de energía eléctrica total consumida por año, desde el 2007 al 2022 desglosadas por grupo de consumo (comercial, industrial, residencial, otros), no se cuenta con un consumo desagregado por sector, por lo cual se usará el consumo energético total anual como proxy para la disponibilidad energética, para reducir el sesgo que se introduce por todos los grupos mencionados, normalizaremos la energía total por la población anual (energía per cápita). El desarrollo de la misma se realizará en la fase de limpieza.

Los datos publicados en el plan de transición energética van desde 2007 hasta 2022. Para cumplir con el intervalo de estudio (2005-2024) y obtener el consumo energético de los años faltantes, se aplica la técnica “Extrapolación Lineal”, la cual se desarrolla en la fase de limpieza.

- Producción de alimentos local (kg), los datos base se obtuvieron de estudios científicos realizados sobre los beneficios de la producción local frente a la importación de alimentos, seguridad hídrica y los sistemas agrícolas bajo escenarios de uso de suelo y cambio climático (Sampedro et al., 2020), el Plan de reactivación Galápagos 2030, en el que se estima una producción anual de 563 toneladas para 2014 (Gobierno de Galápagos, 2020), y análisis centrados en determinar las condiciones del agua de origen natural como base para la formulación de prácticas agrícolas sostenibles, en el que se menciona la producción anual de 575 toneladas para 2017 (Carrión-Mero et al., 2024). Esto en una superficie productiva agropecuaria de 19.010 ha para cultivos de ciclo corto (cebolla, tomate, etc.). Dado que solo se cuenta con dos puntos de anclaje, se procede a completar los datos

aplicando técnicas estadísticas de interpolación y extrapolación lineal, las cuales se desarrollan en la fase de limpieza.

Para completar las series de datos se opta por la aplicación de interpolación lineal y extrapolación simple por sobre otras como series de tiempo o regresión múltiple debido a las siguientes consideraciones propias del estudio:

- Escasez de datos: al no contar con suficiente información la aplicación de métodos más avanzados no sería factible ya que requieren de suficientes observaciones para identificar patrones y tendencias
- Simplicidad y trazabilidad metodológica: estos modelos aseguran resultados consistentes fáciles de interpretar sin sobreajuste.

Limpieza

Durante esta fase se procede con la revisión de inconsistencias y la completitud de valores faltantes, se estandarizan las medidas de todas las variables del nexo, se aplican técnicas estadísticas diferentes para cada caso.

- **Población:** Para completar los datos de la serie (2005- 2024) se aplica el cálculo de la tasa anual compuesta de crecimiento poblacional conforme a lo siguiente (Ecuación 1):
 1. Cálculo de la tasa anual compuesta de crecimiento poblacional r

$$r = \left(\frac{P_1}{P_0} \right)^{\frac{1}{(y_1 - y_0)}} - 1 \quad (1)$$

Sea (y_0, P_0) y (y_1, P_1) dos puntos censales consecutivos, donde:

y_0 : año inicial

P_0 : Población en el año inicial

y_1 : año final

P_1 : población en el año final

2. Estimación de la población en años intermedios: Con la tasa r, la población para un año intermedio y tal que $y_0 < y < y_1$ se estimó como (Ecuación 2)

$$P(y) = P_0 \times (1 + r)^{y - y_0}$$

(2)

Este procedimiento se aplicó de forma independiente para los intervalos: 2001-2010; 2010-2015; 2015-2022

3. Extrapolación para años posteriores (2023, 2024), se aplica la misma tasa anual compuesta r del último intervalo 2015-2022, proyectando la población de la siguiente manera (Ecuación 3):

$$P(2023) = P_{2022} \times (1 + r)$$

$$P(2024) = P_{2022} \times (1 + r)^2$$

(3)

- **Precipitación:** Para completar los datos de los meses faltantes en esta variable, se aplica la técnica: “Interpolación temporal lineal”. (Ecuación 4)

$$P_{faltante\ k} = P_{anterior} + \frac{K}{n+1} \cdot (P_{siguiente} - P_{anterior})$$

Donde K= 1,2, ..., n (4)

- **Consumo energético:** Para determinar el consumo energético realizamos los siguientes pasos:

1. Se aplica la técnica “Extrapolación Lineal” (Ecuación 5) para estimar los datos faltantes y completar la serie en estudio (2005 -2024)

$$P(t) = P(t1) + \frac{t - t1}{t2 - t1} \cdot (P(t2) - P(t1)) \quad (5)$$

2. Se realiza la normalización de la energía total por la población anual (energía per cápita), aplicando la siguiente formula (Ecuación 6):

$$E_t^{pc} = \frac{E_t}{P_t} \quad (6)$$

Donde: E_t = Energía total consumida en Galápagos Enel año t

P_t = Población estimada en Galápagos en el año t

E_t^{pc} = Consumo energético percapita en el año t

- **Producción de alimentos:** Al contar con escasos datos históricos de la producción local en Galápagos, se procede a calcular la pendiente lineal entre los dos puntos de anclaje (Ecuación 7) posterior se realiza la interpolación lineal entre los puntos de anclaje (Ecuación 8), continuamos con la extrapolación hacia adelante (2017 – 2014) y extrapolación hacia atrás (2014 - 2005) (Ecuación 4) sobre estimaciones puntuales, Cálculo de la pendiente lineal entre los puntos de anclaje conocidos

$$m = \frac{Y2 - Y1}{X2 - X1} \quad (7)$$

Donde: x= año, y= producción

Cálculo de la interpolación/extrapolación Lineal

$$P(y) = P(y_0) + m \cdot (y - y_0) \quad (8)$$

Donde:

1. P(y): Producción estimada en el año y
2. P(y₀): Producción conocida en el año base y_0
3. M = pendiente (toneladas año)
4. $y - y_0$ = diferencia de años (positiva si proyectamos hacia adelante, negativa si hacia atras)

Transformación

Se transforma los datos a una escala adecuada para su posterior análisis

- **Precipitación:** Dado que los datos de esta variable vienen en meses y los necesitamos en años, procedemos a cambiar su representación, es decir, realizar el paso de meses a su valor en años. Para lograr este objetivo, se procede a realizar la suma de todos los meses de un año y a sacar el promedio o media aritmética. Este proceso se realiza para todos los años que serán usados en el estudio (2005-2024).
- **Consumo energético:** Los datos del consumo energético se encuentran en MWH por lo cual se procede a realizar la transformación a KWH

$$1\text{Mwh} = 1000 \text{ Kwh}$$

Selección de Variables

Para la selección, se aplica la técnica de correlación de Pearson a las variables del nexo (Agua, Energía, Producción).

La correlación de Pearson es una métrica estadística utilizada para cuantificar qué tan fuerte y en qué dirección se relacionan dos variables continuas a través de una dependencia lineal. En esencia, identifica si el incremento o disminución de una variable se acompaña de cambios similares en la otra.

Su aplicación a las variables del nexo antes de aplicar los modelos permite:

- Detectar las variables que están altamente relacionadas con la variable objetivo
- Ayuda en la optimización del modelo, ya que se puede visualizar qué variables no aportan o qué podría ingresar ruido en el modelo.
- Reduce el riesgo de sobreajuste, ya que al conocer qué variables se relacionan con más fuerza con la variable objetivo (Producción agrícola), se mejora la generalización del modelo.

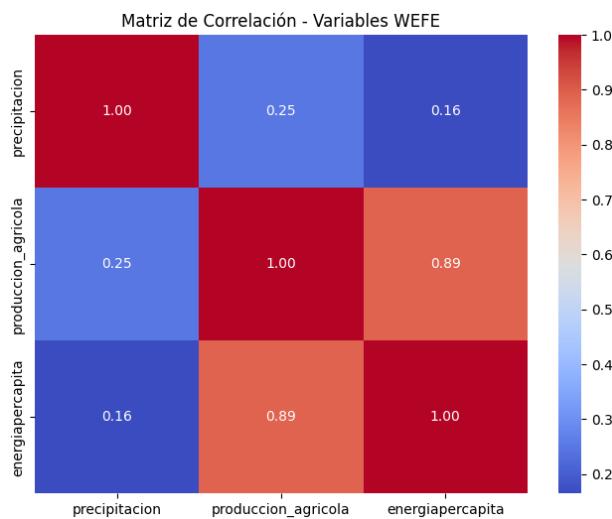
Tabla 1

Matriz de correlación

	Precipitación	Producción Agrícola	Energía Percápita
Precipitación	1.00	0.249	0.164
Producción Agrícola	0.249	1.00	0.8914
Energía Percápita	0.164	0.8914	1.00

Figura 2

Mapa de calor



Fuente: Elaboración propia con librerías de python

Tabla 2

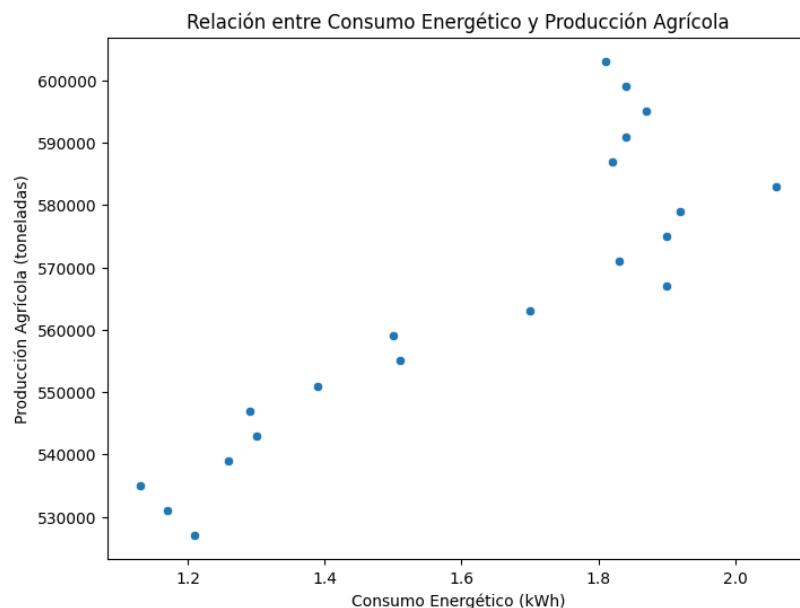
Interpretación de resultados

Relación variables	Valor	Conclusión
Precipitación/ producción_agrcola	0.25	Existe una correlación positiva leve, puede ser debido a que la producción agrícola local, depende también de otras fuentes de agua (Paltán et al., 2023)
Precipitación/energía per cápita	0.164	Correlación positiva leve: al aumentar la lluvia, el consumo energético se incrementa ligeramente por el uso de sistemas eléctricos en varias actividades agrícolas (Jaramillo Díaz et al., 2022).
Producción Agricola /Energía per cápita	0.891	Correlación positiva muy fuerte, en el contexto de Galápagos puede ser que en tiempos de sequía los agricultores acuden al agua transportada y el uso de diferentes tecnologías como Waterboxx (Jaramillo Díaz et al., 2022)

Los resultados reflejan que la producción agrícola mantiene una asociación positiva considerable con la energía per cápita., por lo cual, la producción agrícola local estaría fuertemente vinculada al consumo energético dado la escasez de recursos hidráulicos de buena calidad que apoya al adecuado riego de los cultivos, según un estudio realizado (Carrión-Mero et al., 2024) solo el 55.88% de las 34 fuentes que hay en las islas Galápagos son aptas para riego, lo que conlleva a la importación de productos y agua desde el continente. Para visualizar la relación individual de las dos variables, aplicamos un gráfico de dispersión, teniendo los siguientes resultados:

Figura 3

Gráfico de Dispersion entre las variables Consumo Energético y Producción agrícola



Fuente: Elaboración propia aplicando librerías de Python

En la gráfica no se evidencian outliers graves, se visualiza una ligera dispersión cuando el consumo energético supera los 1,8 Kwh per cápita; sin embargo, los valores siguen el patrón esperado.

