



Maestría en

INTELIGENCIA ARTIFICIAL APLICADA

**Trabajo previo a la obtención de título de
Magister en Inteligencia Artificial
Aplicada**

AUTORES:

Edwin Alexander Orbe Cisneros

Edison Javier Usca Paguay

Byron Geovanny Herrera Mayorga

TUTOR:

Pablo Xavier Andrade Montalvo

Fernanda Paulina Vizcaíno Imacaña

TEMA: Detección de Anomalías Energéticas en Hogares Ecuatorianos mediante
Random Forest: Un Enfoque Híbrido con Datos Sintéticos y Referencias
Internacionales

Certificación de autoría

Nosotros, **Edwin Alexander Orbe Cisneros, Edison Javier Usca Paguay, Byron Geovanny Herrera Mayorga**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
**EDWIN ALEXANDER
ORBE CISNEROS**

Firma

Edwin Alexander Orbe Cisneros



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
**EDISON JAVIER USCA
PAGUAY**

Firma

Edison Javier Usca Paguay



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
**BYRON GEOVANNY
HERRERA MAYORGA**

Firma

Byron Geovanny Herrera Mayorga

Autorización de Derechos de Propiedad Intelectual

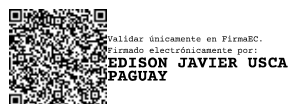
Nosotros, **Edwin Alexander Orbe Cisneros, Edison Javier Usca Paguay, Byron Geovanny Herrera Mayorga**, en calidad de autores del trabajo de investigación titulado ***Detección de Anomalías Energéticas en Hogares Ecuatorianos mediante Random Forest: Un Enfoque Híbrido con Datos Sintéticos y Referencias Internacionales***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, agosto 2026



Firma

Edwin Alexander Orbe Cisneros



Firma

Edison Javier Usca Paguay

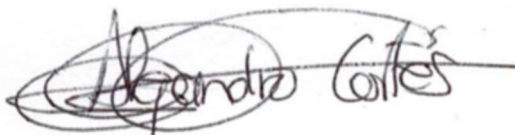


Firma

Byron Geovanny Herrera Mayorga

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López** Director EIG y **Karla Estefanía Mora Cajas** Coordinadora UIDE, declaramos que: **Edwin Alexander Orbe Cisneros, Edison Javier Usca** **Paguay, Byron Geovanny Herrera Mayorga** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López Director de la
Maestría en Inteligencia Artificial
Aplicada



Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Inteligencia Artificial Aplicada

DEDICATORIA

Como creyente en Dios Todopoderoso, dedico en primer lugar este logro a Él, con profunda gratitud por sus infinitas bendiciones, por guiar cada uno de mis pasos y darme la fortaleza, la sabiduría y la perseverancia necesarias para alcanzar esta meta. Sin su amor y su dirección, este camino no habría sido posible.

A mis padres, Edwin y Magola, por su amor incondicional, sus enseñanzas, sus sacrificios y los valores con los que me formaron. Gracias por creer siempre en mí y por convertirme en la persona que soy hoy.

A mi amada hija, Keily Anahí, quien es la mayor inspiración de mi vida, el motor que impulsa cada uno de mis esfuerzos y la fuerza que me motiva a levantarme cada día para seguir adelante y luchar por un mejor futuro.

Y a mi querida abuelita, que hoy descansa en el cielo, por el inmenso cariño, la paciencia y los cuidados que me brindó desde mi infancia. Aunque ya no esté físicamente conmigo, su amor, sus enseñanzas y su recuerdo permanecen para siempre en mi corazón y me acompañan en cada uno de mis logros.

Edwin Alexander Orbe Cisneros

DEDICATORIA

Dedico este trabajo, con todo mi corazón, a Dios, por haberme dado la fuerza para continuar cuando el camino parecía difícil y por permitirme llegar hasta este momento tan importante de mi vida.

A mi esposa, mi compañera de vida, quiero dedicarte un agradecimiento muy especial. Gracias por tu amor, por tu paciencia, por comprender mis momentos de ausencia y por acompañarme silenciosamente en cada esfuerzo. Gracias por creer en mí, por sostenerme cuando las fuerzas faltaban y por recordarme que todo sacrificio tendría su recompensa. Este logro también lleva una parte de ti, porque detrás de cada página y de cada momento de esfuerzo estuvo tu apoyo y tu amor.

A mis hijas, que son una de mis mayores razones para seguir adelante. Ustedes son mi inspiración y la motivación que me impulsa a querer ser cada día una mejor persona. Espero que algún día puedan mirar este logro y entender que nunca debemos dejar de luchar por nuestros sueños, incluso cuando el camino se vuelve difícil. Todo esfuerzo tiene sentido cuando se hace pensando en las personas que amamos.

A mis padres, por haberme enseñado siempre el valor del esfuerzo, la responsabilidad y la perseverancia. Este logro también es una forma de agradecerles todo lo que han hecho por mí.

Y finalmente, me dedico estas palabras a mí mismo. A la persona que comenzó este sueño con ilusión y que hoy puede mirar atrás y sentirse orgulloso de no haberse rendido. Esta tesis representa mucho más que un requisito académico: representa noches de esfuerzo, momentos de incertidumbre, aprendizajes, sacrificios y, sobre todo, la satisfacción de haber llegado hasta aquí.

Hoy cierro una etapa, pero también abro las puertas a nuevos sueños. Y cuando mire estas páginas en el futuro, quiero recordar que todo esfuerzo tuvo un propósito y que, aun en los momentos más difíciles, siempre valió la pena seguir.

Este logro es para todos los que creyeron en mí y, sobre todo, para quienes caminaron a mi lado sin dejarme caer.

Byron Geovanny Herrera Mayorga

DEDICATORIA

Dedico el presente trabajo a Dios en primer lugar y a mi familia: padres, hermanos e hijos y a mi esposa, ellos son el pilar fundamental de mi carrera profesional y mayor motivación a seguir adelante.

Edison Usca Paguay.

AGRADECIMIENTOS

*Expreso, en primer lugar, mi más profundo agradecimiento a **Dios Todopoderoso**, por su infinita misericordia, por acompañarme en cada etapa de mi vida y por brindarme la fortaleza, la sabiduría y la esperanza necesarias para superar los desafíos y alcanzar esta importante meta académica. Su presencia constante ha sido el fundamento de cada uno de mis logros.*

*A **mis padres**, por su amor incondicional, su ejemplo de esfuerzo y dedicación, y por estar a mi lado desde el inicio de mi vida hasta el día de hoy. Gracias por creer siempre en mí, por su apoyo moral y emocional, y por el respaldo económico que, cuando fue necesario, me permitió continuar con mi formación profesional.*

*A la **Universidad Internacional del Ecuador (UIDE)**, por abrirme sus puertas y brindarme un espacio de excelencia académica que fortaleció mis conocimientos, promovió mi crecimiento profesional y enriqueció mi formación integral.*

*A todos los **docentes de la Maestría en Inteligencia Artificial Aplicada de la UIDE**, por compartir generosamente sus conocimientos, experiencias y valores. Su dedicación, compromiso y excelencia académica fueron fundamentales para mi aprendizaje y para la culminación exitosa de este proceso de formación.*

Finalmente, expreso mi sincero agradecimiento a todas las personas que, de una u otra manera, contribuyeron con su apoyo, motivación y confianza durante este camino. Cada palabra de aliento y cada gesto de solidaridad fueron importantes para alcanzar este logro.

Edwin Alexander Orbe Cisneros

AGRADECIMIENTOS

En primer lugar, agradezco a Dios por haberme acompañado y fortalecido durante todo este camino. Gracias por darme la sabiduría, la paciencia y la perseverancia necesarias para superar cada desafío y permitirme alcanzar una meta que en algún momento parecía tan lejana.

A mi esposa, por ser mi compañera incondicional en este proceso. Gracias por tu amor, tu paciencia, tu comprensión y por estar a mi lado incluso en los momentos de mayor cansancio y dificultad. Gracias por entender mis sacrificios, por animarme cuando las fuerzas comenzaban a faltar y por creer en mí aun cuando yo mismo llegué a dudar. Este logro también es tuyo.

A mis hijas, por ser mi mayor inspiración y la razón que muchas veces me impulsó a seguir adelante. Cada sonrisa, cada abrazo y cada momento compartido me recordaron por qué valía la pena continuar. Espero que este logro les enseñe que con esfuerzo, perseverancia y dedicación podemos alcanzar aquello que nos proponemos.

A mis docentes y tutores, por compartir sus conocimientos, experiencias y enseñanzas, y por contribuir significativamente a mi formación académica y profesional. Cada enseñanza recibida ha sido parte importante de este camino.

Finalmente, agradezco a todas aquellas personas que, aunque quizás no aparezcan mencionadas aquí, aportaron de alguna manera para que este sueño pudiera convertirse en realidad. Cada palabra de apoyo, cada consejo y cada gesto de confianza fueron importantes.

A todos ustedes, gracias de corazón.

Byron Geovanny Herrera Mayorga

AGRADECIMIENTOS

Al Dios Todopoderoso quien me permite poder realizar cada logro profesional en mi carrera, a mi familia en especial a mi esposa e hijos, con su apoyo nada de esto sería posible, al Instituto Tecnológico Universitario Euroamericano quien contribuyó con este paso importante con su apoyo en todo momento y en cada fase de la maestría. A la Armada del Ecuador quien me acogió de muy joven y donde pude iniciar y desenvolverme de forma profesional, conocer a personas motivadoras y ejemplares con palabras de aliento como la frase “Eres el mejor de todos, Eres el mejor cabo que conozco y Donde quiera que esté le irá bien porque es un excelente profesional” y a la Universidad Internacional del Ecuador, quien organizó y sostuvo una estructura académica de gran nivel con profesionales capaces y a la vanguardia tecnológica necesaria para esta Maestría.

Edison Usca Paguay.

RESUMEN

El proyecto consiste en el desarrollo de un sistema inteligente para la detección de anomalías en el consumo de energía eléctrica en hogares ecuatorianos mediante técnicas de Aprendizaje Automático. El objetivo es identificar patrones de consumo que se desvían del comportamiento esperado de una vivienda, proporcionando una herramienta de apoyo para la detección temprana de posibles ineficiencias energéticas, errores de medición, consumos atípicos o situaciones que puedan requerir una evaluación técnica.

El sistema ha sido desarrollado en el lenguaje de programación Python y una arquitectura modular que integra diferentes componentes para el procesamiento de datos, entrenamiento del modelo y visualización de resultados. La interfaz de usuario fue implementada mediante Gradio, permitiendo que cualquier usuario pueda ingresar información relacionada con las características de una vivienda y obtener una evaluación automática sobre la probabilidad de presentar un comportamiento energético anómalo.

El modelo de aprendizaje automático utiliza variables representativas del hogar, como ubicación geográfica (costa, sierra, oriente), número de habitantes, cantidad de electrodomésticos, características del consumo y otros factores que influyen en la demanda energética. A partir de estas variables, el algoritmo genera una clasificación que determina si el comportamiento del consumo corresponde a un patrón normal o potencialmente anómalo.

Como parte de la metodología implementada, el sistema incorpora un mecanismo de retroalimentación continua, mediante el cual los nuevos registros generados por los usuarios pueden almacenarse y utilizarse posteriormente para actualizar el conjunto de datos y reentrenar el modelo predictivo. Esta característica permite mejorar progresivamente el desempeño del sistema conforme aumenta la disponibilidad de información real, favoreciendo su capacidad de adaptación a distintos perfiles de consumo presentes en los hogares ecuatorianos.

Los resultados evidenciaron un desempeño satisfactorio del modelo propuesto, alcanzando una precisión superior al 91 % en la detección de anomalías. La evaluación mediante validación cruzada, matriz de confusión, curvas ROC y métricas de desempeño (accuracy, precision, recall y F1-score) confirmó la capacidad del modelo para identificar patrones de consumo anómalos con un alto nivel de confiabilidad. Adicionalmente, se desarrolló un prototipo web implementado con Gradio y FastAPI que permite estimar el consumo mensual esperado, calcular el costo aproximado de la energía conforme a la normativa tarifaria ecuatoriana y comparar dicho valor con el consumo real reportado por el usuario para identificar posibles anomalías.

Palabras Clave: Aprendizaje automático, detección de anomalías, consumo energético residencial, hogares ecuatorianos.

ABSTRACT

The project involves the development of an intelligent system for detecting anomalies in electricity consumption in Ecuadorian households using machine learning techniques. The goal is to identify consumption patterns that deviate from the expected behavior of a household, providing a tool to support the early detection of potential energy inefficiencies, measurement errors, atypical consumption, or situations that may require a technical evaluation.

The system was developed using the Python programming language and a modular architecture that integrates various components for data processing, model training, and result visualization. The user interface was implemented using Gradio, allowing any user to enter information related to a household's characteristics and obtain an automatic assessment of the probability of anomalous energy behavior.

The machine learning model uses variables representative of the household, such as geographic location (coastal, mountainous, east), number of residents, number of appliances, consumption patterns, and other factors that influence energy demand. Based on these variables, the algorithm generates a classification that determines whether the consumption pattern corresponds to a normal or potentially anomalous pattern.

As part of the implemented methodology, the system incorporates a continuous feedback mechanism through which new records generated by users can be stored and later used to update the dataset and retrain the predictive model. This feature allows for the progressive improvement of the system's performance as the availability of real-world information increases, enhancing its ability to adapt to the various consumption profiles found in Ecuadorian households.

From a technological standpoint, the solution integrates specialized libraries for data science and machine learning—such as Pandas, NumPy, and Scikit-Learn—along with tools for model serialization, data management, and application deployment. The architecture was designed in a modular fashion to facilitate maintenance, scalability, and the incorporation of new algorithms or features in future versions.

Keywords: Machine learning, anomaly detection, residential energy consumption, Ecuadorian households.

TABLA DE CONTENIDOS

CAPITULO 1:.....	1
1. INTRODUCCIÓN.....	1
1.1 Definición del proyecto.....	1
1.2 Justificación e importancia del trabajo de investigación	4
1.3 Alcance	6
1.4 Objetivos	8
1.4.1 Objetivo general.....	8
1.4.2 Objetivos específicos	9
CAPITULO 2:.....	10
2. REVISIÓN DE LITERATURA.....	10
2.1. Estado del Arte	10
2.1.1. Detección de anomalías en consumo energético residencial	10
2.1.2. Técnicas de aprendizaje automático aplicadas a la detección de anomalías ...	11
2.1.3. Datasets y características de los datos utilizados	14
2.1.4. Comparación de los enfoques empleados en investigaciones previas	15
2.1.5. Análisis Crítico de investigaciones previas	17
2.1.6. Vacíos de investigación identificados	18
2.1.7. Justificación de la propuesta desarrollada.....	19
2.2. Marco Teórico	20
2.2.1. Inteligencia artificial y aprendizaje automático	20
2.2.2. Detección de anomalías	21
2.2.3. Set de datos o datasets.....	21

2.2.4.	Integración técnica de los datasets (Data Fusion).....	24
2.2.5.	Set de Datos Sintético	25
2.2.6.	Metodología de construcción del set de datos sintético.....	27
2.2.6.1.	Definición de los rangos de las variables.....	28
2.2.6.2.	Generación del consumo esperado.....	29
2.2.6.3.	Generación del consumo reportado	29
2.2.6.4.	Etiquetado de anomalías	30
2.2.6.5.	Validación del conjunto de datos	30
2.2.7.	Random Forest: Fundamentos y funcionamiento	31
2.2.7.1.	Fundamentos del algoritmo.....	31
2.2.7.2.	Funcionamiento del modelo.....	32
2.2.7.3.	Ventajas del Random Forest.....	33
2.2.7.4.	Aplicación en la detección de anomalías	35
2.2.7.5.	Justificación del uso de Random Forest.....	35
2.2.8.	Métricas de evaluación en el aprendizaje automático.....	37
2.2.8.1.	Exactitud (Accuracy)	37
2.2.8.2.	Precisión.....	38
2.2.8.3.	Sensibilidad o Recall.....	38
2.2.8.4.	F1-Score	38
2.2.8.5.	AUC-ROC.....	39
2.2.8.6.	Matriz de Confusión	40
2.2.9.	Restricciones y consideraciones metodológicas	42
CAPÍTULO 3:.....		43
3: DESARROLLO		43
3.1.	Arquitectura del Sistema	43
3.1.1.	Componentes del Sistema.....	43

3.1.2. Diagrama de componentes UML.....	45
3.1.3. Diagrama de casos de uso UML.....	48
3.2. Desarrollo del set de datos sintético.....	49
3.2.1. Metodología de Construcción del Set de Datos Sintético	50
3.2.2. Integración de fuentes de información	50
Dataset Household Power Consumption	50
Dataset UK-DALE	51
3.2.3. Generación de los registros sintéticos	51
3.2.4. Cálculo de Consumo Esperado.....	52
3.2.5. Generación de Anomalías.....	52
3.3. Sistema de Tarifas Eléctricas.....	53
3.3.1. Estructura Tarifaria	54
3.3.2. Beneficios y Descuentos.....	54
3.4. Entrenamiento del Modelo Random Forest	55
3.4.1. Preparación de Datos	55
3.4.2. Configuración del Modelo.....	55
3.4.3. Análisis de Importancia de Features	56
3.5. Aplicación Web con FastAPI y Gradio	56
3.5.1. Arquitectura de la Aplicación	56
3.5.2. Funcionalidades Principales	56
3.5.3. Diseño de Interfaz de Usuario	57
CAPÍTULO 4:.....	63
ANÁLISIS DE RESULTADOS.....	63
4.1. Procesamiento de los Datasets de Referencia	63
4.1.1. Household Power Consumption	63
4.1.2. UK-DALE y valores de referencia implementados.....	63

4.2. Construcción y Validación del Dataset Integrado	64
4.2.2. Validación de Coherencia	65
4.3. Diseño Experimental	67
4.3.1. Protocolo de Evaluación	69
4.3.2. Formulaciones del problema	69
4.4. Análisis del Desempeño del Modelo	70
4.4.1. Análisis de comparativo de algoritmos	71
4.4.2. Matriz de Confusión	73
4.4.3. Curvas ROC y precisión- recall	74
4.4.4. Significancia estadística de las diferencias	76
4.4.5. Calibración del umbral de decisión	78
4.4.6. Curva de aprendizaje y mejora continua	80
4.4.7. Análisis de sobreajuste y límites de la generalización	82
4.5. Análisis de Importancia de Variables	85
4.5.1. Ranking de variables predictoras	86
4.5.2. Interpretación y validación cruzada entre modelos	86
4.6. Evaluación del Sistema Tarifario	87
4.6.1. Costos por tipo de usuario y pruebas de frontera	87
4.6.2. Análisis de sensibilidad de los beneficios	87
4.7. Verificación y Validación de la Aplicación	88
4.7.1. Casos de prueba funcionales	89
4.7.2. Pruebas de robustez y rendimiento	90
4.7.3. Prueba de extremo a extremo en producción	90
4.7.4. Validación de las recomendaciones	91
4.8. Discusión y Comparación con Estudios Previos	92
4.9. Limitaciones Identificadas	92