La correlación observada de Pearson con un valor igual a 0.891 entre la producción agrícola y la energía per cápita es consistente con la necesidad energética elevada en el sistema agrícola de Galápagos, por lo cual nuestra variable independiente será la energía per cápita

Minería de datos

Esta etapa es el corazón de la metodología, ya que en ella se aplican los modelos en estudio a los datos previamente seleccionados para realizar predicciones basadas en un modelo establecido.

Con la finalidad de tener una amplia gama de resultados y seleccionar el modelo que nos permita alcanzar nuestro objetivo “Predecir la producción agrícola local de Galápagos de los próximos 10 años con base en la energía per cápita utilizada”, se proponen 7 modelos de regresión supervisada, todos eficaces en el análisis de relaciones lineales y no lineales.

Regresión Lineal: Consiste en una aproximación lineal destinada a explicar cómo una variable dependiente se relaciona con una o varias variables independientes. El proceso determina la recta óptima que describe de manera más precisa los datos disponibles. (Seber & Lee, 2012)

Random Forest: Corresponde a un método común en aprendizaje automático que fusiona las decisiones generadas por múltiples árboles para generar un único resultado, permitiendo abordar tanto análisis de clasificación como de regresión. (Breiman, 2001)

Support Vector Regression (SVR): Es un modelo basado en los mismos algoritmos de las máquinas de vectores de soporte, SVR encuentra un hiperplano con el margen máximo entre los puntos de datos, todo dentro de un margen de tolerancia (Drucker et al., 1997)

Red Neuronal Artificial (ANN): Es un sistema capaz de tomar decisiones de manera parecida al cerebro humano, ya que utiliza mecanismos inspirados en el trabajo conjunto de las neuronas para reconocer patrones, evaluar alternativas y generar conclusiones. (Haykin, 1999)

Gradient Boosting: Es un método de aprendizaje automático aplicado tanto en regresión como en clasificación estadística, que construye un modelo predictivo combinando múltiples predictores simples —generalmente árboles de decisión (Friedman, 2001).

XGBoost: Este método se basa en árboles de decisión y supone una mejora sobre otros métodos, como el bosque aleatorio y refuerzo de gradientes. Funciona bien con datasets grandes y complejos al utilizar varios métodos de optimización. (Chen & Guestrin, 2016).

K-Nearest Neighbors (KNN): Es un método de clasificación que decide la categoría de un dato nuevo basándose en las clases a las que pertenecen sus vecinos más próximos dentro del conjunto de datos. (Cover & Hart, 1967)

Con los algoritmos seleccionados se procede con la implementación de cada uno mediante la herramienta Google Colab.

Para el entrenamiento se utilizó el 80% del conjunto de datos y el 20% para la validación. Para cada algoritmo, se entrenó el modelo con los datos históricos de consumo de la energía per cápita como variable predictora y la producción como variable objetivo.

A cada modelo se le calcularon las métricas de evaluación MAE, RMSE y R², las cuales se emplean en análisis de datos y estadística para valorar la precisión y desempeño de los modelos de regresión y predicción.

Mae (Mean Absolute Error) resume cuánto difieren, en promedio, los valores que predice un modelo respecto a los observados, tomando únicamente la distancia absoluta entre ambos (Ecuación 9). Esta medida refleja la magnitud típica de los errores cometidos durante la predicción y mantiene las mismas unidades que la variable analizada. (Cover & Hart, 1967)

$$MAE = \frac{1}{norte} \sum_{yo=1}^{norte} |y_i - \hat{y}_i| \quad (9)$$

Dónde y_i = valor real; \hat{y}_i = valor previsto; norte = número total de observaciones.

RMSE (Root Mean Squared Error): Es una métrica de evaluación de precisión comúnmente utilizada en el análisis de regresión que mide la magnitud promedio de los errores en un modelo de regresión (Ecuación 10). Esta métrica indica la desviación estándar de los errores del modelo.

Un RMSE más bajo sugiere menores errores de predicción promedio y, por lo tanto, predicciones más precisas (Chai & Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

(10)

Donde:

- y_i , es el valor real de la observación i .
- \hat{y}_i , es el valor previsto para la observación i .
- n es el número total de observaciones.

R2 (Coeficiente de determinación): es un indicador numérico que muestra qué tan bien las variables explicativas de un modelo logran representar las variaciones observadas en la variable que se desea predecir (Ecuación 11). Su rango oscila entre 0 y 1, cuanto más cerca está de 1, mejor es el ajuste del modelo (Theil, 1971).

La fórmula para calcular es:

$$R^2 = 1 - \frac{\text{Variación inexplicable}}{\text{Variación total}}$$

(11)

Estas métricas permiten comparar la precisión y capacidad de generalización de cada algoritmo, con la finalidad de identificar cuál proporciona la mejor predicción para el caso de estudio en Galápagos.