4.10. Síntesis de Resultados	93
CAPITULO 5:.....	95
CONCLUSIONES Y COMENDACIONES	95
5.1 CONCLUSIONES	95
5.1.1. Verificación del cumplimiento de los objetivos.....	95
5.1.2. Conclusiones generales.....	97
5.2. El sistema en perspectiva	98
5.2.1. Línea base del prototipo	98
5.2.2. Estado validado del sistema.....	99
5.2.3. Futuro: hoja de ruta.....	100
5.3. Recomendaciones.....	101
Referencias Bibliográficas	102

LISTA DE TABLAS (Índice de tablas)

TABLA 1	15
COMPARACIÓN DE ENFOQUES PARA LA DETECCIÓN DE ANOMALÍAS EN EL CONSUMO ENERGÉTICO	15
TABLA 2	26
DESCRIPCIÓN DE LAS VARIABLES DEL SET DE DATOS UTILIZADO	26
TABLA 3	34
COMPARATIVA ENTRE RANDOM FOREST Y XG BOOST	34
TABLA 4	54
FUNCIONES DEL MÓDULO TARIFAS.PY	54
TABLA 5	64
CONSUMOS Y POTENCIAS DE REFERENCIA POR EQUIPO, IMPLEMENTADOS EN EL SISTEMA	64
TABLA 6	64
ESTADÍSTICAS DESCRIPTIVAS DEL DATASET INTEGRADO (VERSIÓN CONGELADA 1.0).....	64
TABLA 7	71
VALIDACIÓN CRUZADA ESTRATIFICADA 5-FOLD - EXPERIMENTO A: CONSUMO ALTO (MEDIA $\pm$ DESVIACIÓN ESTÁNDAR)	71
TABLA 8	71
VALIDACIÓN CRUZADA ESTRATIFICADA 5-FOLD — EXPERIMENTO B: DETECCIÓN DE ANOMALÍA (MEDIA $\pm$ DESVIACIÓN ESTÁNDAR)	71
TABLA 9	76
PRUEBA EXACTA DE MCNEMAR: RANDOM FOREST FRENTE A CADA COMPETIDOR — EXPERIMENTO A	76
TABLA 10	77
PRUEBA EXACTA DE MCNEMAR: RANDOM FOREST FRENTE A CADA COMPETIDOR — EXPERIMENTO B.....	77
TABLA 11	87
COSTO CALCULADO POR EL SISTEMA PARA 150 KWH/MES (REGIÓN SIERRA)	87
TABLA 12	89
CASOS DE PRUEBA FUNCIONALES: SEIS ESCENARIOS SOBRE TRES PERFILES REPRESENTATIVOS	89
TABLA 13	90

PRUEBA DE EXTREMO A EXTREMO CONTRA EL DESPLIEGUE EN PRODUCCIÓN	90
TABLA 14	91
POTENCIAL DE AHORRO ACUMULADO DE LAS RECOMENDACIONES GENERADAS, POR PERFIL	91
TABLA 15	95
CUMPLIMIENTO DE LOS OBJETIVOS ESPECÍFICOS Y EVIDENCIA QUE LO SUSTENTA	95
TABLA 16	99
DEFECTOS DE LA LÍNEA BASE DETECTADOS POR LA EVALUACIÓN SISTEMÁTICA Y SU RESOLUCIÓN	99

LISTA DE FIGURAS (Índice de figuras)

FIGURA 1	20
INTELIGENCIA ARTIFICIAL Y SUS RAMAS	20
FIGURA 2	22
EJEMPLO DE UN SET DE DATOS CON FILAS Y COLUMNAS	22
FIGURA 3	28
CLASIFICACIÓN DE LAS VARIABLES DEL SET DE DATOS SINTÉTICO	28
FIGURA 4	32
ALGORITMO RANDOM FOREST	32
FIGURA 5	33
EVOLUCIÓN DE LOS ALGORITMOS BASADOS EN ÁRBOLES DE DECISIÓN	33
FIGURA 6	40
CURVA AUC ROC	40
FIGURA 7	41
MATRIZ DE CONFUSIÓN	41
FIGURA 8	44
ARQUITECTURA GENERAL DEL SISTEMA	44
FIGURA 9	47
DIAGRAMA DE COMPONENTES DEL SISTEMA	47
FIGURA 10	48
DIAGRAMA DE FLUJO DEL SISTEMA.....	48
FIGURA 11	49
DIAGRAMA DE CASO DE USO UML.....	49
FIGURA 12	50
METODOLOGÍA DE CONSTRUCCIÓN DE DATOS SINTÉTICOS	50
FIGURA 13	53
FLUJO DEL MÓDULO TARIFAS.PY	53
FIGURA 14	58

INTERFAZ DE USUARIO – PESTAÑA PRINCIPAL “DETECCIÓN” DEL SISTEMA DETECTOR DE ANOMALÍAS ENERGÉTICAS	58
FIGURA 15	59
INTERFAZ DE USUARIO – PESTAÑA “HISTORIAL” DEL SISTEMA DETECTOR DE ANOMALÍAS ENERGÉTICAS	59
FIGURA 16	60
INTERFAZ DE USUARIO – PESTAÑA “MÉTRICAS DEL MODELO” DEL SISTEMA DETECTOR DE ANOMALÍAS ENERGÉTICAS	60
FIGURA 17	61
INTERFAZ DE USUARIO – PESTAÑA “COMPARATIVA MODELOS” DEL SISTEMA DETECTOR DE ANOMALÍAS ENERGÉTICAS	61
FIGURA 18	62
INTERFAZ DE USUARIO – PESTAÑA “ACERCA DE” DEL SISTEMA DETECTOR DE ANOMALÍAS ENERGÉTICAS	62
FIGURA 19	65
DISTRIBUCIÓN DEL CONSUMO REPORTADO Y COMPARACIÓN POR REGIÓN	65
FIGURA 20	66
COHERENCIA ENTRE LA FÓRMULA DE LA APLICACIÓN Y EL DATASET SINTÉTICO (N = 1 500; R = 0.877)	66
FIGURA 21	67
CONSUMO ESPERADO FRENTE A CONSUMO REPORTADO SEGÚN LA ETIQUETA DE ANOMALÍA	67
FIGURA 22	73
COMPARATIVA VISUAL DE LOS OCHO MODELOS EN LAS DOS FORMULACIONES DEL PROBLEMA (F1 Y AUC-ROC, VALIDACIÓN CRUZADA 5-FOLD)	73
FIGURA 23	74
MATRICES DE CONFUSIÓN DE LOS OCHO MODELOS — EXPERIMENTO B (CONJUNTO DE PRUEBA, N = 1 503) ..	74
FIGURA 24	75
CURVAS ROC DE LOS OCHO MODELOS — EXPERIMENTO A (HOLDOUT 70/30)	75
FIGURA 25	75
CURVAS ROC DE LOS OCHO MODELOS — EXPERIMENTO B (HOLDOUT 70/30)	75
FIGURA 26	76
CURVAS DE PRECISIÓN-RECALL DE LOS OCHO MODELOS — EXPERIMENTO B (HOLDOUT 70/30)	76
PRUEBA EXACTA DE MCNEMAR: RANDOM FOREST FRENTE A CADA COMPETIDOR — EXPERIMENTO B	77
FIGURA 27	79
PRECISIÓN, RECALL Y F1 EN FUNCIÓN DEL UMBRAL DE DECISIÓN — RANDOM FOREST, EXPERIMENTO A	79
FIGURA 28	79
ANTES Y DESPUÉS DE LA CALIBRACIÓN DEL UMBRAL DE DECISIÓN (0.50 → 0.81)	79
FIGURA 29	81
CURVA DE APRENDIZAJE DEL MODELO RANDOM FOREST (VALIDACIÓN CRUZADA 5-FOLD)	81
FIGURA 30	82

CONTROL DE ACEPTACIÓN DEL REENTRENAMIENTO: COMPARACIÓN ENTRE EL MODELO EN PRODUCCIÓN Y EL
CANDIDATO RECHAZADO (PRUEBA EN SECO) 82

FIGURA 31 86

IMPORTANCIA DE VARIABLES: RANDOM FOREST FRENTE A XGBOOST — EXPERIMENTO A..... 86

FIGURA 32 88

COSTO MENSUAL SEGÚN TIPO DE USUARIO, CALCULADO CON EL MÓDULO TARIFARIO REAL (REGIÓN SIERRA)88

CAPITULO 1:

1. INTRODUCCIÓN

1.1 Definición del proyecto

El presente proyecto tiene como finalidad desarrollar un prototipo basado en inteligencia artificial orientado a la detección de anomalías en el consumo energético residencial de hogares ecuatorianos. La solución propuesta integra técnicas de aprendizaje automático para analizar el comportamiento del consumo eléctrico y proporcionar a los usuarios una herramienta de apoyo para identificar patrones de consumo potencialmente anómalos, además de promover un uso más eficiente de la energía.

El sistema utilizará un conjunto de datos sintéticos, generados a partir de patrones de consumo energético mediante el análisis de datasets internacionales de referencia, principalmente UK-DALE (2012–2017) (Kelly, y College, 2017) y Household Power Consumption (2006–2010) (Hebrail y Berard, 2006). Estos datasets servirán como base para modelar escenarios representativos de consumo residencial, adaptado al contexto del proyecto y orientado al entrenamiento de los modelos de aprendizaje automático que conforman el sistema de detección de anomalías y de predicción de consumo.

La unidad de análisis corresponde a un hogar residencial ecuatoriano y la información procesada considera el consumo mensual de energía eléctrica expresado en kilovatios-hora (kWh), así como su costo estimado en dólares estadounidenses, calculado conforme a la normativa tarifaria vigente en el Ecuador. La variable dependiente principal de la investigación corresponde a la variable **anomalía**, la cual clasifica el comportamiento del consumo como normal o potencialmente anómalo. Para apoyar este proceso, el sistema incorpora un modelo de regresión (Random Forest Regressor), cuya función es estimar el consumo energético esperado de la vivienda a partir de sus características demográficas y de equipamiento. Esta estimación constituye una referencia utilizada posteriormente por el modelo de clasificación basado en Random Forest para determinar la existencia de posibles anomalías en el consumo reportado.

El modelo de clasificación supervisada constituye el componente central de la investigación. Su finalidad es identificar patrones de consumo energético anómalos mediante el análisis conjunto del consumo esperado, el consumo reportado y las características del hogar. De esta manera, el modelo permite clasificar cada registro como consumo normal o consumo potencialmente anómalo, proporcionando un mecanismo inteligente de apoyo para el análisis del comportamiento energético residencial.

El consumo energético residencial representa en el Ecuador un componente significativo de la demanda total de electricidad en el Sistema Nacional Interconectado (SNI). Sin embargo, existen diversas problemáticas identificadas estadísticamente en el consumo habitual, tales como el uso ineficiente de energía, pérdidas económicas no técnicas (como conexiones ilegales o errores de medición), así como la falta de herramientas avanzadas confiables que permitan monitorear y analizar el comportamiento del consumo en tiempo real de manera inteligente. Estas situaciones no solo afectan la sostenibilidad del sistema eléctrico nacional, sino que también generan impactos económicos negativos para las empresas distribuidoras, así como a usuarios finales.

Actualmente, los métodos tradicionales de análisis del consumo eléctrico se basan en revisiones manuales o en reglas fijas, lo que limita la capacidad de detectar anomalías o cambios en el comportamiento del consumo; situaciones que pasan desapercibidas o son detectadas de manera tardía e incompleta, reduciendo la efectividad de las acciones correctivas que permitan prevenir futuras mediciones incorrectas y sobre todo que deriven en excesos en los tarifarios mensuales.

Este problema merece ser estudiado debido a su impacto directo en la eficiencia energética, la reducción de pérdidas económicas y la sostenibilidad del SNI. Además, en un contexto donde la cuarta revolución industrial 4.0 conlleva a una transformación digital y el uso de tecnologías inteligentes que están en crecimiento, resulta fundamental incorporar soluciones basadas en análisis de datos y aprendizaje automático que permitan optimizar la gestión del consumo energético.

Como solución, se propone el desarrollo de un sistema inteligente de detección de anomalías energéticas basado en aprendizaje automático. La arquitectura incorpora

dos modelos supervisados con funciones complementarias. El primero corresponde a un modelo de regresión encargado de estimar el consumo energético esperado de una vivienda a partir de sus características. El segundo, implementado mediante Random Forest, constituye el componente principal de la investigación y utiliza dicha estimación junto con el consumo reportado para clasificar el comportamiento energético como normal o potencialmente anómalo. Esta estrategia permite disponer de una referencia objetiva para apoyar la detección de desviaciones significativas en el consumo residencial.

El proyecto busca proporcionar una herramienta que permita a los usuarios:

- Detectar posibles anomalías en el consumo energético residencial mediante técnicas de aprendizaje automático.
- Estimar el consumo energético esperado de acuerdo con las características del hogar y la tarifa eléctrica vigente en el Ecuador.
- Comparar el consumo esperado con el consumo reportado para determinar el nivel de anomalía y generar alertas cuando se identifiquen desviaciones significativas.
- Emitir recomendaciones personalizadas orientadas a mejorar la eficiencia energética.
- Informar sobre los beneficios tarifarios disponibles conforme a la normativa ecuatoriana.
- Registrar las interacciones de los usuarios para ampliar progresivamente el conjunto de datos y facilitar el reentrenamiento periódico del modelo.
- Almacenar registro estadístico mensual a partir de la interacción del usuario con su número de cédula (Dato no obligatorio).

En consecuencia, la presente investigación aborda un problema de clasificación supervisada, en el cual la estimación del consumo esperado mediante un modelo de regresión constituye una etapa auxiliar que proporciona la referencia necesaria para que el modelo de clasificación identifique patrones de consumo energético potencialmente anómalos.

Como trabajo futuro, se contempla la incorporación de un chatbot interactivo que pueda responder preguntas frecuentes de los usuarios, explicar los resultados de las predicciones y ofrecer consejos adicionales de ahorro energético en lenguaje natural.

1.2 Justificación e importancia del trabajo de investigación

El análisis del consumo energético residencial constituye un tema de alta relevancia en la actualidad, especialmente en países en desarrollo como el Ecuador, donde la eficiencia en el uso de la energía eléctrica y la reducción de pérdidas representan desafíos clave para la sostenibilidad del sistema eléctrico. En este contexto, la detección de anomalías en el consumo energético se convierte en un problema crítico, debido a que permite identificar comportamientos atípicos asociados a fallas técnicas o usos ineficientes de la energía.

La relevancia de este problema radica en su impacto directo tanto a nivel económico como operativo. Las pérdidas en el consumo de energía eléctrica, especialmente las no técnicas (como fraudes o errores de medición), generan importantes afectaciones para las empresas distribuidoras y, en consecuencia, para el sistema eléctrico nacional. Además, el consumo ineficiente por parte de los usuarios residenciales incrementa la demanda energética, lo cual puede derivar en mayores costos de generación y en una presión adicional sobre la infraestructura existente. En estos últimos años, donde se han presentado problemas en el ámbito energético del Sistema Nacional Interconectado, es necesario cuidar de este recurso para contribuir y mejorar la salud del Sistema Eléctrico del Ecuador.

Resolver este problema mediante el uso de técnicas de inteligencia artificial puede tener un impacto significativo en tres niveles identificados. En primer lugar, permitiría mejorar la capacidad de detección temprana de anomalías, optimizando los procesos de supervisión y control del consumo eléctrico, realizando una comparativa basados en un modelo confiable con alta precisión. En segundo lugar, contribuiría a la reducción del consumo y demanda energética poblacional, mediante una mejor planificación del sistema eléctrico. Finalmente, promovería una cultura de uso eficiente de la energía, beneficiando a los usuarios finales mediante la obtención de información objetiva sobre su consumo real, proyecciones y estimaciones que optimicen su economía.

El proyecto beneficia a múltiples actores. Por un lado, a las empresas distribuidoras de energía eléctrica, que podrán contar con herramientas más avanzadas para el monitoreo del consumo mediante el presente estudio que busca ser el punto de partida en el análisis energético a nivel residencial de Ecuador Continental. Por otro lado, a los organismos de regulación y control, como ARCONEL, que podrían apoyarse en este tipo de soluciones para mejorar la supervisión del sector. Asimismo, los usuarios residenciales se verán beneficiados al poder identificar patrones de consumo ineficiente y tomar decisiones informadas para optimizar su uso de energía.

Desde el punto de vista académico, este proyecto aporta al desarrollo y aplicación de técnicas de inteligencia artificial en el análisis de datos históricos particularmente en el ámbito energético. Además, contribuye a la generación de conocimiento aplicado en un contexto local, lo cual es fundamental para fortalecer la investigación en el país. En el ámbito tecnológico, el proyecto impulsa la implementación de modelos avanzados de aprendizaje automático, el uso de árboles y bosques de decisión, aplicados a problemas reales. En cuanto al impacto social y empresarial, la propuesta promueve la eficiencia energética, la sostenibilidad y la modernización del sector eléctrico desde el consumo energético residencial que a futuro optimice el recurso implementando tecnología moderna de menor potencia.

En términos de viabilidad, el proyecto cuenta con acceso a múltiples fuentes de datos relevantes. A nivel internacional, existen datasets ampliamente utilizados en la investigación, como el High Resolution Smart Meter Dataset (Bensalah y Abdellatif, 2025), el Electricity Load Diagrams 2011-2014 (Trindade, 2015) y el UK-DALE Dataset (Kelly, y College, 2017), los cuales permiten validar y comparar modelos de consumo energético a nivel de electrodomésticos. A nivel nacional, se dispone de información proporcionada por instituciones como el Instituto Nacional de Estadística y Censos (INEC), a través de su plataforma de datos abiertos, así como información del sector eléctrico gestionada por ARCONEL, lo que fortalece el enfoque contextualizado del estudio, que, aunque no cuentan con información consistente en el plano de consumo nacional, sirven como base en la generación de datos sintéticos respaldados en esta literatura.

Asimismo, el desarrollo del proyecto es factible desde el punto de vista tecnológico, ya que se dispone de herramientas modernas para el diseño, implementación y despliegue de modelos de inteligencia artificial. Entre estas herramientas se incluyen entornos de desarrollo como Visual Studio Code, tecnologías de contenedorización como Docker, y plataformas para despliegue y monitoreo de modelos como Hugging Face. Estas tecnologías permiten construir soluciones escalables, reproducibles y alineadas con las prácticas actuales de la industria.

Finalmente, el proyecto es realista dentro del tiempo establecido para una maestría, ya que se enfoca en el desarrollo de un modelo específico de detección de anomalías, utilizando datasets disponibles y herramientas accesibles. La delimitación clara del alcance, junto con el uso de metodologías estructuradas como CRISP-DM, permite garantizar una ejecución ordenada y viable del proyecto.

En conclusión, la presente investigación es pertinente, relevante y viable, ya que aborda un problema real del sector energético ecuatoriano residencial mediante el uso de tecnologías innovadoras, con el potencial de generar beneficios significativos a nivel académico, tecnológico y social.

1.3 Alcance

El presente proyecto tiene como alcance el desarrollo y validación de un sistema inteligente para la detección de anomalías en el consumo energético residencial de hogares ecuatorianos mediante técnicas de aprendizaje automático. La investigación se centra en el análisis del comportamiento del consumo eléctrico mensual a nivel de vivienda, considerando variables sociodemográficas, de equipamiento eléctrico y tarifarias que influyen en el consumo de energía.

La unidad de análisis corresponde a un hogar residencial y cada registro del conjunto de datos representa un período mensual de facturación. En este contexto, la variable dependiente principal de la investigación es la clasificación del consumo como normal o potencialmente anómalo, objetivo que se aborda mediante un modelo supervisado basado en Random Forest Classifier. Como parte del proceso de detección, el sistema incorpora adicionalmente un modelo de regresión cuya finalidad es estimar el consumo energético esperado de la vivienda. Esta estimación constituye un mecanismo auxiliar que permite comparar el consumo esperado con el consumo

reportado por el usuario, proporcionando información de apoyo para la clasificación de anomalías, sin constituir el objetivo principal de la investigación.

El entrenamiento y la evaluación de los modelos se realizarán empleando un conjunto de datos sintético construido a partir de información estadística nacional proveniente del Instituto Nacional de Estadística y Censos (INEC), complementada con datasets internacionales de referencia, entre ellos UK-DALE y Household Power Consumption. La generación de estos datos permitirá representar escenarios de consumo residencial acordes con el contexto ecuatoriano y disponer de un conjunto de entrenamiento adecuado para el desarrollo y validación inicial de la propuesta.

El proyecto comprende las etapas de exploración y preparación de datos, generación del dataset sintético, ingeniería de características, entrenamiento y evaluación de modelos de aprendizaje automático, validación de resultados y desarrollo de un prototipo funcional. La evaluación del modelo de clasificación se realizará mediante métricas de desempeño apropiadas para este tipo de problemas, complementadas con técnicas de validación que permitan verificar su capacidad para identificar patrones de consumo anómalos.

Como resultado de la investigación se desarrollará un prototipo web implementado mediante Gradio y FastAPI, orientado exclusivamente a la validación técnica y metodológica de la propuesta. El sistema permitirá al usuario ingresar las características de su vivienda, estimar el consumo energético esperado, calcular el costo aproximado de la energía conforme a la normativa tarifaria vigente en el Ecuador, comparar dicho resultado con el consumo registrado en su factura eléctrica y determinar si existen posibles anomalías en el comportamiento del consumo.

Adicionalmente, el prototipo almacenará las interacciones de los usuarios en un repositorio de datos, quienes podrán generar un control estadístico de su consumo mediante el registro de su cédula con el propósito de facilitar futuras investigaciones relacionadas con el reentrenamiento del modelo y la ampliación del conjunto de datos sintéticos.

El ámbito de aplicación del estudio corresponde al consumo energético residencial del Ecuador continental y considera exclusivamente ciclos mensuales de facturación. En consecuencia, la investigación no contempla el análisis de series temporales de alta

frecuencia ni el procesamiento de información proveniente de medidores inteligentes en tiempo real, debido a la disponibilidad limitada de este tipo de infraestructura en el contexto nacional.

El presente trabajo tampoco incluye la integración del modelo con sistemas operativos de empresas distribuidoras de energía, plataformas comerciales o infraestructuras de medición inteligente. Asimismo, no contempla procesos de certificación de software, pruebas piloto con usuarios finales ni actividades propias de un sistema productivo, ya que el propósito de la investigación es demostrar la viabilidad técnica y metodológica de una solución basada en inteligencia artificial para apoyar la detección de anomalías en el consumo energético residencial.

Consideraciones metodológicas y restricciones del estudio. La disponibilidad de datos residenciales reales de alta frecuencia en el contexto ecuatoriano condiciona el diseño de la investigación, por lo que se empleará un conjunto de datos sintético construido a partir de información nacional e internacional de referencia. En consecuencia, los resultados obtenidos estarán limitados a los escenarios representados por dicho conjunto de datos y no podrán interpretarse como una validación operacional del comportamiento energético de la totalidad de los hogares ecuatorianos.

Asimismo, debido a que la unidad de análisis corresponde a un hogar durante un período mensual de facturación, el estudio no pretende modelar el comportamiento temporal horario o diario del consumo ni realizar detección de anomalías en tiempo real. La clasificación de anomalías se realizará a partir de la comparación entre el consumo reportado y el comportamiento esperado definido por el modelo, utilizando las variables disponibles en el conjunto de datos.

Estas restricciones se incorporan como condiciones de diseño de la investigación y serán consideradas durante la selección de los datos, entrenamiento, validación e interpretación de los resultados.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un sistema inteligente para la detección de anomalías en el consumo

energético residencial de hogares ecuatorianos mediante técnicas de aprendizaje automático, integrando un modelo de regresión para estimar el consumo energético esperado y un modelo de clasificación basado en Random Forest para identificar consumos potencialmente anómalos, utilizando datos sintéticos contruidos a partir de información nacional y datasets internacionales de referencia.

1.4.2 Objetivos específicos

- Realizar una revisión sistemática de la literatura sobre técnicas de inteligencia artificial aplicadas a la detección de anomalías en el consumo energético residencial, mediante el análisis de modelos, variables predictoras y datasets de referencia nacionales e internacionales, con el propósito de establecer los fundamentos metodológicos para el desarrollo de la propuesta.
- Desarrollar un modelo supervisado basado en Random Forest para la clasificación del nivel de consumo eléctrico residencial a partir de variables sociodemográficas y de equipamiento eléctrico, contrastándolo con algoritmos de otras familias mediante validación cruzada estratificada y pruebas de significancia estadística, y analizando qué parte de su desempeño es atribuible a la información disponible en los datos.
- Incorporar al sistema la normativa tarifaria vigente del sector eléctrico ecuatoriano, incluyendo la tarifa residencial escalonada, la Tarifa Dignidad y los beneficios para adultos mayores, con el fin de estimar el costo del consumo energético de acuerdo con las condiciones particulares del usuario.
- Diseñar un mecanismo de actualización del conjunto de entrenamiento mediante la incorporación de nuevos registros reales y la generación de datos sintéticos representativos, que facilite el reentrenamiento periódico del modelo y contribuya a mantener su capacidad para detectar anomalías en el consumo energético residencial.

- Desarrollar una interfaz web utilizando Gradio que permita al usuario ingresar la información de su hogar, visualizar el consumo energético esperado, el costo estimado, la clasificación del nivel de anomalía respecto al consumo reportado, recomendaciones orientadas a mejorar la eficiencia energética y análisis del historial del consumo individual del usuario.
- Containerizar el prototipo mediante Docker y desplegarlo en una plataforma en la nube para validar técnicamente la solución propuesta y facilitar su acceso con fines académicos y demostrativos.
- Evaluar el desempeño del modelo de clasificación mediante métricas de aprendizaje automático, incluyendo accuracy, precision, recall, F1-score, matriz de confusión, curva ROC, valor AUC e importancia de variables, con el propósito de analizar su capacidad para detectar anomalías en el consumo energético residencial y realizar la comparativa con otros modelos de clasificación.

CAPITULO 2:

2. REVISIÓN DE LITERATURA

2.1. Estado del Arte

El análisis del consumo energético residencial y la detección de anomalías mediante técnicas de inteligencia artificial ha sido ampliamente estudiado en los últimos años, debido al crecimiento en la disponibilidad de datos provenientes de medidores inteligentes y al avance de algoritmos de aprendizaje automático. Estas investigaciones se han centrado principalmente en el uso de modelos de series temporales, aprendizaje profundo y técnicas no supervisadas para identificar patrones atípicos en el consumo eléctrico.

2.1.1. Detección de anomalías en consumo energético residencial

El consumo energético residencial constituye un ámbito de interés para la aplicación de técnicas de inteligencia artificial debido a la variabilidad que presentan los patrones de consumo entre viviendas y a la posibilidad de identificar comportamientos que se apartan de los patrones considerados normales. En este contexto, la detección de anomalías busca identificar desviaciones que pueden estar relacionadas con

comportamientos atípicos, ineficiencias en el uso de la energía, fallas técnicas u otras situaciones que requieren análisis.

Diversas investigaciones han abordado la detección de anomalías en el consumo eléctrico mediante técnicas de aprendizaje automático. (Zhang et al., 2021) y (Himeur et al., 2021) destacan la utilidad de estas técnicas para identificar comportamientos atípicos asociados al consumo energético y evidencian la necesidad de emplear métodos capaces de adaptarse a la variabilidad propia de los patrones de consumo.

Una característica común entre estos estudios es el reconocimiento de las limitaciones de los métodos tradicionales basados exclusivamente en reglas o umbrales fijos. Estos enfoques pueden resultar insuficientes cuando el comportamiento energético presenta variaciones significativas entre usuarios o a lo largo del tiempo, debido a que un valor que puede considerarse normal para una vivienda podría representar un comportamiento inusual para otra.

Desde otra perspectiva, (MZH et al., 2021) y (Ghanim y Awad, 2024) plantean enfoques de aprendizaje no supervisado para identificar comportamientos inusuales sin depender de conjuntos de datos previamente etiquetados. Esta característica representa una ventaja cuando las anomalías no se encuentran claramente identificadas en los registros históricos. Sin embargo, este tipo de metodología también implica que la definición de lo que constituye un comportamiento anómalo depende de la capacidad del algoritmo para aprender adecuadamente el comportamiento considerado normal.

En consecuencia, los antecedentes revisados permiten identificar dos características relevantes del problema. En primer lugar, la detección de anomalías puede abordarse mediante diferentes paradigmas de aprendizaje automático dependiendo de la disponibilidad de datos etiquetados. En segundo lugar, la naturaleza del consumo energético y la variabilidad entre viviendas hacen necesario seleccionar modelos que sean compatibles con las características específicas de los datos disponibles.

2.1.2. Técnicas de aprendizaje automático aplicadas a la detección de anomalías

Las investigaciones revisadas muestran la utilización de diferentes enfoques de aprendizaje automático para abordar la detección de anomalías en el consumo

energético. Entre los principales se encuentran los métodos supervisados, los métodos no supervisados y las técnicas de aprendizaje profundo.

Los métodos supervisados requieren una variable objetivo previamente definida que permita diferenciar entre comportamientos normales y anómalos. Su principal ventaja radica en que el proceso de aprendizaje se orienta directamente hacia la clasificación establecida en los datos de entrenamiento. Sin embargo, su aplicación depende de la disponibilidad de registros adecuadamente etiquetados.

Por otro lado, los métodos no supervisados permiten identificar patrones atípicos sin requerir una clasificación previa de los registros. Este enfoque resulta especialmente relevante en escenarios donde las anomalías son difíciles de identificar manualmente o donde no existe un conjunto de datos suficientemente etiquetado. Los trabajos de (MZH et al., 2021) y (Ghanim y Awad, 2024) constituyen antecedentes dentro de esta línea.

Una tercera alternativa corresponde al aprendizaje profundo, particularmente mediante arquitecturas diseñadas para el procesamiento de datos secuenciales. (Schmidhuber, 2003) constituye una referencia teórica para el desarrollo de arquitecturas de redes neuronales recurrentes y, posteriormente, modelos como LSTM han sido utilizados para representar dependencias temporales en diferentes problemas relacionados con el consumo energético.

En esta línea, (Xu et al., 2022) emplean modelos basados en LSTM para la detección de anomalías en el consumo energético. La principal ventaja de este tipo de arquitectura se relaciona con su capacidad para representar dependencias temporales y patrones complejos presentes en datos secuenciales.

Asimismo, (Pang et al., 2020) presentan el uso de autoencoders como una alternativa para la detección de anomalías mediante el análisis del error de reconstrucción. Bajo este enfoque, el modelo aprende una representación del comportamiento considerado normal y las desviaciones pueden identificarse mediante errores de reconstrucción elevados.

Análisis comparativo

Los enfoques revisados presentan ventajas diferentes y su conveniencia depende principalmente de las características del problema. Los métodos no supervisados son apropiados cuando no existen etiquetas confiables, mientras que los modelos supervisados permiten establecer explícitamente una clasificación cuando se dispone de una variable objetivo. Por su parte, los modelos de aprendizaje profundo presentan ventajas para el análisis de relaciones temporales complejas, pero generalmente requieren mayores volúmenes de información y recursos computacionales.

Esta diferencia resulta relevante para la presente investigación, debido a que el problema se aborda mediante información residencial de periodicidad mensual y variables de naturaleza numérica y categórica. Por tanto, el modelo seleccionado debe ser capaz de trabajar con datos tabulares y representar relaciones entre las características del hogar y la clasificación de su consumo.

Bajo estas condiciones, Random Forest constituye una alternativa pertinente para el componente de clasificación de la propuesta, debido a su capacidad para trabajar con múltiples variables y representar relaciones no lineales. Además, su utilización permite analizar la importancia relativa de las variables empleadas durante el proceso de clasificación.

Es importante señalar que esta selección no implica que Random Forest sea superior de manera general a los modelos de aprendizaje profundo o a los métodos no supervisados. La selección responde a las características y restricciones particulares del problema abordado en esta investigación.

2.1.3. Datasets y características de los datos utilizados

El desarrollo de modelos de inteligencia artificial para el análisis del consumo energético ha sido favorecido por la disponibilidad de conjuntos de datos públicos provenientes de diferentes contextos residenciales.

Entre los conjuntos de datos utilizados como referencia en las investigaciones revisadas se encuentran **UK-DALE**, **Pecan Street Dataset**, **Electricity Load Diagrams 2011–2014** y **High Resolution Smart Meter Dataset**. Estos conjuntos proporcionan información de consumo energético con diferentes niveles de granularidad y características, permitiendo estudiar el comportamiento de los hogares y, en determinados casos, el consumo asociado a electrodomésticos.

En particular, el conjunto UK-DALE constituye una referencia relevante para el análisis del consumo residencial debido a que proporciona información de demanda eléctrica de viviendas y de determinados electrodomésticos. (Kelly y Knottenbelt, 2015) describen este conjunto de datos como una fuente para estudiar la demanda eléctrica a nivel de vivienda y electrodomésticos.

Los datasets internacionales presentan una ventaja importante: permiten disponer de información real y suficientemente detallada para desarrollar y evaluar modelos de aprendizaje automático. Sin embargo, su utilización directa en un contexto diferente puede presentar limitaciones de representatividad, debido a que los patrones de consumo están condicionados por factores como las características de las viviendas, los hábitos de los usuarios, el equipamiento eléctrico, las condiciones climáticas y la estructura tarifaria de cada país.

Esta situación resulta especialmente relevante para el contexto ecuatoriano. La disponibilidad de información nacional procedente de fuentes como el Instituto Nacional de Estadística y Censos permite incorporar características relacionadas con la realidad sociodemográfica del país. No obstante, la ausencia de un conjunto nacional específicamente etiquetado para la detección de anomalías en el consumo residencial constituye una limitación para la aplicación directa de modelos supervisados.

Por esta razón, la presente investigación plantea la construcción de un conjunto de datos sintético contextualizado al Ecuador, utilizando información nacional como referencia y datasets internacionales como fuentes para representar patrones y escenarios de consumo residencial. Este enfoque permite disponer de información estructurada para el entrenamiento y evaluación del modelo, aunque sus resultados deberán interpretarse considerando la naturaleza sintética de los datos.

2.1.4. Comparación de los enfoques empleados en investigaciones previas

A partir de la revisión realizada, es posible establecer diferencias importantes entre las principales estrategias utilizadas para la detección de anomalías.

Tabla 1

Comparación de enfoques para la detección de anomalías en el consumo energético

Enfoque	Característica Principal	Ventaja	Limitación	Pertinencia para esta investigación
Aprendizaje Supervisado	Utiliza una variable objetivo previamente definida	Permite orientar directamente la clasificación	Requiere datos etiquetados	Alta, debido a la definición de la variable anomalía
Aprendizaje no Supervisado	Identifica desviaciones sin etiquetas previas	Útil cuando no existen registros etiquetados	La definición del comportamiento anómalo depende del modelo	Permite establecer un referente metodológico, pero no corresponde al enfoque principal
LSTM modelos recurrentes	/ Analizan dependencias temporales	Adecuados para patrones temporales complejos	Requieren datos secuenciales y mayores recursos	Limitada, debido a la granularidad mensual y

					estructura tabular
Autoencoders	Detectan desviaciones mediante reconstrucción	No requieren necesariamente etiquetas explícitas	Dependen de una adecuada representación del comportamiento normal	Alternativa metodológica, pero no seleccionada	
Random Forest	Clasificación mediante múltiples árboles de decisión	Adecuado para variables heterogéneas y relaciones no lineales	Requiere una definición adecuada de la variable objetivo	Alta, por las características de los datos y el enfoque supervisado	

Interpretación de la comparación

La comparación evidencia que la selección de una técnica de inteligencia artificial no puede realizarse únicamente considerando los resultados de precisión reportados por otros estudios. También deben considerarse la naturaleza de los datos, la disponibilidad de etiquetas, la granularidad temporal, los recursos computacionales y el objetivo específico de la investigación.

Los modelos recurrentes y otras arquitecturas de aprendizaje profundo resultan especialmente apropiados cuando el problema requiere representar dependencias temporales complejas. Sin embargo, la presente investigación no pretende analizar el consumo horario o subhorario, sino trabajar con registros residenciales correspondientes a períodos mensuales de facturación.