RESULTADOS Y DISCUSIÓN

Tras definir MAE, RMSE y R^2 como las métricas que permitirán evaluar el desempeño estadístico y predictivo de cada modelo de regresión, se exponen a continuación los resultados alcanzados por cada uno de los algoritmos analizados:

Tabla 3

Resultados de cada algoritmo para cada una de las métricas de evaluación

Modelo	MAE ↓	RMSE ↓	$R^2 \uparrow$	Desempeño
Regresión Lineal	10849.40	11272.72	0.86	Bueno, modelo base sólido
Random Forest	11961.19	13578.55	0.81	Desempeño más bajo que Regresión Lineal, puede ser porque son pocos datos
SVR	30994.60	31553.53	-0.025	Muy mal desempeño
Red Neuronal	559963.31	560829.63	-322.92	Muy mal desempeño
Gradient Boosting	8003.55	10156.32	0.8938	Buen desempeño, mejor que Regresión Lineal y Random Forest
XGBoost	8000.06	8485.33	0.9258 R^2	Muy buen desempeño, bajo error y muy buen R^2

Modelo	MAE ↓	RMSE ↓	R ² ↑	Desempeño
KNN (k=3)	16,000.00	17,400.61	0.9527	Muy buen desempeño, captura muy bien la varianza de los datos, pero los errores en escala son más grandes

Figura 4

Gráfico de los datos obtenidos de la métrica MAE

Comparación de Modelos de Machine Learning

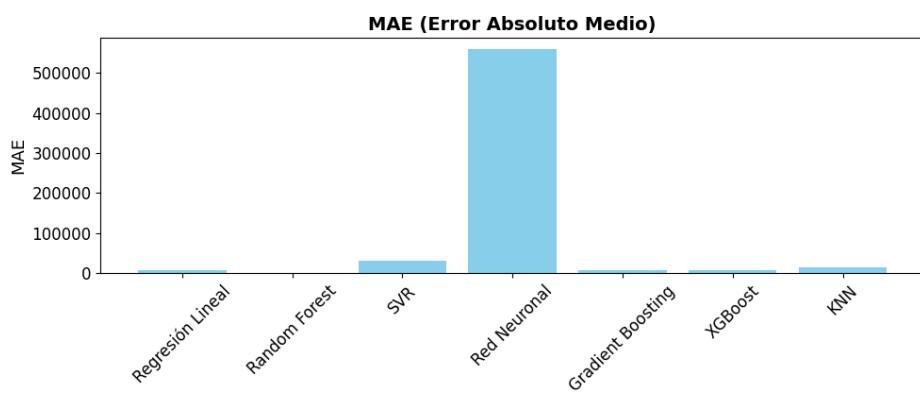


Figura 5

Gráfico de los datos obtenidos de la métrica RMSE

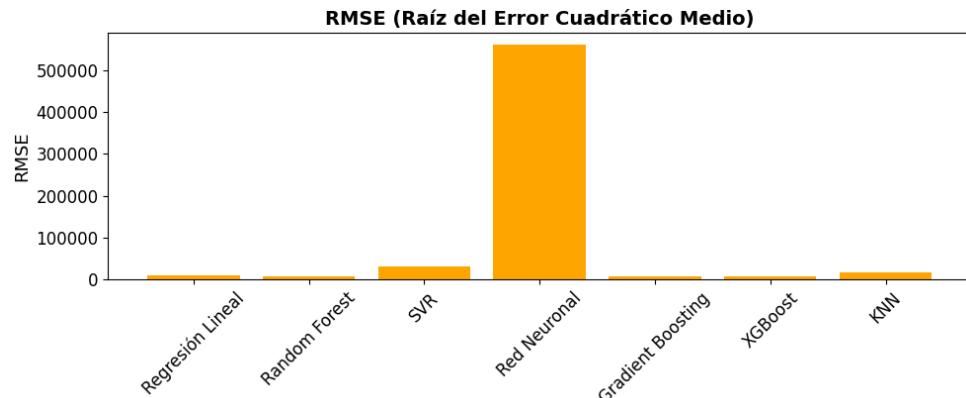


Figura 6

Gráfico de los datos obtenidos de la métrica R^2

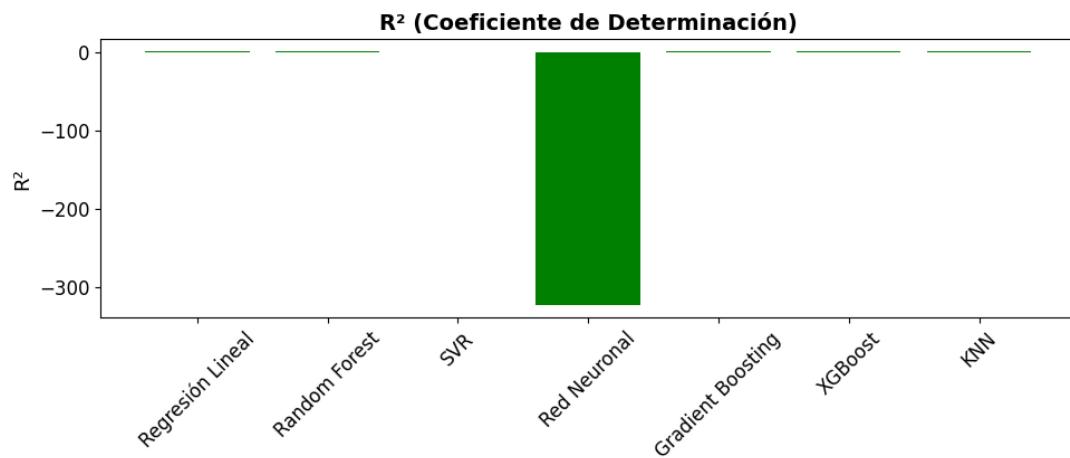
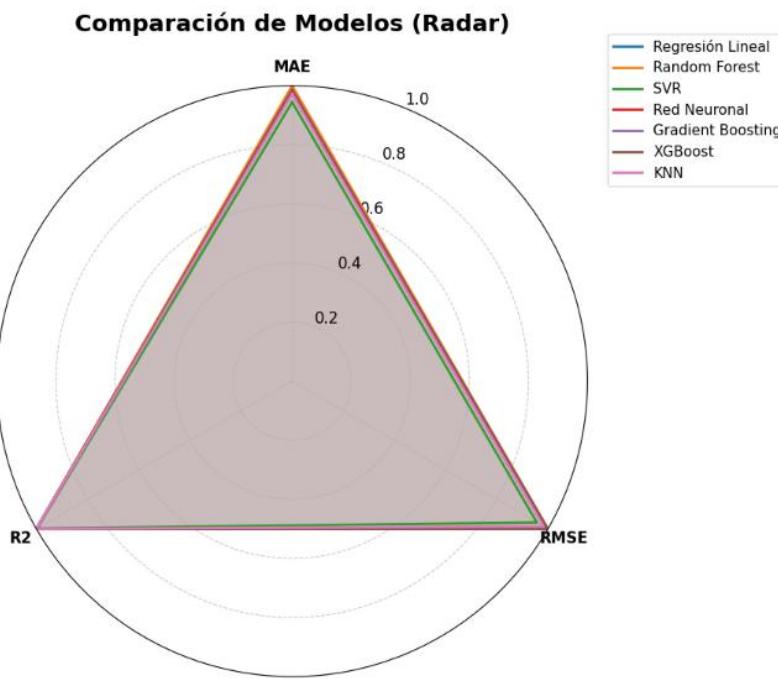


Figura 7

Comparación de modelos



Este tipo de gráficos nos indica que cuanto más cerca del borde externo estén las líneas del modelo, mejor es su desempeño general.

Conforme a los datos obtenidos, observamos que KNN ofrece un mejor R^2 pero con errores muy altos, por lo cual podríamos concluir que el modelo es arriesgado para predicciones a futuro. Sin embargo, **XGBoost** generaliza bien con pocos datos y puede extrapolar hacia el futuro con mayor estabilidad que KNN. Por lo mencionado, el modelo más adecuado es XGBoost.

Por lo expuesto y los resultados positivos, se procede con la aplicación del algoritmo XGBoost como modelo para la predicción de la producción agrícola local en Galápagos para los próximos 10 años, ya que se confirma que el modelo es muy robusto en problemas de predicción en el tiempo, incluso con pocos datos.

Para generar las predicciones con el modelo XGBoost se aplican las librerías xgboost, XGBRegressor desde xgboost.sklearn, MinMaxScale de sklearn.preprocessing para realizar la normalización de escalas y prevenir datos planos por la diferencia de escalas entre la variable objetivo (Producción) y la variable independiente (energiapercapita). Se trabaja con el entorno de desarrollo Google Colab, teniendo los siguientes resultados:

Tabla 4

Resultado de aplicar ML

	Consumo	Producción
	Energético	Agrícola Predicha
2025	1.81	593899.0625
2026	1.86	591001.9375
2027	1.92	579381.3125
2028	1.97	579381.3125
2029	2.03	579381.3125
2030	2.08	579381.3125
2031	2.14	579381.3125
2032	2.19	579381.3125
2033	2.25	579381.3125
2034	2.31	579381.3125

Figura 8

Consumo energético Proyectado

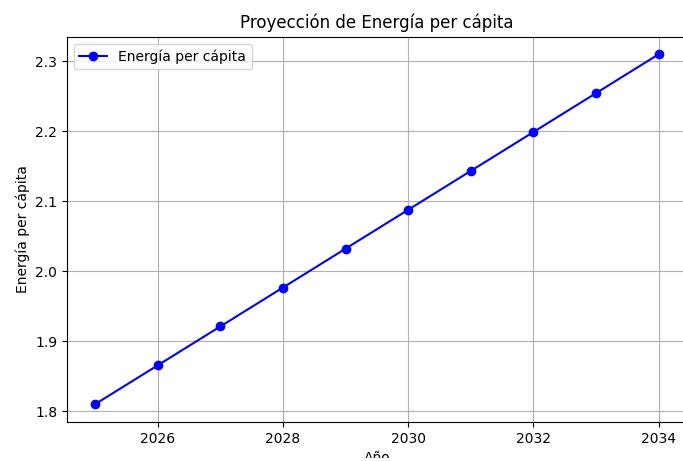
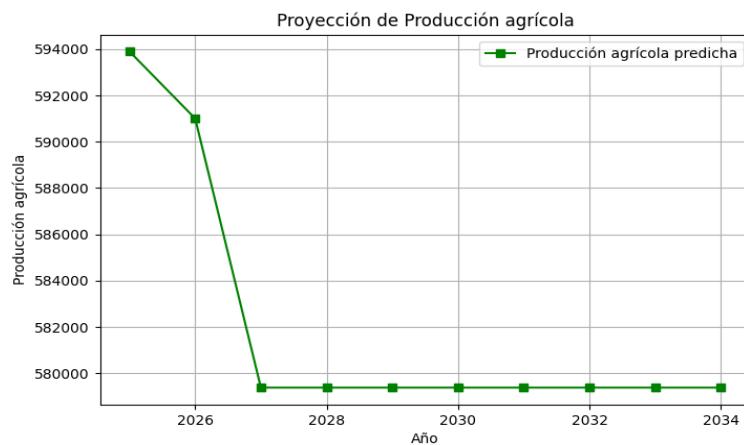


Figura 9

Producción Agrícola proyectada



Como se puede observar en la gráfica (Figura 9), la energía per cápita tiene un crecimiento gradual en los próximos 10 años; sin embargo, la producción agrícola (Figura 10) muestra una ligera caída, y a partir de 2027 se mantiene constante sin variaciones significativas a pesar de que la energía per cápita aumenta, lo que nos muestra que el conjunto de datos actual no permite realizar predicciones significativas y realistas.