Por otra parte, los métodos no supervisados constituyen una alternativa cuando no se dispone de etiquetas de anomalía. No obstante, el presente proyecto establece una variable dependiente que permite formular el problema como una clasificación supervisada. Esta diferencia justifica que el modelo principal seleccionado sea Random Forest.

2.1.5. Análisis Crítico de investigaciones previas

El análisis conjunto de las investigaciones revisadas permite identificar varias convergencias. En primer lugar, existe consenso respecto de la utilidad de las técnicas de aprendizaje automático para identificar patrones que pueden resultar difíciles de detectar mediante reglas estáticas. En segundo lugar, los estudios coinciden en que la calidad, cantidad y granularidad de los datos constituyen factores determinantes para el desempeño de los modelos.

Sin embargo, también existen diferencias importantes. Los estudios basados en aprendizaje profundo se orientan principalmente hacia problemas donde la dimensión temporal de los datos desempeña un papel fundamental, mientras que los enfoques no supervisados responden a escenarios donde la disponibilidad de etiquetas constituye una limitación. Los modelos supervisados, por el contrario, requieren una definición explícita del comportamiento que se desea clasificar, pero ofrecen una formulación más directa del problema cuando dicha información puede establecerse.

Otra diferencia significativa corresponde al contexto de los datos. Los principales datasets utilizados en las investigaciones revisadas provienen de otros países y, aunque constituyen fuentes valiosas para el desarrollo y evaluación de modelos, no necesariamente representan las condiciones específicas de los hogares ecuatorianos. Variables como la composición del hogar, el equipamiento eléctrico, los hábitos de consumo y la estructura tarifaria pueden modificar sustancialmente el comportamiento energético observado.

En este sentido, se identifica una diferencia entre la disponibilidad de metodologías avanzadas para detección de anomalías y su adaptación a contextos donde no existen datasets nacionales etiquetados con las características requeridas. Esta situación constituye uno de los principales elementos que motivan el planteamiento de la presente investigación.

Además, una parte de los trabajos revisados se concentra principalmente en la evaluación del algoritmo. La presente propuesta busca complementar esta perspectiva mediante el desarrollo de un prototipo funcional que integre el procesamiento de datos, el modelo de aprendizaje automático, el cálculo tarifario y la interacción con el usuario.

2.1.6. Vacíos de investigación identificados

A partir de la comparación crítica realizada, se identifican los siguientes vacíos:

1. Falta de contextualización al entorno ecuatoriano

Los estudios revisados utilizan principalmente datasets internacionales. Si bien estos conjuntos constituyen referencias metodológicas relevantes, existe una limitada representación de las características específicas de los hogares ecuatorianos.

2. Limitada disponibilidad de datos nacionales etiquetados

La aplicación de métodos supervisados requiere una variable objetivo que permita distinguir entre comportamientos normales y anómalos. La ausencia de un dataset nacional con este tipo de etiquetas constituye una dificultad para entrenar directamente modelos supervisados.

3. Diferencia en la granularidad de los datos

Una parte importante de los estudios analizados trabaja con datos obtenidos mediante sistemas de medición inteligente y con granularidad horaria o subhoraria. Este tipo de infraestructura no corresponde al alcance de la presente investigación, que trabaja con información mensual asociada a períodos de facturación.

4. Distancia entre el modelo y su aplicación

Algunas investigaciones se concentran en el desarrollo y evaluación de algoritmos, mientras que existe una oportunidad para integrar el modelo en una herramienta funcional que permita demostrar su aplicación desde la perspectiva del usuario residencial.

5. Necesidad de adaptar el análisis a variables heterogéneas

El problema abordado en esta investigación integra variables sociodemográficas, características del equipamiento eléctrico, consumo y aspectos tarifarios. Esto requiere un enfoque capaz de procesar información heterogénea y establecer relaciones entre estas variables y la clasificación del comportamiento energético.

2.1.7. Justificación de la propuesta desarrollada

Los vacíos identificados permiten establecer la necesidad de una propuesta que considere las características particulares del contexto ecuatoriano y las restricciones relacionadas con la disponibilidad de datos. En respuesta a esta situación, la presente investigación plantea un sistema inteligente orientado a la detección de anomalías en el consumo energético residencial mediante aprendizaje automático supervisado.

La propuesta utiliza un conjunto de datos sintético construido a partir de información nacional y referencias internacionales, con el propósito de representar escenarios de consumo residencial en hogares ecuatorianos. El conjunto considera variables relacionadas con las características del hogar, su equipamiento eléctrico y el comportamiento mensual del consumo.

Desde el punto de vista metodológico, el sistema integra dos componentes con funciones diferenciadas. El modelo de regresión se utiliza como componente auxiliar para estimar el consumo energético esperado de acuerdo con las características del hogar. Esta estimación proporciona una referencia para el análisis posterior. El modelo de clasificación basado en Random Forest constituye el componente principal de la investigación y tiene como finalidad clasificar el consumo como normal o potencialmente anómalo.

Esta diferenciación permite evitar que la estimación del consumo y la detección de anomalías sean consideradas como dos objetivos independientes. La regresión cumple una función de apoyo dentro del flujo de procesamiento, mientras que la clasificación constituye el problema principal abordado por la investigación.

Finalmente, la propuesta integra el modelo desarrollado en un prototipo web que incorpora el cálculo del costo energético conforme a la normativa tarifaria ecuatoriana y proporciona al usuario información relacionada con su consumo. De esta manera, el proyecto busca demostrar no solamente la viabilidad del modelo de inteligencia artificial, sino también la factibilidad técnica de integrar dicho modelo en una solución funcional orientada al análisis del consumo energético residencial.

2.2.Marco Teórico

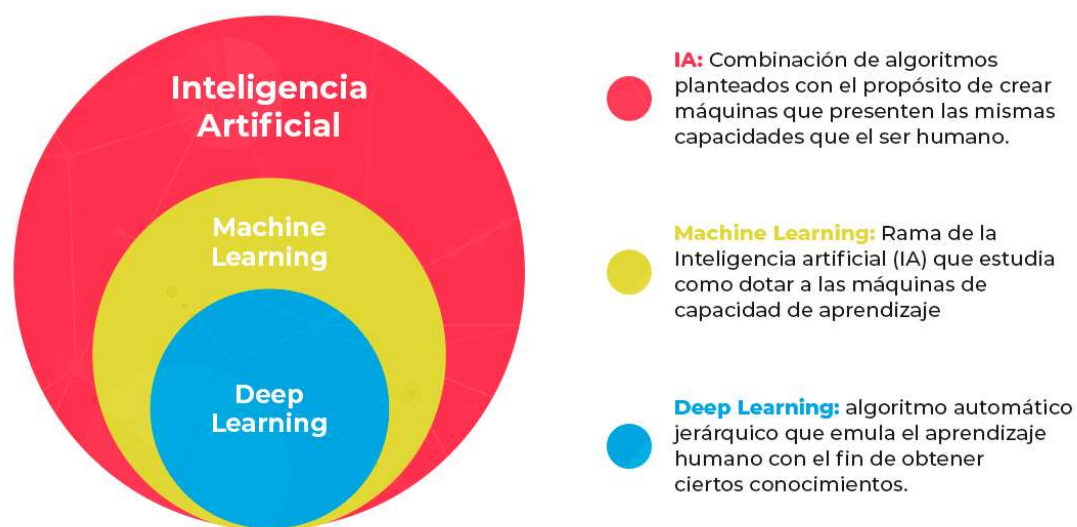
2.2.1. Inteligencia artificial y aprendizaje automático

La Inteligencia Artificial es una rama de la informática que busca desarrollar sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento y la toma de decisiones (A. Boden, 2016).

Dentro de esta disciplina se encuentra el Aprendizaje Automático, el cual permite a los sistemas aprender a partir de datos sin ser programados explícitamente. En lugar de definir reglas fijas, los modelos identifican patrones y relaciones dentro de los datos. Para presentar un panorama más amplio de la inteligencia artificial y su relación con el aprendizaje automático se presenta la siguiente figura, donde se puede observar como la inteligencia artificial es una disciplina amplia que contiene otras dos como el aprendizaje automático (machine learning) y aprendizaje profundo (Deep learning).

Figura 1

Inteligencia Artificial y sus ramas



Nota: La inteligencia artificial es una disciplina muy amplia que contiene al machine learning y a la vez este, contiene al Deep learning que es una rama de las otras dos. (Madrid, 2021)

El aprendizaje automático se divide principalmente en:

- Supervisado: usa datos etiquetados
- No supervisado: encuentra patrones sin etiquetas

- Semi-supervisado: combina ambos enfoques

En este proyecto se utiliza aprendizaje supervisado, ya que se busca clasificar el consumo energético como normal o anómalo.

2.2.2. Detección de anomalías

La detección de anomalías es una técnica que consiste en identificar datos que se desvían significativamente del comportamiento normal. Otros autores (Feng y otros, 2020) definen las anomalías como valores atípicos, desviaciones, discordancias o excepciones. Clasificando las anomalías en los siguientes tipos:

- **Anomalías puntuales:** Esta anomalía se asocia con un punto de datos puntual de un conjunto de datos, que representa el caso más simple y común.
- **Anomalías contextuales:** Un dato puede considerarse anómalo en un contexto específico, como el caso de este trabajo, el contexto energético.
- **Anomalías colectivas:** Una anomalía deja de ser puntual cuando los datos están referidos a un conjunto espacio – temporal.

Hablando del contexto energético de este trabajo, una anomalía puede representar:

- Fraude eléctrico
- Errores de medición
- Consumo ineficiente
- Fallas en equipos

Este problema es complejo porque:

- Las anomalías son poco frecuentes
- No siempre están etiquetadas
- El comportamiento normal puede variar

2.2.3. Set de datos o datasets

En el campo de la inteligencia artificial y el machine learning, es fundamental disponer de conjuntos de datos que permitan el estudio, entrenamiento y validación de los modelos computacionales desarrollados. Estos conjuntos de datos, conocidos

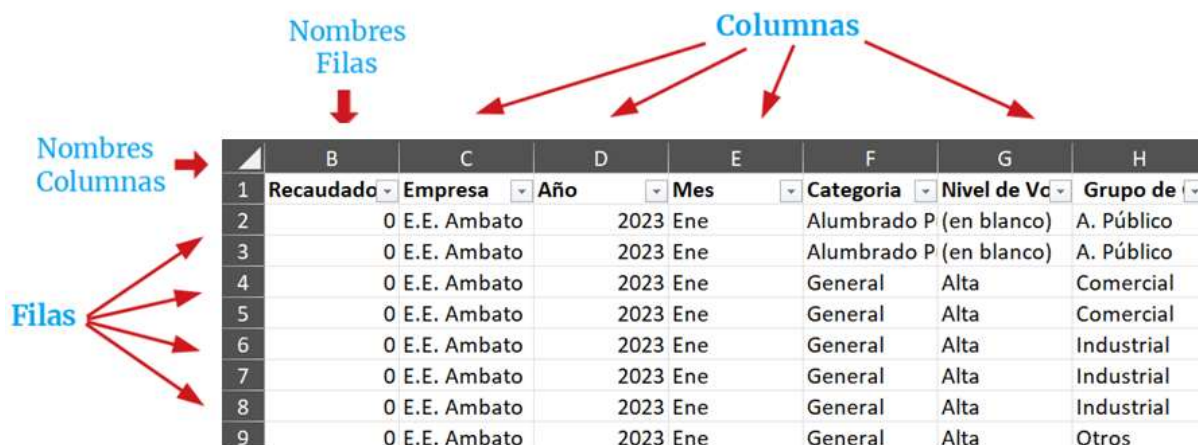
como datasets, constituyen la base para la construcción de algoritmos capaces de identificar patrones, realizar predicciones y detectar anomalías en diferentes contextos.

Dentro de los distintos tipos de datasets, destaca la estructura tabular, la cual organiza la información en filas y columnas, facilitando su procesamiento, análisis e interpretación. En este formato, cada fila representa una observación o instancia, mientras que cada columna corresponde a una variable o característica, lo que permite una adecuada representación de datos estructurados para su posterior modelado. En otras palabras y en el contexto de este trabajo las filas y columnas tendrán la información similar a lo siguiente:

- **Cada fila** representa una observación (ejemplo: un hogar, persona o cliente).
- **Cada columna** representa una variable o atributo (ejemplo: consumo eléctrico, ingreso, tipo de vivienda).

Figura 2

Ejemplo de un set de datos con filas y columnas



El diagrama muestra una tabla de datos con las siguientes columnas: Recaudado, Empresa, Año, Mes, Categoría, Nivel de Vc, Grupo de. Las filas están numeradas del 1 al 9. Las etiquetas 'Nombres Filas' y 'Nombres Columnas' están en la parte superior, y 'Filas' está a la izquierda. Las flechas indican la correspondencia entre las etiquetas y los datos en la tabla.

	B	C	D	E	F	G	H
1	Recaudado	Empresa	Año	Mes	Categoría	Nivel de Vc	Grupo de
2		0 E.E. Ambato	2023	Ene	Alumbrado P (en blanco)	A. Público	
3		0 E.E. Ambato	2023	Ene	Alumbrado P (en blanco)	A. Público	
4		0 E.E. Ambato	2023	Ene	General	Alta	Comercial
5		0 E.E. Ambato	2023	Ene	General	Alta	Comercial
6		0 E.E. Ambato	2023	Ene	General	Alta	Industrial
7		0 E.E. Ambato	2023	Ene	General	Alta	Industrial
8		0 E.E. Ambato	2023	Ene	General	Alta	Industrial
9		0 E.E. Ambato	2023	Ene	General	Alta	Otros

Nota: El dataset de la imagen fue tomado de (Renovables-ARCERNNR, 2023) y elaborado por el autor Alexander Orbe.

En el presente proyecto, el modelo de aprendizaje automático se entrena utilizando un conjunto de datos sintético, diseñado para representar patrones de consumo eléctrico residencial acordes con las características de los hogares ecuatorianos. La generación de este conjunto de datos se fundamenta en información proveniente de diversas fuentes de referencia, las cuales aportan conocimiento técnico, regulatorio y

sociodemográfico necesario para construir un conjunto de entrenamiento representativo.

Como principal referencia para la modelación del consumo energético se emplea el dataset, UK-DALE (UK Domestic Appliance-Level Electricity) (Kelly, Ukerc, 2017), el cual contiene mediciones detalladas del consumo eléctrico de electrodomésticos en viviendas reales del Reino Unido. Este conjunto de datos permite identificar patrones de uso, perfiles de consumo y relaciones entre los diferentes equipos eléctricos, constituyendo la base para definir el comportamiento energético de los hogares representado posteriormente en el conjunto de datos sintético.

Adicionalmente, se utiliza información publicada por la Agencia de Regulación y Control de Electricidad (ARCONEL) (Renovables-ARCERNR, 2023), , correspondiente a la estructura tarifaria del servicio eléctrico residencial en Ecuador, incluyendo rangos de consumo, tarifas escalonadas, subsidios y beneficios sociales como la Tarifa Dignidad y los descuentos para adultos mayores. Estos datos no forman parte del entrenamiento del modelo, sino que se incorporan al sistema para estimar el costo mensual del consumo eléctrico conforme a la normativa vigente.

Por otra parte, se consideran los datos de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) (INEC, 2025), elaborada por el Instituto Nacional de Estadística y Censos, con el propósito de caracterizar variables sociodemográficas representativas de los hogares ecuatorianos, tales como región geográfica, número de habitantes y condiciones socioeconómicas. Esta información sirve como referencia para definir distribuciones realistas de las variables empleadas durante la generación del conjunto de datos sintético.

A partir de estas fuentes de información se construye un conjunto de datos sintético que integra variables relacionadas con las características del hogar, el equipamiento eléctrico, los hábitos de uso de los electrodomésticos y el consumo mensual estimado de energía. Este proceso permite disponer de un volumen suficiente de registros para el entrenamiento y validación del modelo de aprendizaje supervisado, preservando las relaciones estadísticas observadas en las fuentes de referencia y adaptándolas al contexto ecuatoriano.

2.2.4. Integración técnica de los datasets (Data Fusion)

La integración de los conjuntos de datos **UK-DALE**, **ARCONEL** y **ENEMDU** se realiza mediante un proceso de data fusion, cuyo objetivo es unificar información heterogénea - temporal, técnica, socioeconómica y regulatoria, en una única estructura analítica que permita el entrenamiento de modelos de machine learning.

En primer lugar, se identifican las características propias de cada dataset. El dataset **UK-DALE** contiene series temporales de consumo eléctrico a nivel de electrodomésticos con alta granularidad (segundos o minutos), mientras que los datos de **ARCONEL** presentan información estructurada sobre tarifas eléctricas, subsidios y categorías de usuarios. Por su parte, la **ENEMDU** del Instituto Nacional de Estadística y Censos proporciona variables sociodemográficas y económicas organizadas en formato tabular a nivel de hogar y persona.

Dado que no existe una clave primaria común entre estos datasets, la integración se realiza mediante un enfoque de fusión por enriquecimiento contextual, en lugar de una unión directa (join tradicional). Este proceso se desarrolla en las siguientes etapas:

1. **Preprocesamiento y limpieza de datos:** Se realiza la depuración de valores nulos, tratamiento de datos atípicos y estandarización de formatos. En el caso de **UK-DALE**, se agregan las series temporales a intervalos consistentes (por ejemplo, consumo horario o diario), mientras que en **ENEMDU** y **ARCONEL** se normalizan las variables categóricas y numéricas.
2. **Transformación y generación de variables (feature engineering):** Se construyen variables derivadas que permitan la compatibilidad entre datasets. Por ejemplo:
 - De **UK-DALE**: consumo promedio, demanda máxima, factor de carga, patrones horarios.
 - De **ARCONEL**: categoría tarifaria, nivel de subsidio, costo estimado del consumo.
 - De **ENEMDU**: nivel socioeconómico, región geográfica, tamaño del hogar.
3. **Homologación de niveles de agregación:** Se establece un nivel común de análisis, generalmente a nivel de hogar. Dado que **UK-DALE** no corresponde a hogares ecuatorianos, se realiza una simulación o asignación probabilística donde

los perfiles de consumo eléctrico se asocian a tipologías de hogares definidas a partir de ENEMDU.

4. **Integración mediante enriquecimiento contextual:** Se asignan a cada perfil de consumo (UK-DALE) variables externas provenientes de ENEMDU y ARCONEL. Este proceso puede realizarse mediante:
 - Reglas heurísticas (ej. asignar tarifas según nivel de consumo).
 - Muestreo estadístico condicionado.
 - Técnicas de matching por similitud (clustering o nearest neighbors).
5. **Construcción del dataset final:** Se obtiene un dataset unificado en formato tabular donde cada registro representa un hogar (real o simulado) con:
 - Variables de consumo energético
 - Variables tarifarias y regulatorias.
 - Variables sociodemográficas.