La proyección uniforme surge como resultado de la limitada disponibilidad de información y del uso de métodos de interpolación para llenar los valores faltantes. Este resultado permitió analizar las siguientes limitaciones con las que se contó al comenzar el estudio:

1. **Escasez y discontinuidad de datos:** Este factor implicó la necesidad de recurrir a interpolaciones y extrapolaciones para completar los valores faltantes y así mantener la continuidad de la serie, sin embargo, se denota una afectación en la capacidad predictiva del modelo a largo plazo, lo cual se podría mejorar aplicando métodos como Bootstrapping o imputación múltiple en futuros estudios.
2. **Análisis técnico:** Se realiza un diagnóstico comprensivo de la Calidad de Datos que incluye técnicas como: Análisis de patrones de missing values, Test de aleatoriedad de datos faltantes (Little's MCAR) y Evaluación de consistencia temporal.
3. **Resultados Obtenidos:** El análisis reveló un puntaje de calidad de 0.75 (riesgo medio), confirmando que las interpolaciones introducen un margen de error del 20-25% en las predicciones.
4. **Posibles relaciones espurias entre las variables:** Por la correlación Alta entre energía per cápita y producción agrícola
5. **Ánalisis Técnico:** Se aplica Análisis Causal con Knowledge Graphs que incorpora: Grafos causales basados en conocimiento experto, Test de causalidad de Granger y Modelos de ecuaciones estructurales.

6. **Resultados Obtenidos:** Se identificó que aproximadamente entre un 15 % y un 25% de la correlación observada representa una relación causal real. Mientras que el 75 – 85 % restante se explica por tendencias temporales paralelas y variables omitidas.
7. **Proyecciones Planas no Realistas:** Proyección plana hasta el 2034 por falta de variabilidad en los datos de entrada.
8. **Análisis técnico:** Para validar la robustez del modelo se aplica el modelo bayesiano con: efectos jerárquicos por año, Splines cúbicos y componentes estacionales y de tendencia.
Resultados Obtenidos: A continuación, los escenarios proyectados para 2025-2034
 - Estabilidad productiva estimada del 65%: este escenario refleja equilibrio entre energía per capita y producción agrícola.
 - Crecimiento moderado estimado en 25%: con un crecimiento gradual de la producción de 1 a 2% anual si se mantiene un suministro de energía confiable.
 - Posible declive alrededor del 10%: puede darse un declive en la producción si existieran limitaciones energéticas o eventos climáticos adversos.

Incertidumbre no cuantificada

Análisis Técnico: Para cuantificar la incertidumbre de las proyecciones se implementa la inferencia bayesiana con cuantificación de incertidumbre mediante: distribuciones posteriores completas, intervalos de credibilidad del 89% y análisis de sensibilidad a priori.

Resultados obtenidos: Para 2034, la producción agrícola proyectada es $564,400 \pm 20,000$ kg, indicando que el rango probable oscila entre 544,400 y 584,400 kg.

Validación insuficiente: Dada la escasez de datos, es recomendable aplicar métricas para datos escasos.

Análisis técnico: Se diseña un framework de validación para series cortas que incluye: validación cruzada temporal bloqueada, pruebas con datos sintéticos, métricas de negocio específicos

Resultados obtenidos: El modelo muestra estabilidad alta media (puntaje $\approx 0.78/1.0$) y una capacidad sólida de generalización temporal, manteniendo un RMSE promedio de ≈ 1848 entre bloques y un R^2 sintético = 0.965 ± 0.025 . Demostrando robustez ante datos incompletos (22,5 % de faltantes).

Validación con escenarios hipotéticos: No se realiza la aplicación del modelo con datos externos o escenarios hipotéticos, por lo cual los resultados obtenidos son válidos dentro de la serie de datos establecida.

Cada limitación técnica identificada tiene una respuesta concreta en las recomendaciones. La aplicación conjunta de estas acciones permitirá que el modelado del nexo WEF en Galápagos evolucione de un simple análisis académico a un instrumento sólido que respalde decisiones estratégicas en materia de sostenibilidad