Este dataset final es utilizado como entrada para los modelos de machine learning, permitiendo incorporar múltiples dimensiones del problema en el proceso de entrenamiento.

En conjunto, esta estrategia de integración permite superar la limitación de no contar con datos unificados reales, generando un entorno analítico robusto que combina información técnica, económica y social, lo cual resulta fundamental para la detección de anomalías en el consumo eléctrico residencial.

2.2.5. Set de Datos Sintético

El modelo de aprendizaje automático fue desarrollado utilizando un conjunto de datos estructurado que representa las características de consumo eléctrico de hogares residenciales. El dataset constituye la principal fuente de información para el entrenamiento, validación y posterior actualización del modelo de detección de anomalías energéticas.

La versión actual del conjunto de datos contiene **5 010 registros y 16 variables**, donde cada registro corresponde a un hogar y describe tanto sus características generales como su comportamiento de consumo energético. En la siguiente tabla

se puede observar una descripción de cada una de las variables utilizadas en el set de datos final.

Tabla 2

Descripción de las variables del set de datos utilizado

ítem	Variable	Descripción
1	id_hogar	Identificador único del hogar dentro del conjunto de datos.
2	Región	Región geográfica del Ecuador donde se ubica la vivienda (Costa, Sierra u Oriente).
3	num_habitantes	Número de personas que habitan la vivienda.
4	tiene_refrigeradora	Indica si el hogar dispone de refrigeradora (Sí/No).
5	tiene_lavadora	Indica si el hogar dispone de lavadora (Sí/No).
6	tiene_microondas	Indica si el hogar dispone de microondas (Sí/No).
7	tiene_computadora	Indica si el hogar dispone de computador (Sí/No).
8	tiene_aire_acondicionado	Indica si el hogar dispone de aire acondicionado (Sí/No).
9	tiene_calefon_electrico	Indica si el hogar dispone de calefón eléctrico (Sí/No).
11	num_televisores	Número de televisores presentes en la vivienda.
12	consumo_esperado	Consumo eléctrico estimado para el hogar de acuerdo con sus características.

13	Anomalia	Variable objetivo utilizada para indicar si el consumo corresponde a un comportamiento normal (0) o anómalo (1).
14	consumo_reportado	Consumo eléctrico registrado o ingresado por el usuario.
15	Mes	Mes correspondiente al registro del consumo energético.
16	Cédula	Identificador anonimizado del usuario o del origen del registro.

2.2.6. Metodología de construcción del set de datos sintético

La construcción del conjunto de datos sintético responde a la necesidad de disponer de información representativa sobre el consumo energético residencial en Ecuador, debido a la ausencia de bases de datos públicas que integren variables demográficas, geográficas y de equipamiento doméstico adecuadas para el entrenamiento de un modelo de detección de anomalías energéticas.

Para ello, se definieron un conjunto de variables consideradas relevantes en el comportamiento del consumo eléctrico de una vivienda, tomando como referencia estudios relacionados con eficiencia energética, consumo residencial y monitoreo de carga eléctrica, además de las características propias de los hogares ecuatorianos.

Las variables seleccionadas se agruparon en cuatro categorías principales:

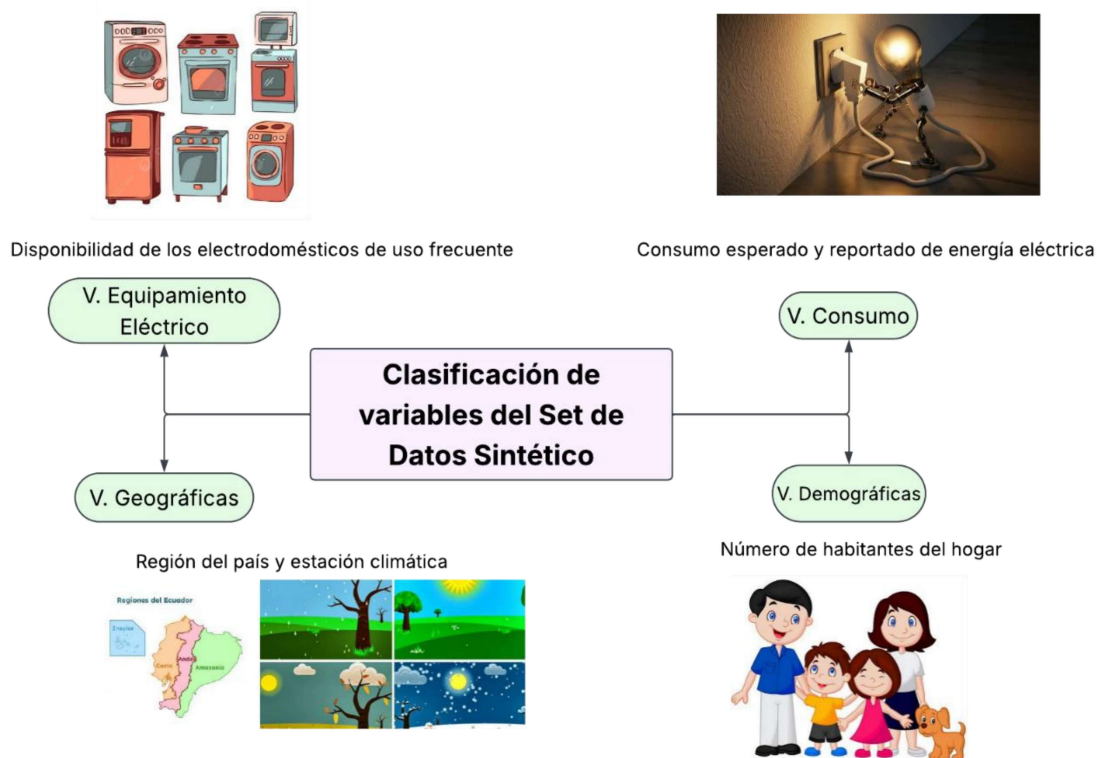
- **Variables geográficas**, como la región del país.
- **Variables demográficas**, representadas por el número de habitantes del hogar.
- **Variables de equipamiento eléctrico**, relacionadas con la disponibilidad

de electrodomésticos de uso frecuente y de alto consumo energético.

- **Variables de consumo**, correspondientes al consumo esperado y al consumo reportado de energía eléctrica.

Figura 3

Clasificación de las variables del Set de Datos Sintético



Nota: La clasificación de las variables del set de datos sintético, agrupa las variables desde el aspecto geográfico como la región sierra costa u oriente, el aspecto demográfico como el número de habitantes y llegando hasta la parte técnica como es el consumo esperado y reportado de energía eléctrica.

Elaborado por el Autor Alexander Orbe.

2.2.6.1. Definición de los rangos de las variables

Cada variable fue definida considerando valores compatibles con el contexto residencial ecuatoriano.

La variable región contempla las principales zonas geográficas del país (Costa, Sierra y Oriente) se exceptúa Galápagos, ya que es más difícil poder recolectar información en esta región.

El número de habitantes fue definido dentro de un rango representativo de los hogares ecuatorianos, permitiendo simular viviendas unifamiliares y familias numerosas.

Las variables correspondientes a la presencia de electrodomésticos (refrigeradora, lavadora, microondas, computadora, aire acondicionado y calefón eléctrico) fueron modeladas mediante variables binarias que indican la existencia o ausencia de cada equipo dentro del hogar.

Adicionalmente, se incorporó el número de televisores como una variable discreta para representar diferentes niveles de equipamiento tecnológico.

2.2.6.2. Generación del consumo esperado

El consumo esperado representa una estimación del consumo mensual de energía eléctrica que debería presentar una vivienda de acuerdo con sus características.

Su cálculo considera la combinación de las variables demográficas y del equipamiento eléctrico, asumiendo que el incremento en el número de habitantes y la disponibilidad de electrodomésticos producen un aumento gradual en la demanda energética.

Esta variable constituye la referencia utilizada por el sistema para evaluar el comportamiento energético de cada vivienda.

2.2.6.3. Generación del consumo reportado

El consumo reportado corresponde al valor observado o registrado para cada vivienda.

Durante la construcción del dataset sintético, este valor fue generado simulando diferentes escenarios de consumo alrededor del consumo esperado, incorporando variaciones que representan cambios normales en los hábitos de uso de la energía eléctrica.

Asimismo, se incluyeron escenarios con incrementos significativamente superiores al comportamiento esperado, con el propósito de representar posibles anomalías de consumo.

Esta estrategia permitió disponer de ejemplos correspondientes tanto a condiciones normales como a situaciones anómalas necesarias para el entrenamiento supervisado del modelo.

2.2.6.4. Etiquetado de anomalías

Una vez generados los valores de consumo esperado y consumo reportado, se procedió a asignar la variable objetivo denominada anomalía.

Cada registro fue clasificado en una de dos categorías:

- 0 (consumo normal): cuando el consumo reportado se mantiene dentro de un rango aceptable respecto al consumo esperado.
- 1 (consumo anómalo): cuando la razón entre el consumo reportado y el consumo esperado es igual o superior a 1.30, es decir, cuando el consumo facturado supera en al menos un 30 % al esperado para ese perfil de hogar.

El umbral de 1.30 constituye el criterio único del sistema y se implementa en un solo módulo del código, empleado por la interfaz, por el almacenamiento de interacciones y por la evaluación del modelo. Esta unificación corrige una inconsistencia de la implementación previa, en la que convivían tres definiciones distintas —1.30 en el etiquetado del conjunto, un umbral adaptativo de 1.10, 1.20 o 1.30 en la interfaz según la probabilidad del clasificador, y 1.20 al persistir el registro—, de modo que una lectura con razón 1.25 podía mostrarse al usuario como normal y almacenarse simultáneamente como anómala en el conjunto de entrenamiento.

Este proceso permitió construir un conjunto de datos etiquetado para entrenar un algoritmo de clasificación supervisada capaz de identificar patrones asociados a posibles anomalías energéticas.

2.2.6.5. Validación del conjunto de datos

El conjunto de datos sintético fue revisado para garantizar la consistencia entre las variables, evitando registros con combinaciones imposibles o inconsistentes.

Posteriormente, el dataset fue utilizado para entrenar el modelo inicial de aprendizaje automático y evaluar su capacidad para distinguir entre patrones normales y anómalos de consumo energético.

Finalmente, el sistema fue diseñado para incorporar nuevos registros generados durante su utilización, permitiendo actualizar el conjunto de datos y reentrenar periódicamente el modelo. Esta estrategia favorece un proceso de aprendizaje continuo, en el que el conocimiento inicial obtenido mediante datos sintéticos puede enriquecerse progresivamente con información proveniente de escenarios reales de consumo energético residencial.

2.2.7. Random Forest: Fundamentos y funcionamiento

El algoritmo Random Forest es una técnica de machine learning basada en el aprendizaje por ensamblado (ensemble learning), introducida por Leo Breiman en 2001 (Breiman, 2001). Este método combina múltiples árboles de decisión con el objetivo de mejorar la precisión predictiva y reducir el sobreajuste (overfitting), siendo ampliamente utilizado en problemas de clasificación y regresión.

El principio fundamental de Random Forest radica en la construcción de un conjunto de árboles de decisión independientes, donde cada uno es entrenado a partir de diferentes subconjuntos de datos y variables. Este enfoque permite que el modelo capture diversas relaciones presentes en los datos, aumentando su capacidad de generalización (Sistemas Inteligentes, 2021).

2.2.7.1. Fundamentos del algoritmo

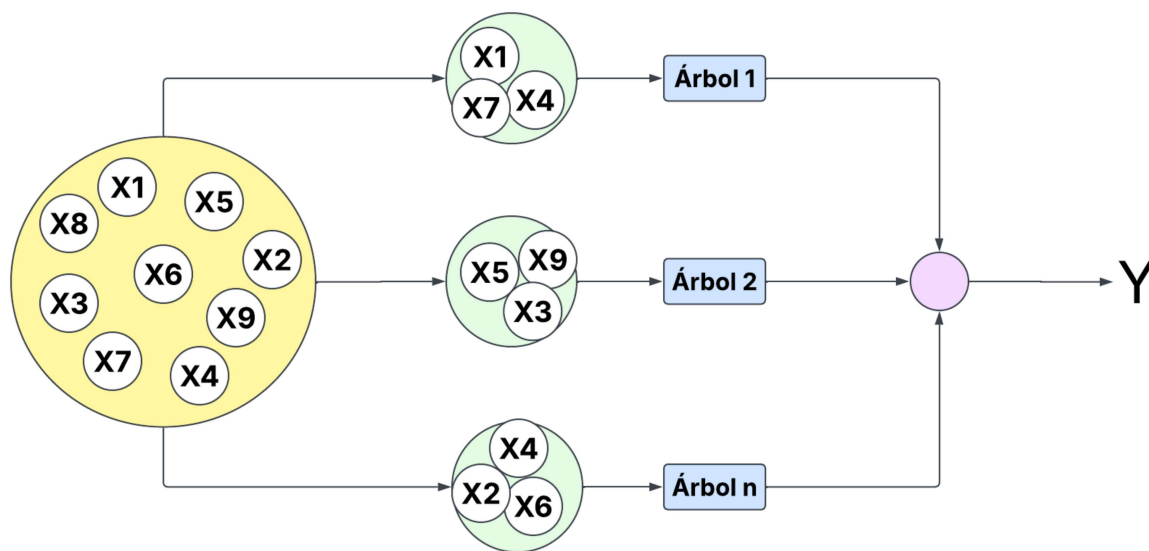
El algoritmo Random Forest pertenece a una técnica de aprendizaje supervisado que se basa en árboles de decisión. El mismo que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento. La ventaja de este algoritmo es que el resultado de cada árbol de decisión es combinado con el objetivo de obtener un resultado más robusto (Espinoza Zuñiga, 2020). Cada árbol se obtiene mediante un proceso de dos etapas:

1. Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables m (predictores) de forma que $m < M$ (donde M = total de predictores).
2. Cada árbol crece hasta su máxima extensión.

En la siguiente figura se aprecia el fundamento del algoritmo Random Forest; donde la idea fundamental de este algoritmo es combinar varios modelos para obtener un modelo final más robusto y preciso.

Figura 4

Algoritmo Random Forest



Nota: El algoritmo Random Forest es un árbol de decisión compuesto por varios árboles de decisión los cuales determinan un resultado individual, que al final es modelado en un resultado final para obtener mejores resultados (Espinoza Zuñiga, 2020). Elaborado por el autor Alexander Orbe.

2.2.7.2. Funcionamiento del modelo

El proceso de funcionamiento de Random Forest puede describirse en las siguientes etapas (Jain, 2024) (geeksforgeeks, 2025):

1. **Construcción de múltiples árboles de decisión:** Se generan N árboles, cada uno entrenado con un subconjunto bootstrap del dataset.
2. **Crecimiento de los árboles:** Cada árbol se construye dividiendo recursivamente los datos en función de la mejor variable seleccionada (según criterios como Gini o entropía), considerando únicamente un subconjunto aleatorio de características en cada división.
3. **Predicción individual:** Cada árbol produce una predicción independiente.

4. Agregación de resultados

- En clasificación: se utiliza votación mayoritaria.
- En regresión: se calcula el promedio de las predicciones.

2.2.7.3. Ventajas del Random Forest

Entre las principales ventajas del algoritmo se destacan:

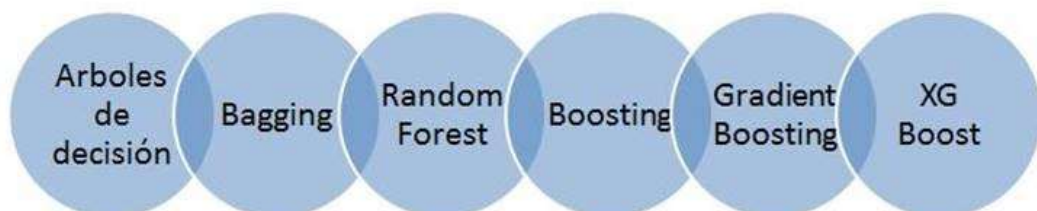
- Alta precisión en comparación con modelos individuales.
- Reducción del sobreajuste debido al promedio de múltiples árboles.
- Capacidad para manejar grandes volúmenes de datos y variables.
- Robustez frente a valores atípicos y ruido.
- Capacidad de estimar la importancia de las variables.

Estas características lo convierten en un modelo adecuado para problemas complejos con datos heterogéneos, como el análisis del consumo eléctrico residencial.

En el estado del arte y evolución de los árboles de decisión existen otros algoritmos que pueden brindar mejores resultados que el Random Forest, pero que sin embargo por su complejidad no son considerados. En la siguiente figura, se puede observar cómo ha sido la evolución de los árboles de decisión.

Figura 5

Evolución de los algoritmos basados en árboles de decisión



Nota: En la evolución de los algoritmos basados en árboles de decisión, se puede observar dos algoritmos interesantes; el que se está implementando en este trabajo (Random Forest) y el XG Boost (Espinoza Zuñiga, 2020).

A continuación, se presenta una comparación de las principales ventajas y limitaciones de Random Forest y XGBoost, considerando sus características para el tratamiento de datos tabulares y su posible aplicación en problemas de clasificación y regresión.

Tabla 3

Comparativa entre Random Forest y XGBoost

Algoritmo	Ventajas	Desventajas
Random Forest	• Permite resolver problemas de clasificación y regresión. • Presenta buen desempeño con variables numéricas y categóricas adecuadamente codificadas. • Es relativamente robusto frente al ruido y al sobreajuste debido al uso de múltiples árboles. • Permite estimar la importancia de las variables, facilitando la interpretación de los factores asociados a la predicción. • Su configuración y entrenamiento son relativamente sencillos para conjuntos de datos tabulares.	• Puede requerir un número elevado de árboles para obtener un buen desempeño, aumentando el costo computacional. • La interpretación individual del modelo puede resultar compleja debido a la combinación de múltiples árboles. • Puede presentar menor desempeño que métodos de boosting en determinados problemas. • No realiza extrapolaciones fuera del rango observado en los datos de entrenamiento.
XGBoost	• Presenta un alto desempeño predictivo en problemas de clasificación y regresión. • Incorpora mecanismos de regularización que ayudan a controlar el sobreajuste.	• Su configuración y ajuste de hiperparámetros pueden ser más complejos. • Puede requerir mayor conocimiento técnico para obtener un desempeño óptimo.