Recomendaciones

- A nivel del Gobierno, establecer una política para el registro anual de fuentes hídricas y el volumen que utilizan los agricultores; de esta forma, garantizar que futuros análisis contarán con series completas de datos, reduciendo la incertidumbre de las predicciones del actual 25% a menos del 10%, permitiendo modelar con precisión la relación real entre disponibilidad hídrica y producción agrícola.
- Dar mayor relevancia al Nexo WEF, sustentado en el hallazgo de que, al controlar por variables de confusión mediante modelos causales, la energía explica solo el 25% de la variación en producción, esta estimación podría estar influenciada por variables omitidas que estarían afectando tanto a la energía como a la producción, entre las que podrían estar políticas agrícolas o cambios en el tipo de cultivo. Esto destaca la necesidad de políticas integradas que consideren múltiples factores simultáneamente que pueden contribuir a la toma de decisiones orientadas a la sostenibilidad.
- Incentivar tecnologías que promuevan el uso de energía renovable que asegure un suministro constante de energía, esto como respuesta al escenario de mayor probabilidad identificado con el modelo bayesiano, donde el crecimiento moderado de la producción solo es posible si se mantiene el acceso energético actual. Las energías renovables emergen como factor estabilizador ante posibles crisis en suministro de combustibles.
- Implementar dispositivos IoT que permitan monitorear de manera permanente y en tiempo real los parámetros clave del nexo, de esta forma, reducir la banda de incertidumbre identificada de ± 20.000 kg en un 70%, permitiendo ajustes dinámicos en políticas agrícolas.
- En base al análisis de estabilidad que muestra confianza media alta de 78%, contar con bases de datos robustas, desagregadas por sectores, mejoraría la calidad de datos, aumentando la confianza por encima del 88%, haciendo las predicciones aptas para planificación estratégica en lugar de solo alertas tempranas.
- Se sugiere efectuar evaluaciones de validación a futuro empleando datos generados mediante escenarios hipotéticos simulados. Para garantizar la reproducibilidad del modelo, puede utilizarse el código disponible en el enlace proporcionado: https://github.com/RouseCC/ModelosML_Galapagos.git,

CONCLUSIONES

La aplicación del mapa de calor mostró una correlación alta de 0,89 entre energía per cápita y producción agrícola, bajo lo cual se continuó con el análisis. Sin embargo, al aplicar el modelo XGBoost, para realizar predicciones hacia adelante, este no reproduce la misma fuerza en esta relación debido a la mínima variabilidad en los datos de la energía per cápita. Esto demuestra que, aunque dos variables estén fuertemente relacionadas, dicha asociación no debe interpretarse como

una relación causal, mientras que la correlación mide la relación en los datos observados, el modelo XGBoost necesita que las variables proyectadas también varíen lo suficiente para capturar esa relación y amplificarla.

El modelo planteado nos deja ver una proyección plana de la producción agrícola, lo cual nos muestra que la energía per cápita, en los niveles proyectados, no genera un impacto sustancial en la producción en los próximos 10 años. El análisis predictivo de XGBoost muestra que bajo escenarios de poca variabilidad futura en energía y clima, esa relación pierde fuerza.

El diagnóstico de calidad reveló que nuestras interpolaciones introducen hasta 25% de incertidumbre. Por ello, la Recomendación 1 (registros anuales de fuentes hídricas) es crucial: garantiza que en 3 años tendremos series completas, reduciendo el error a menos del 10% y permitiendo modelar con precisión las relaciones hídrico-productivas.

Clarificación de Relaciones Causales: Los grafos causales demostraron que solo el 25% de la correlación energía-producción es causal real. Por esto, la Recomendación 2 (enfoque integral del nexo) es esencial dado que políticas que solo optimicen energía ignorarán el 75% de factores determinantes, llevando a intervenciones subóptimas.

Escenarios Realistas de Futuro: El modelado bayesiano identificó tres futuros plausibles, siendo el más probable (65%) un crecimiento moderado condicionado a acceso energético estable. Por tanto, la Recomendación 3 (energías renovables) asegura el escenario favorable y mitiga riesgos de los escenarios desfavorables.

Toma de Decisiones con Incertidumbre Cuantificada: Nuestras proyecciones incluyen ahora bandas de incertidumbre ($\pm 20,000$ kg). Por esto, la Recomendación 4 (tecnología IoT) es prioritaria: reduce esta incertidumbre en 70%, transformando predicciones de alertas generales a herramientas de gestión precisa.

Cimientos para Futuras Investigaciones: La validación rigurosa muestra confianza media-alta de (78%) en predicciones. Por ello, la Recomendación 5 (bases de datos robustas) es fundamental: eleva la confianza sobre 85%, haciendo posible usar estos modelos para asignación presupuestaria y no solo para orientación general.

Cada limitación técnica identificada tiene una respuesta concreta en nuestras recomendaciones. La implementación de este paquete de medidas transformará el modelado del nexo WEF en Galápagos pasar de ser un análisis teórico a convertirse en un instrumento sólido que respalde decisiones orientadas a la sostenibilidad"

De manera general, se evidencia que el nexo constituye una herramienta analítica útil para comprender las interdependencias entre recursos que son considerados críticos en Galápagos y su impacto en la producción agrícola local. Sin embargo, la escasez de datos sectorizados y la necesidad de interpolaciones resaltan la urgencia de fortalecer los sistemas de monitoreo e información de la región.

REFERENCIAS

- Bagal, Y. V., Pandey, A. R., Pednekar, S. V., & Dhamdhere, T. B. (2020). *Data mining in agriculture—A novel approach*. International Journal of Engineering Research & Technology (IJERT), 9(8). <https://doi.org/10.17577/ijertv9is080107>
- Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Carrión-Mero, P., Morante-Carballo, F., Briones-Bitar, J., Jaya-Montalvo, M., Sánchez-Zambrano, E., Solórzano, J., Malavé-Hernández, J., Montalván Toala, F. J., Proaño, J., Flor-Pineda, Á., & Espinel, R. (2024). *Water quality from natural sources for sustainable agricultural development strategies: Galápagos, Ecuador*. Water, 16(11), 1516. <https://doi.org/10.3390/w16111516>
- Chai, T., & Draxler, R. R. (2014). *Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature*. Geoscientific Model Development, 7(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Comisión Económica para América Latina y el Caribe (CEPAL). (2016, 1 de julio). *Horizontes 2030: La igualdad en el centro del desarrollo sostenible* (LC/G.2660(SES.36/3) Rev. 1). CEPAL.
- Cover, T., & Hart, P. (1967). *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13(1), 21–27. <https://doi.org/10.1109/TIT.1967.1053964>
- Danubianu, M. (2014). *Step by step data preprocessing for data mining: A case study*. In *Proceedings of the International Conference on Information Technologies (InfoTech-2015)* (pp. 17–18). Romania: “Stefan cel Mare” University of Suceava.
- Ding, T., Chen, J., Fang, L., Ji, J., & Fang, Z. (2023). *Urban ecosystem services supply-demand assessment from the perspective of the water-energy-food nexus*. Sustainable Cities and Society, 90, 104401. <https://doi.org/10.1016/j.scs.2023.104401>
- Douglas, C. H. (2006). *Small island states and territories: sustainable development issues and strategies—Challenges for changing islands in a changing world*. Sustainable Development, 14(2), 75–80. <https://doi.org/10.1002/sd.297>
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). *Support vector regression machines*. Advances in Neural Information Processing Systems, 9, 155–161.
- Echeverría, P., Domínguez, Ch., Villacís, M., & Violette, S. (2024). *Fog harvesting potential for domestic rural use and irrigation in San Cristóbal Island, Galápagos, Ecuador*. Cuadernos de Investigación Geográfica. <https://doi.org/10.18172/cig.4382>.

- Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Gil, M., & Llavona, A. (2020, 9 de octubre). *Implementación de políticas con enfoque nexo en América Latina y el Caribe (ALC): Indicadores y políticas basadas en evidencia para el caso de Bolivia* [Presentación]. CEPAL, División de Recursos Naturales.
- Gobierno de Galápagos. (2020, junio). *Información base para el diálogo: Plan Galápagos 2030* [Documento PDF]. https://www.gobiernogalapagos.gob.ec/wp-content/uploads/downloads/2020/06/Informacion_base_para_dialogo_plan_galapagos_2030.pdf
- Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). Prentice Hall.
- International Partnership for the Satoyama Initiative (IPSI). (2025). *Farm to table, leakages, and challenges for food security in Galapagos Islands*. Satoyama Initiative Case Study Series. https://satoyamainitiative.org/case_studies/farm-to-table-leakages-and-challenges-for-food-security-in-galapagos-islands/
- Jaramillo Díaz, C. A., Mena, C. F., Burneo, J. I., & Arias, M. (2022). *Adoption of sustainable agriculture practices through participatory research: A case study on Galápagos Islands farmers using water-saving technologies*. Plants, 11(21), 2848. <https://doi.org/10.3390/plants11212848>
- Llerena Pizarro, O. R., Micena, R. P., Tuna, C. E., & Silveira, J. L. (2019). *Electricity sector in the Galapagos Islands: Current status, renewable sources, and hybrid power generation system proposal*. Renewable and Sustainable Energy Reviews, 108, 65–75. <https://doi.org/10.1016/j.rser.2019.03.043>
- Naciones Unidas. (2018). *La Agenda 2030 y los Objetivos de Desarrollo Sostenible: Una oportunidad para América Latina y el Caribe* (LC/G.2681-P/Rev.3). Naciones Unidas.
- Paltán, H., Estrella, C., Córdova, M., Mena, C. F., Benítez, F., & Buytaert, W. (2023). *Water security and agricultural systems in the Galápagos Islands: Vulnerabilities under uncertain future climate and land use pathways*. Frontiers in Water, 5, 1170857. <https://doi.org/10.3389/frwa.2023.1170857>
- Sampedro, C., Pizzitutti, F., Quiroga, D., Walsh, S. J., & Mena, C. F. (2020). *Food supply system dynamics in the Galapagos Islands: Agriculture, livestock and imports*. Renewable Agriculture and Food Systems, 35(3), 234–248. <https://doi.org/10.1017/S1742170518000534>
- Sánchez-Zarco, X. G., González-Bravo, R., & Ponce-Ortega, J. M. (2021). *Multi-objective optimization approach to meet water, energy, and food needs in an arid region involving security assessment*. ACS Sustainable Chemistry & Engineering, 9(13), 4754–4767. <https://doi.org/10.1021/acssuschemeng.0c08305>

- Seber, G. A. F., & Lee, A. J. (2012). *Linear regression analysis* (2nd ed.). Wiley.
<https://doi.org/10.1002/9781118274422>
- Secretaría Nacional de Planificación. (2021, 13 de julio). *Ecuador presentó su experiencia y desafíos en la implementación de la Agenda 2030 en foro de la ONU* [Noticia].
<https://www.planificacion.gob.ec/ecuador-presento-su-experiencia-y-desafios-en-la-implementacion-de-la-agenda-2030-en-foro-de-la-onu/>
- Shen, F., Zhao, L., Du, W., Zhong, W., Peng, X., & Qian, F. (2022). *Data-driven stochastic robust optimization for industrial energy system considering renewable energy penetration*. ACS Sustainable Chemistry & Engineering, 10, 3690–3703.
<https://doi.org/10.1021/acssuschemeng.2c00211>
- Simpson, G. B., & Jewitt, G. P. W. (2019). *The development of the water energy food nexus as a framework for achieving resource security: A review*. Frontiers in Environmental Science, 7, Article 8. <https://doi.org/10.3389/fenvs.2019.00008>
- Theil, H. (1971). *Principles of econometrics*. John Wiley & Sons.
- Weitz, N., Nilsson, M., & Davis, M. (2014). *A nexus approach to the post-2015 agenda: Formulating integrated water, energy, and food SDGs*. SAIS Review of International Affairs, 34(2), 37–50. <https://doi.org/10.1353/sais.2014.0022>
- Willmott, C. J., & Matsuura, K. (2005). *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance*. Climate Research, 30(1), 79–82. <https://doi.org/10.3354/cr030079>