-
- | | |
|--|---|
| <ul style="list-style-type: none"> • Puede manejar conjuntos de datos de alta dimensionalidad y relaciones no lineales. • Permite obtener medidas de importancia de las variables. • Ofrece opciones avanzadas de optimización y ajuste de hiperparámetros. | <ul style="list-style-type: none"> • El entrenamiento puede demandar mayores recursos computacionales dependiendo del tamaño y complejidad del conjunto de datos. • Su comportamiento puede resultar menos sencillo de interpretar que el de modelos más simples. |
|--|---|
-

Nota: De la tabla comparativa entre los algoritmos de Random Forest y XG Boost, se puede concluir, que el último obtiene mejores resultados, pero puede consumir más recursos computacionales, por lo cual se justifica en este trabajo la utilización del algoritmo de Random Forest sobre XG Boost.

2.2.7.4. Aplicación en la detección de anomalías

En el contexto de este proyecto, Random Forest se emplea para clasificar patrones de consumo eléctrico como normales o anómalos. La integración de los datasets UK-DALE, los datos tarifarios de ARCONEL y la información sociodemográfica de la ENEMDU del Instituto Nacional de Estadística y Censos permite construir un modelo que considera múltiples dimensiones del problema.

El modelo aprende patrones de comportamiento energético a partir de variables como consumo promedio, variabilidad temporal, nivel socioeconómico y estructura tarifaria. De esta manera, es capaz de identificar desviaciones significativas respecto a patrones esperados, lo cual resulta clave para la detección de anomalías, posibles fraudes o consumos ineficientes.

2.2.7.5. Justificación del uso de Random Forest

La elección de Random Forest en este estudio se fundamenta en su capacidad para:

- Integrar variables de distinta naturaleza (numéricas y categóricas).
- Manejar relaciones no lineales entre variables.
- Ofrecer interpretabilidad mediante la importancia de características.

- Mantener un buen desempeño sin requerir un ajuste excesivo de hiperparámetros.

Estas propiedades lo convierten en una alternativa adecuada frente a otros modelos más complejos, especialmente en escenarios donde se combinan datos técnicos, económicos y sociales.

Esta elección se contrasta empíricamente en la sección 4.4. Bajo la formulación que corresponde al sistema desplegado, los ocho algoritmos evaluados se agrupan en una banda estrecha de F1 (entre 0.7115 y 0.7883 sobre la partición de retención) pese a que su capacidad de representación varía entre una frontera lineal y un aproximador universal (Hornik, Stinchcombe y White, 1989). Cuando familias con mecanismos de inducción tan distintos convergen a un rendimiento equivalente, el límite observado es atribuible a la información contenida en las variables predictoras y no a la elección del algoritmo.

En ese contexto la interpretabilidad deja de ser una preferencia metodológica y se convierte en un criterio de decisión: el sistema debe explicar al usuario por qué su consumo se considera elevado, y la importancia de variables de Random Forest (Breiman, 2001) proporciona esa explicación de forma directa, mientras que los modelos de mayor capacidad requerirían métodos de atribución adicionales. A ello se suma un hallazgo del análisis de sensibilidad de la sección 4.4.7: al introducir dispersión realista en el consumo de los hogares sin anomalía, Random Forest degrada menos que el perceptrón multicapa y termina por superarlo. Dado que los datos de campo presentarán dispersión, esta propiedad resulta más relevante que el rendimiento sobre el conjunto sintético actual.

Se deja constancia de que, en la formulación que incorpora el consumo facturado como variable predictora, SVM (RBF), MLP, XGBoost y LightGBM superan a Random Forest con diferencias estadísticamente significativas tras la corrección de Holm-Bonferroni. La elección de Random Forest se sostiene pese a ello, sobre los criterios anteriores, y no a costa de omitirlo.

2.2.8. Métricas de evaluación en el aprendizaje automático

En aprendizaje automático (Machine Learning) una vez que se ha diseñado el modelo y se lo ha entrenado, es necesario evaluar que tan bueno es el modelo, y validar si cumple con los objetivos delineados.

Existen técnicas de validación y evaluación que permiten medir el rendimiento de un modelo de aprendizaje automático, estas técnicas varían dependiendo de la tarea y objetivo que cumple el modelo, esto es, tareas de clasificación, regresión y agrupamiento. En este trabajo nos enfocaremos específicamente en las métricas de evaluación centradas en las tareas de clasificación.

En las tareas de clasificación los modelos se encargan de predecir categorías discretas, como específicamente hablando en este trabajo, indicar si existe anomalía o no. Dentro de las métricas de clasificación, encontramos las siguientes técnicas:

2.2.8.1. Exactitud (Accuracy)

Esta métrica indica la proporción de predicciones correctas sean positivas o negativas respecto al total de predicciones.

$$Accuracy = \frac{\text{Clasificaciones Correctas}}{\text{Clasificaciones Totales}} = \frac{VP + VN}{VP + VN + FP + FN}$$

Donde:

- VP: Verdaderos Positivos
- VN: Verdaderos Negativos
- FP: Falsos Positivos
- FN: Falsos Negativos

Es importante indicar que esta métrica puede ser engañosa para evaluar modelos de aprendizaje automático en casos donde los sets de datos se encuentran desbalanceados, ya que puede presentar resultados buenos, pero que en la práctica no lo son. Por ejemplo, en un set de datos con un 95 % de clase X y un 5 % de clase Y, un modelo que prediga solo la clase X seguirá alcanzando una precisión

del 95 %, pero no logrará identificar ninguna instancia de la clase Y (Campo Yepes, 2026).

2.2.8.2. Precisión

Esta métrica mide cuantas de las clasificaciones positivas son realmente positivas. Esto es útil en un escenario donde los falsos positivos tienen un costo bastante alto. Por ejemplo, en el campo de la medicina, un falso positivo sería detectar cáncer en un paciente que realmente no lo tiene, esto tendría graves consecuencias en el paciente (Campo Yepes, 2026).

$$\text{Precisión} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos positivos} + \text{Falsos Positivos}} = \frac{VP}{VP + FP}$$

2.2.8.3. Sensibilidad o Recall

Esta métrica es algo parecida a la precisión, pero en vez de enfocarse en los falsos negativos. En otras palabras, la sensibilidad mide cuantos de los casos positivos reales fueron identificados realmente por el modelo. Poniendo el mismo ejemplo anterior, en la medicina puede tener un alto costo, decirle a un paciente que no tiene nada, cuando en realidad si lo tiene, falso negativo, imagínense las consecuencias igual de graves que serían (Campo Yepes, 2026).

$$\text{Sensibilidad o Recall} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos positivos} + \text{Falsos Negativos}} = \frac{VP}{VP + FN}$$

2.2.8.4. F1-Score

Anteriormente vimos la precisión y sensibilidad, dos métricas igual de importantes, una concentrada en la importancia de los falsos positivos y otra en los falsos negativos respectivamente, pero que al final sabemos que las dos son importantes.

Es aquí donde entra a trabajar esta métrica F1-Score, cuando se requiere un equilibrio entre precisión y sensibilidad. Y el F1-Score se define de esa forma como la media armónica de la precisión y sensibilidad (Campo Yepes, 2026).

$$F1 - Score = 2 * \frac{\text{Precisión} * \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

2.2.8.5. AUC-ROC

Para entender esta métrica conocida como curva AUC-ROC, es importante conocer dos conceptos, uno de ellos que ya se lo menciono anteriormente, y es la sensibilidad o recall, que en este caso lo vamos a llamar como la tasa de verdaderos positivos (TPR). Y el otro concepto que es la tasa de falsos positivos (FPR).

La Tasa de Falsos Positivos (FPR) es la proporción de todos los negativos reales que fueron declarados incorrectamente como positivos, de mejor forma también conocido como **la probabilidad de falsa alarma**.

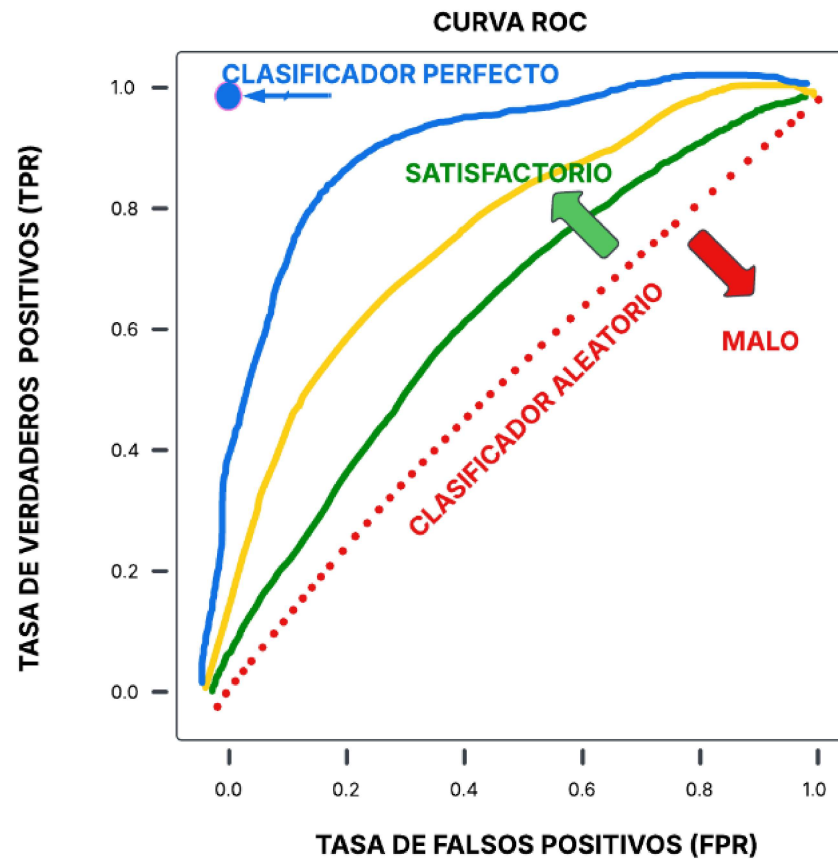
$$FPR = \frac{\text{Falsos Positivos}}{\text{Falsos positivos} + \text{Verdaderos Negativos}} = \frac{FP}{FP + VN}$$

Ahora comprendidos estos conceptos nos adentramos en la curva ROC y el área bajo esta curva se conoce como AUC, que generalmente oscila entre 0.0 a 1.0. Representándose un 0.0 como un modelo bastante malo y un 1.0 como un modelo satisfactorio (Bouza-Herrera, 2021).

Ahora entendiendo más gráficamente la curva ROC y su entorno mencionamos lo siguiente:

- En el eje Y se encuentra representado por la tasa de verdaderos positivos (TPR).
- En el eje X se encuentra representado por la tasa de falsos positivos (FPR)
- La curva comienza en la esquina inferior izquierda con coordenadas (0,0) y se extiende hasta la esquina superior derecha (1,1). Es decir, que gráficamente los ejes X y Y están comprendidos desde 0 hasta 1.

En la siguiente figura se observa algunas curvas ROC y su comportamiento.

Figura 6*Curva AUC ROC*

Nota: En la figura se puede observar diferentes curvas ROC, de las cuales la de color azul es la que tiene mejor comportamiento, entre más se acerca al punto azul, mejor será la curva ROC, y obviamente tendrá un AUC más cercano a 1.0. Elaborado por el autor Alexander Orbe.

2.2.8.6. Matriz de Confusión

Una matriz de confusión es una métrica de evaluación muy útil para evaluar modelos; su principal objetivo es mostrar de forma clara y estructurada, el comportamiento de los algoritmos de inteligencia artificial al comparar las predicciones realizadas con los valores reales del set de datos (Barrios Arce, 2019).

En una matriz de confusión en forma generalizada se encuentran cuatro escenarios:

- **Verdadero Positivo – True Positive VP/ TP:** El valor real es positivo y el modelo también predice positivo.
- **Verdadero Negativo – True Negative VN/TN:** El valor real es negativo y el modelo predice negativo.
- **Falso Negativo – False Negative FN:** El valor real es positivo, pero el modelo predice negativo.
- **Falso Positivo – False Positive FP:** El valor real es negativo, pero el modelo predice positivo.

En la siguiente figura se puede ver en forma general la matriz de confusión:

Figura 7

Matriz de confusión

		CLASES REALES	
		REAL POSITIVO	REAL NEGATIVO
CLASES PREDICHAS	PREDICHO POSITIVO	VERDADERO POSITIVO (VP)	FALSO POSITIVO (FP)
	PREDICHO NEGATIVO	FALSO NEGATIVO (FN)	VERDADERO NEGATIVO (VN)

2.2.9. Restricciones y consideraciones metodológicas

El diseño metodológico de la investigación se encuentra condicionado por la disponibilidad y naturaleza de los datos utilizados, así como por el propósito de desarrollar un prototipo funcional para la validación técnica y metodológica de un sistema de detección de anomalías en el consumo energético residencial. Estas condiciones fueron consideradas desde la definición del conjunto de datos, la selección de los modelos y el proceso de evaluación, con el propósito de mantener coherencia entre los objetivos planteados, los recursos disponibles y el alcance de los resultados.

En primer lugar, debido a la disponibilidad limitada de registros reales de consumo eléctrico residencial ecuatoriano con el nivel de detalle requerido para el entrenamiento supervisado, se optó por construir un conjunto de datos sintético a partir de información estadística nacional y datasets internacionales de referencia. En consecuencia, los registros empleados para el entrenamiento y evaluación representan escenarios de consumo residencial contruidos mediante reglas y distribuciones definidas metodológicamente, por lo que los resultados obtenidos deben interpretarse dentro de las condiciones de generación del conjunto de datos y no como una representación exhaustiva de todos los hogares del Ecuador.

En segundo lugar, la investigación establece como unidad de análisis el hogar residencial durante un período mensual de facturación. Por esta razón, el estudio no contempla el análisis de series temporales de alta frecuencia, patrones horarios o diarios ni detección de anomalías en tiempo real. Esta decisión responde a la disponibilidad de información y al alcance definido para el prototipo, por lo que el modelo se orienta a identificar desviaciones respecto al comportamiento esperado durante un ciclo mensual.

En tercer lugar, debido a que el modelo principal corresponde a un algoritmo de aprendizaje supervisado, es necesario disponer de una variable objetivo que permita distinguir entre comportamientos normales y potencialmente anómalos. En el conjunto de datos sintético, estas etiquetas se generan mediante criterios definidos durante la construcción de los registros y representan escenarios de desviación respecto al consumo esperado. Por tanto, las etiquetas no deben interpretarse como confirmaciones de anomalías determinadas mediante inspecciones técnicas de una empresa distribuidora.

En cuarto lugar, la evaluación del modelo se realizará mediante métricas de aprendizaje automático y procedimientos de validación aplicados al conjunto de datos disponible. En consecuencia, las métricas obtenidas representan el desempeño experimental del modelo bajo las condiciones del estudio y no constituyen una garantía de desempeño en sistemas eléctricos reales ni ante poblaciones de usuarios no representadas en los datos de entrenamiento.

Finalmente, el prototipo se desarrolla con fines académicos y de validación técnica, sin integración con sistemas operativos de empresas distribuidoras, infraestructura de medición inteligente ni procesos de certificación o pruebas piloto propias de sistemas productivos. Estas restricciones delimitan la interpretación de los resultados y establecen que la investigación busca demostrar la viabilidad metodológica de la propuesta, dejando la validación con datos reales y la evaluación operativa como líneas de trabajo posteriores.

CAPÍTULO 3:

3: DESARROLLO

3.1. Arquitectura del Sistema

El sistema de detección de anomalías en consumo energético residencial se compone de múltiples módulos interconectados que trabajan en conjunto para procesar datos, entrenar modelos y presentar resultados. La arquitectura sigue el patrón de diseño modular, separando responsabilidades claras entre componentes. En otras palabras, el presente sistema corresponde a una arquitectura modular por capas (Layered Architecture) con componentes de Machine Learning integrados (Gómez Palomo y Moraleda Gil, 2020) (Arias y Durango, 2016).

3.1.1. Componentes del Sistema

El sistema está compuesto por las siguientes capas principales:

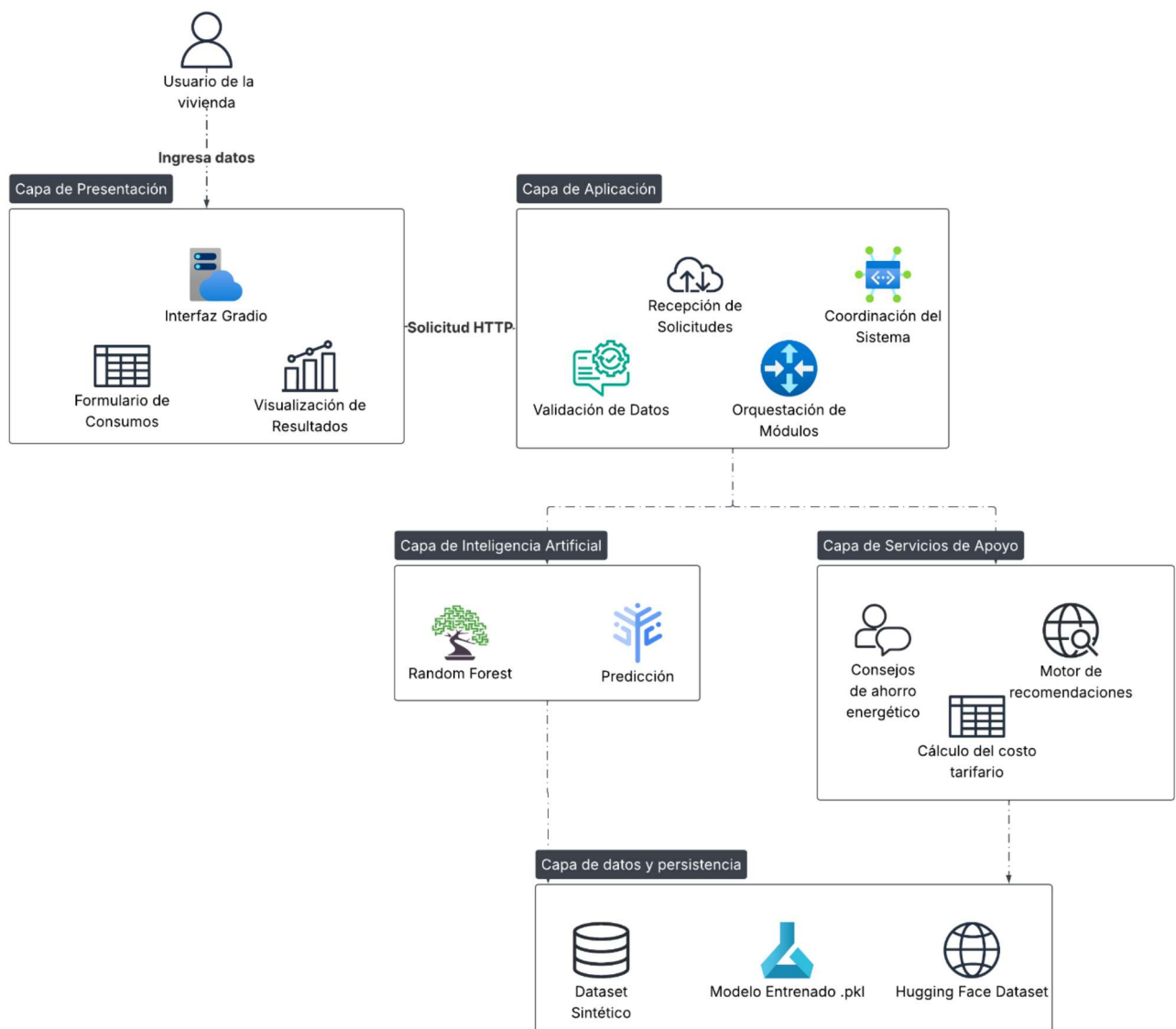
1. **Capa de presentación:** interfaz desarrollada con Gradio para la interacción con el usuario.
2. **Capa de aplicación:** FastAPI, encargada de la validación de datos, coordinación del flujo y comunicación entre módulos.
3. **Capa de inteligencia artificial:** modelo de Random Forest responsable de la

predicción del consumo esperado y la detección de anomalías.

4. **Capa de servicios de apoyo:** módulos de cálculo tarifario, generación de recomendaciones y acceso a enlaces oficiales de distribuidoras eléctricas.
5. **Capa de datos y persistencia:** dataset sintético, modelo entrenado (.pkl) y almacenamiento de nuevas interacciones en archivos CSV o en un repositorio de Hugging Face.

Figura 8

Arquitectura General del Sistema



Nota: La arquitectura general del “Sistema de Detección de Anomalías Energéticas en Hogares Ecuatorianos mediante Random Forest: Un Enfoque Híbrido con Datos Sintéticos y

Referencias Internacionales”, muestra cinco componentes fundamentales modulares interconectados, los mismos que cumplen una función en específico para cumplir con el objetivo de este proyecto. Elaborado por el autor Alexander Orbe.

3.1.2. Diagrama de componentes UML

Un diagrama de componentes – UML, es perfecto para identificar cada uno de los componentes principales que interactúan en este sistema.

Usuario. - El usuario representa el actor externo del sistema. No forma parte del software; simplemente interactúa con él proporcionando información sobre su vivienda y su consumo eléctrico.

Por ejemplo:

- Región;
- Número de habitantes;
- Electrodomésticos;
- Consumo mensual.

Toda esta información constituye la entrada del sistema.

Componente de FastAPI. – Este es el componente más importante de toda la arquitectura. En UML suele denominarse componente coordinador porque comunica todos los módulos.

Cuando Gradio envía la información, FastAPI:

1. Recibe la solicitud;
2. Valida los datos;
3. Decide qué módulos deben ejecutarse;
4. Recopila todos los resultados;
5. Responde nuevamente a Gradio.

En términos de arquitectura, FastAPI implementa la lógica de negocio del sistema.

Componente de Interfaz Gradio. - Este componente corresponde a la capa de presentación.

Su responsabilidad es únicamente interactuar con el usuario. Recibe los datos introducidos en el formulario y posteriormente muestra:

- La predicción;
- El costo estimado;
- Las recomendaciones;
- La clasificación del consumo.

Es importante destacar que Gradio no realiza cálculos, únicamente actúa como intermediario entre el usuario y el backend.

Componente de IA. – Este componente contiene todo lo relacionado con la predicción. Su única responsabilidad es responder preguntas como:

"¿Existe una anomalía en este consumo?"

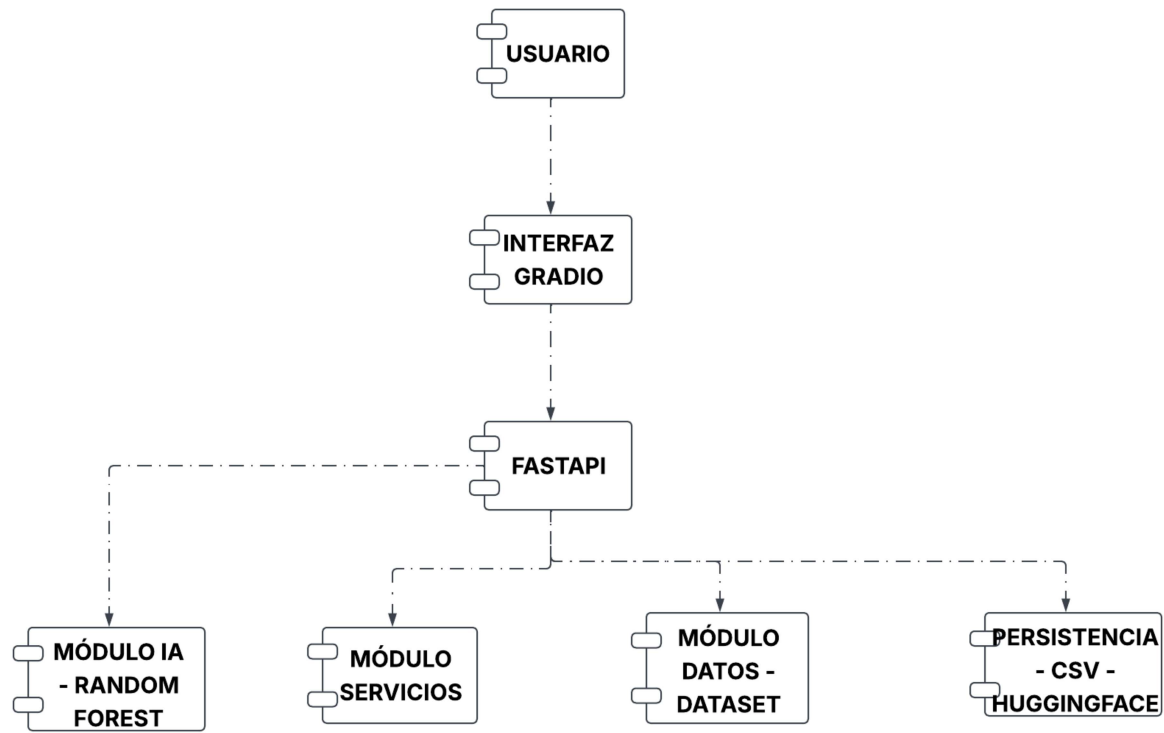
No calcula costos ni almacena información.

Componente de Servicios. – Este módulo recibe el resultado del modelo de IA y, a partir de él:

- Calcula el costo del consumo;
- Genera recomendaciones de ahorro;
- Identifica la empresa distribuidora correspondiente.

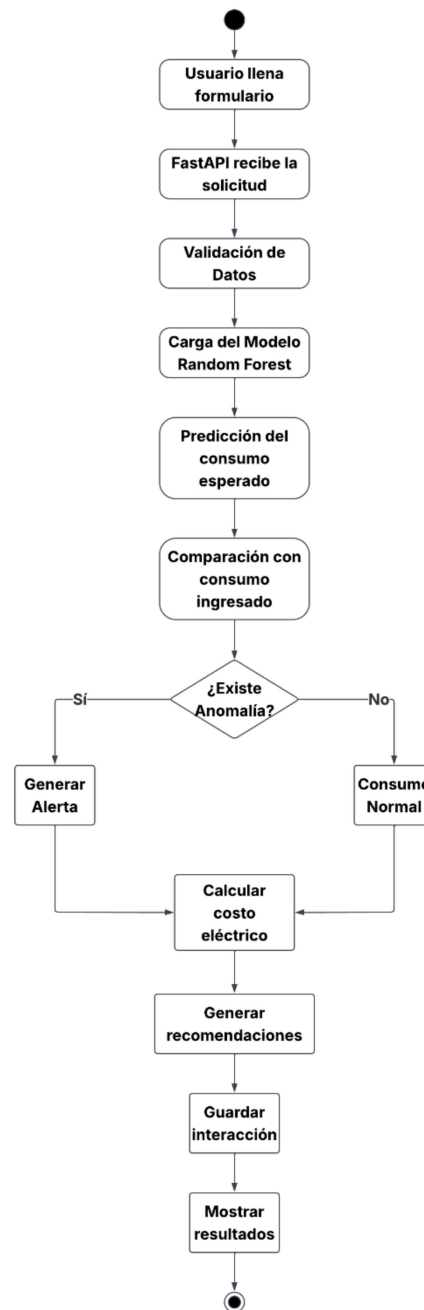
Componente de Datos. – Este componente administra toda la información utilizada por el sistema. Aquí no se realizan predicciones. Su responsabilidad es proporcionar datos al modelo y gestionar el acceso a la información.

Componente de Persistencia. – Finalmente, este componente almacena la información. Es independiente del resto. Su única función es guardar las nuevas interacciones.

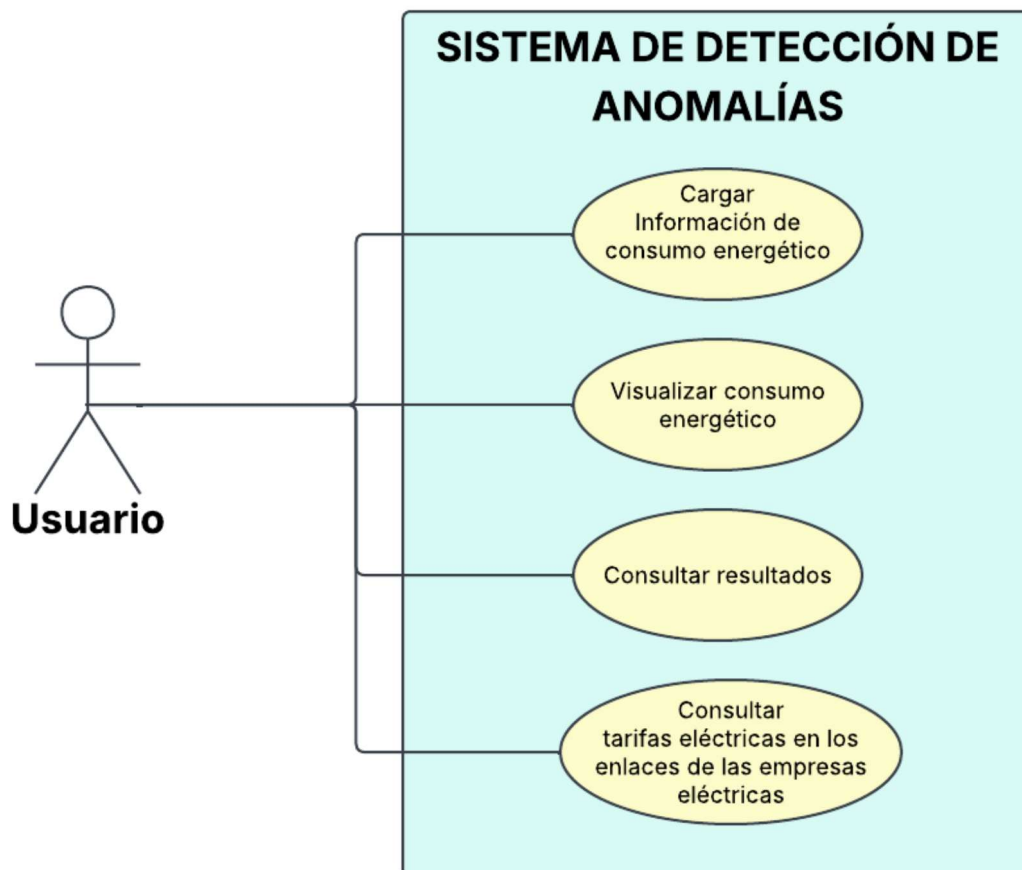
Figura 9*Diagrama de Componentes del Sistema*

Nota: El diagrama de componentes del sistema muestra todos los actores que participan en la entrada, procesamiento y salida del sistema. Principalmente estos son los componentes que en forma general constituyen el sistema, puede haber otros elementos que no se los vea explícitamente, pero que se encuentran constituidos dentro de estos componentes. Elaborado por Alexander Orbe.

Finalmente, y no menos importante se presenta el diagrama de flujo interno de procesamiento.

Figura 10*Diagrama de Flujo del Sistema***3.1.3. Diagrama de casos de uso UML**

Un diagrama de caso de uso UML representa la relación que existe entre un actor o actores que interactúan con el sistema. De aquí que los elementos que componen este diagrama son: actores, casos de uso y relaciones (Vega, 2010).

Figura 11*Diagrama de Caso de Uso UML*

Nota: En el diagrama de uso se puede observar la interacción del actor con el sistema, y los casos de uso asociados. Elaborado por el autor Alexander Orbe.

3.2. Desarrollo del set de datos sintético

El punto de partida para elaborar un proyecto de inteligencia artificial son los datos. La búsqueda de información que nos sirva de base para el desarrollo de un proyecto de detección de anomalías en el sector energético residencial es limitada. Aunque existen conjuntos de datos internacionales ampliamente utilizados en investigación, estos fueron recopilados en contextos geográficos, climáticos, socioeconómicos y regulatorios diferentes al Ecuador, por lo que su utilización directa podría introducir sesgos y disminuir la capacidad de generalización del modelo.

Por esta razón, el proyecto adopta una estrategia basada en la construcción de un dataset sintético, entendido como un conjunto de registros generados artificialmente mediante reglas estadísticas y de negocio, preservando el comportamiento esperado del consumo energético, utilizando información real y verídica que la proporcionan los usuarios que interactúan con el sistema desarrollado.

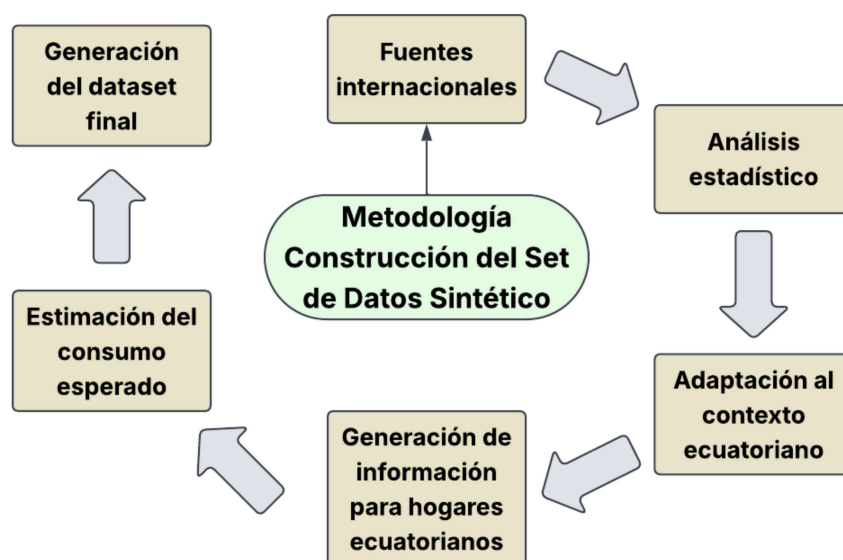
El objetivo principal del dataset sintético es proporcionar una base suficientemente amplia y representativa para entrenar un modelo de aprendizaje supervisado capaz de estimar el consumo mensual esperado de una vivienda y detectar posibles anomalías de consumo.

3.2.1. Metodología de Construcción del Set de Datos Sintético

Para la metodología de construcción del set de datos sintético se ha considerado las fuentes internacionales y la normativa eléctrica ecuatoriana, adaptados al comportamiento energético de los hogares ecuatorianos. El proceso de construcción se divide en seis etapas principales, conforme se indica en la siguiente figura.

Figura 12

Metodología de Construcción de Datos Sintéticos



3.2.2. Integración de fuentes de información

Dataset Household Power Consumption

El primer paso del pipeline de procesamiento consiste en transformar el dataset Household Power Consumption de su formato TXT original a un formato estructurado en CSV. Este

dataset contiene mediciones de consumo eléctrico con granularidad minuto a minuto (Hebrail & Berard, 2012).

Proceso de transformación:

1. Lectura del archivo TXT con separador punto y coma (;)
2. Conversión de fecha y hora a formato datetime
3. Transformación de valores '?' a NaN y conversión a tipo numérico
4. Eliminación de registros con valores nulos
5. Agregación horaria para reducir granularidad y ruido
6. Conversión de Wh a kWh para estandarización

Este procesamiento genera perfiles horarios de consumo eléctrico que representan patrones temporales reales de hogares. Los datos agregados incluyen consumo total, consumo por cocina, lavandería y sistemas térmicos (calefón/aire acondicionado).

Dataset UK-DALE

El procesamiento del dataset UK-DALE (Kelly, Ukerc, 2017) permite obtener consumos específicos por electrodoméstico. Este dataset está almacenado en formato HDF5, lo que requiere un procesamiento especializado.

Metodología de extracción:

- Apertura del archivo HDF5 mediante la librería h5py
- Iteración sobre cada edificio (building1-building5) contenido en el archivo
- Lectura de canales de medición (elec) correspondientes a diferentes electrodomésticos
- Extracción de mediciones con filtrado de valores válidos (>0)
- Clasificación automática de electrodomésticos mediante análisis del nombre del canal
- Cálculo de consumo mensual promedio en kWh

En caso de que no se puedan extraer datos del archivo HDF5, el sistema aplica valores de respaldo provenientes de la literatura científica (Kelly & Knottenbelt, 2015), asegurando la continuidad del procesamiento.

3.2.3. Generación de los registros sintéticos

Los registros sintéticos se consideran a partir de las variables del set de datos descritas en la [sección 2.2.5 Set de Datos Sintético](#):

- Región geográfica: Costa (45%), Sierra (40%), Oriente (15%)
- Estrato socioeconómico: Con distribuciones diferenciadas por región
- Número de habitantes: De 1 a 8 personas con distribución probabilística
- Electrodomésticos presentes: Probabilidad de posesión según estrato socioeconómico
- Número de televisores: De 1 a 5, dependiendo del estrato socioeconómico

3.2.4. Cálculo de Consumo Esperado

Para cada perfil generado, se calcula el consumo esperado mensual mediante un modelo determinativo que considera:

- **Consumo base por electrodomésticos (valores actualizados 2024-2025):**
 - Refrigeradora: 35 kWh/mes (tecnología inverter)
 - Lavadora: 18 kWh/mes (motores eficientes)
 - Microondas: 8 kWh/mes
 - Televisor: 10 kWh/mes (tecnología LED)
 - Computadora: 6 kWh/mes
- **Equipos de alto consumo con horas de uso variables:**
 - Aire acondicionado: 1.0 kW (tecnología inverter)
 - Calefón eléctrico: 3.0 kW
- **Factores de ajuste:**
 - Factor por habitantes: $0.7 + (\text{num_habitantes} \times 0.1)$
 - Factor regional: Costa +15%, Oriente +10%, Sierra base
 - Factor estacional: Invierno en Costa +20%, en Sierra -5%
 - Variación aleatoria: $\pm 10\%$ para simular diferencias naturales

3.2.5. Generación de Anomalías

Para entrenar el modelo de detección, se genera un conjunto sintético de anomalías que representa el 15% del dataset total. Estas anomalías se crean mediante:

- Selección aleatoria de hogares (seed=42 para reproducibilidad)
- Multiplicación del consumo esperado por un factor aleatorio uniforme, que produce

razones observadas entre 1.300 y 1.974 sobre el conjunto generado

- Etiquetado como anomalía=1 (hogares normales tienen anomalía=0)

Este enfoque permite que el modelo aprenda a distinguir entre consumos normales y anormalmente elevados, simulando situaciones como fallas en equipos, fraudes eléctricos o usos ineficientes.

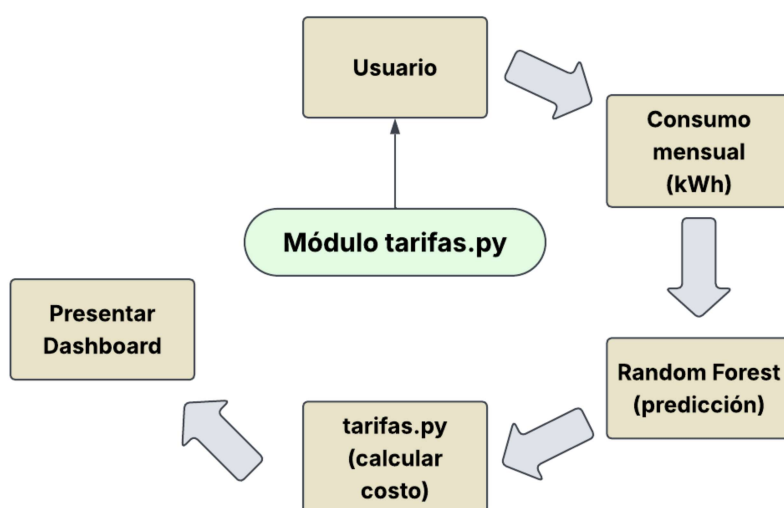
3.3. Sistema de Tarifas Eléctricas

El módulo `tarifas.py` implementa el sistema tarifario ecuatoriano según normativa de ARCONEL, incluyendo subsidios y descuentos aplicables a diferentes tipos de usuarios. El propósito principal de este módulo es administrar las tarifas eléctricas y calcular el costo estimado de un consumo residencial.

Dentro de la arquitectura del proyecto, este módulo no realiza predicciones; esa tarea corresponde al modelo Random Forest. Su función es tomar el consumo (en kWh) generado por el modelo o ingresado por el usuario y convertirlo en un valor económico.

Figura 13

Flujo del Módulo Tarifas.py



En relación a las funciones implementadas en el módulo `tarifas.py`, se describen en la tabla siguiente:

Tabla 4*Funciones del módulo tarifas.py*

Función	Acción
cargar_tarifas()	Obtiene las tarifas vigentes
guardar_tarifas()	Guarda nuevas tarifas
calcular_costo()	Calcula el costo de un consumo
obtener_info_tarifas_html()	General el panel HTML del dashboard

La fórmula matemática que se aplica en el módulo tarifas.py es la siguiente:

$$\text{Costo} = \text{Consumo [kWh]} \times \text{Tarifa [\$]}$$

3.3.1. Estructura Tarifaria

El sistema implementa tarifas escalonadas:

- Rango 1: 0-100 kWh → \$0.09/kWh
- Rango 2: 101-300 kWh → \$0.10/kWh
- Rango 3: 301-500 kWh → \$0.12/kWh
- Rango 4: >500 kWh → \$0.15/kWh

3.3.2. Beneficios y Descuentos

El sistema contempla tres categorías de usuarios con beneficios diferenciados:

- **Adulto Mayor (65+ años):**
 - 50% de descuento en los primeros 138 kWh
 - Consumos superiores pagan tarifa normal escalonada
- **Tarifa Dignidad:**
 - \$0.04/kWh hasta límite regional:
 - Costa: 130 kWh
 - Sierra: 110 kWh

- Oriente: 130 kWh
- Excedentes pagan tarifa normal escalonada
- **Residencial General:**
 - Tarifa normal escalonada sin descuentos

3.4. Entrenamiento del Modelo Random Forest

3.4.1. Preparación de Datos

El proceso de preparación incluye:

- Carga del dataset integrado (dataset_integrado_final.csv)
- Creación de variable objetivo: consumo_alto = 1 si consumo > percentil 80
- Selección de features predictoras (sin incluir 'estrato' para evitar sesgos)
- Codificación one-hot de variables categóricas (región, num_habitantes)
- División train/test: 70% entrenamiento, 30% prueba (stratified)

3.4.2. Configuración del Modelo

Hiperparámetros seleccionados para el Random Forest:

- n_estimators: 150 árboles de decisión
- max_depth: 12 niveles máximos

Los hiperparámetros se fijaron en valores estándar y se mantuvieron constantes en todas las comparaciones entre algoritmos. No se realizó búsqueda de hiperparámetros: el objeto de estudio es el efecto de la información disponible sobre el rendimiento y no el máximo alcanzable por cada modelo. Optimizar cada algoritmo por separado introduciría una variable de confusión en la comparación y, al hacerse sobre el mismo conjunto, un riesgo adicional de sobreajuste al protocolo de evaluación.

- random_state: 42 (reproducibilidad)
- class_weight: 'balanced' (compensar desbalance de clases)

La configuración equilibrada evita que el modelo favorezca la clase mayoritaria y permite detectar eficientemente consumos anómalos que son minoritarios en el dataset.

3.4.3. Análisis de Importancia de Features

El modelo Random Forest proporciona métricas de importancia que indican qué variables contribuyen más a la predicción. Este análisis permite:

- Identificar los factores más determinantes del consumo alto
- Validar la coherencia del modelo con el conocimiento del dominio
- Generar recomendaciones focalizadas en los equipos más relevantes

3.5. Aplicación Web con FastAPI y Gradio

El módulo app.py implementa una interfaz web completa que permite a usuarios finales interactuar con el sistema de detección de anomalías.

3.5.1. Arquitectura de la Aplicación

La aplicación utiliza dos frameworks complementarios:

- **FastAPI: Backend RESTful para endpoints de API**
 - Endpoint /health: Monitoreo de estado del sistema
 - Endpoint /detectar: Procesamiento de perfiles de usuario
 - Endpoint /: Página principal con información del proyecto
- **Gradio: Interfaz gráfica interactiva montada en /ui**
 - Pestaña 'Detección': Formulario de ingreso de datos
 - Pestaña 'Métricas del Modelo': Visualización de rendimiento
 - Pestaña 'Acerca de': Documentación del proyecto

3.5.2. Funcionalidades Principales

- **Detección de Anomalías:**
 - Formulario para ingresar características del hogar
 - Cálculo de consumo esperado en tiempo real
 - Comparación con consumo real (opcional)
 - Clasificación de nivel de anomalía (BAJA/MEDIA/ALTA)
 - Cálculo de costo estimado con tarifas actuales
- **Generación de Recomendaciones:**
 - Consejos personalizados según equipos presentes
 - Estimación de ahorro potencial en kWh/mes

- Priorización de acciones según impacto
- **Visualización de Métricas:**
 - Matriz de confusión del modelo
 - Curva ROC con área bajo la curva (AUC)
 - Gráfico de importancia de features
 - Métricas de rendimiento (Accuracy, Precision, Recall, F1-Score)

3.5.3. Diseño de Interfaz de Usuario

La interfaz utiliza los colores institucionales de la Universidad Internacional del Ecuador (UIDE):

- Azul corporativo (#003366, #0066cc): Headers y títulos
- Amarillo institucional (#ffcc00): Acentos y bordes
- Verde (#4caf50, #2e7d32): Indicadores positivos y ahorro
- Rojo (#f44336): Alertas de consumo alto
- Amarillo-alerta (#ffc107): Consumo medio

Figura 14

Interfaz de usuario – pestaña principal “Detección” del sistema Detector de Anomalías Energéticas

Detector de Anomalías Energéticas
Universidad Internacional del Ecuador

Maestría en Inteligencia Artificial
Despliegue de Modelos en la nube

Detección Historial Métricas del Modelo Comparativa Modelos Acerca de

Datos Generales

Región
Costa **Sierra** Oriente

Año
Selecciona el año del consumo: 2026

Mes
Selecciona el mes del consumo: Julio

Cédula del propietario (opcional)
Con tu cédula podrás ver tu historial de consumo.
Ingresa tu número de cédula (10 dígitos)

Número de habitantes que pasan promedio todo el día
4

Tipo de usuario
Residencial General Adulto Mayor (65+ años) Tarifa Dignidad

Beneficios disponibles:

- Adulto Mayor: 50% descuento en primeros 138 kWh.
- Tarifa Dignidad: \$0.04/kWh hasta 130 kWh (Costa) o 110 kWh (Sierra).
- Residencial General: Tarifa normal escalonada.
- Consumos actualizados con tecnologías modernas (2024-2025).

Equipos de Alto Consumo

Horas de uso de aire acondicionado por día
4

Minutos de uso de calefón/ducha por día
10

Electrodomésticos

☒ Refrigeradora

☐ Lavadora

☐ Microondas

☐ Computadora

☐ Aire acondicionado

☐ Calefón eléctrico

Televisores encendidos promedio
2

Consumo Real

Consumo según factura (kWh/mes) *OBLIGATORIO*
0

¿No tienes tu consumo a la mano? Escoge la región en Datos Generales.
Consulta tu consumo en el sitio oficial de tu distribuidora eléctrica.
El enlace se abre en una nueva pestaña para que puedas volver fácilmente.

Sierra - Distribuidora: EEQ o CNEL EP

Consultar consumo en EEQ (Pichincha)

Consultar consumo en EMELNORTE (Norte de la Sierra)

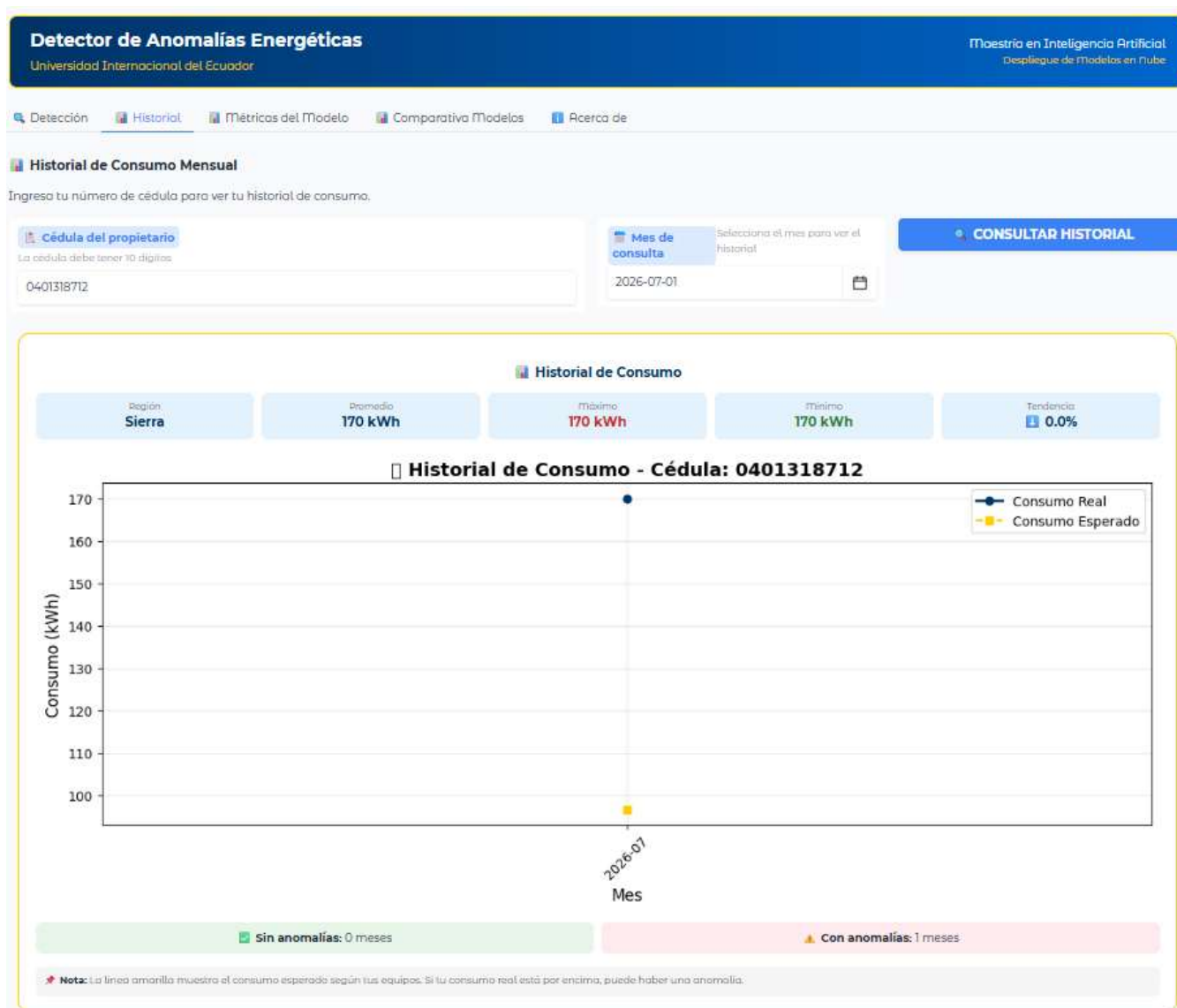
¡Necesitas tu Cédula (EEQ) o CUEN (CNEL) (aparece en tu factura)!

DETECTAR ANOMALÍA

*Nota: La interfaz web principal del sistema de detección de anomalías energéticas en su pestaña **Detección** muestra al usuario la información respecto a las características del hogar relacionadas a las variables: geográfica, demográficas, de consumo energético y de equipamiento eléctrico que caracterizan a cada hogar.*

Figura 15

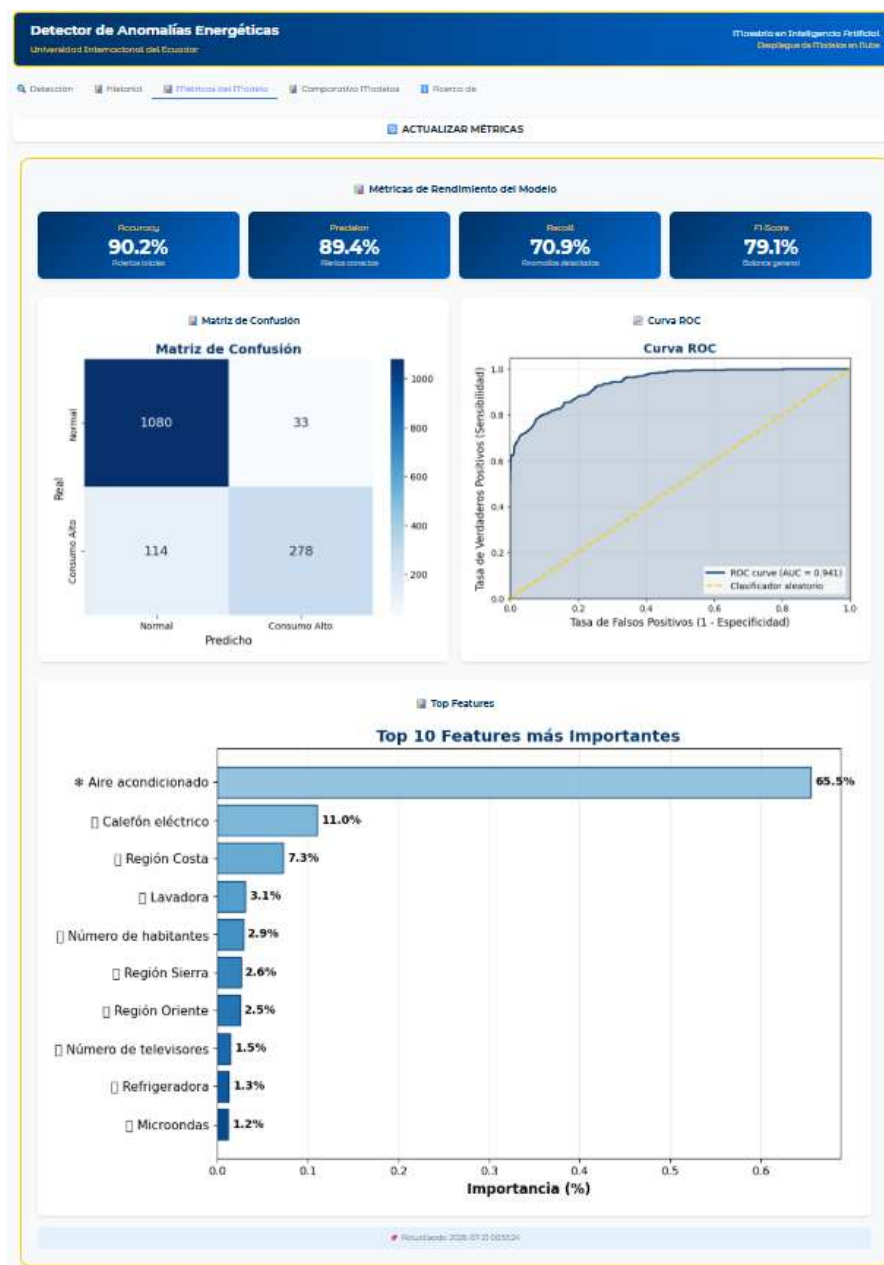
Interfaz de usuario – pestaña “Historial” del sistema Detector de Anomalías Energéticas



*Nota: La interfaz web del sistema de detección de anomalías energéticas en su pestaña **Historial** muestra al usuario información respecto al historial de consumo, conforme el usuario lo haya ido registrado mensualmente en el sistema de detección de anomalías. Esta información lo muestra tanto gráfica como textualmente.*

Figura 16

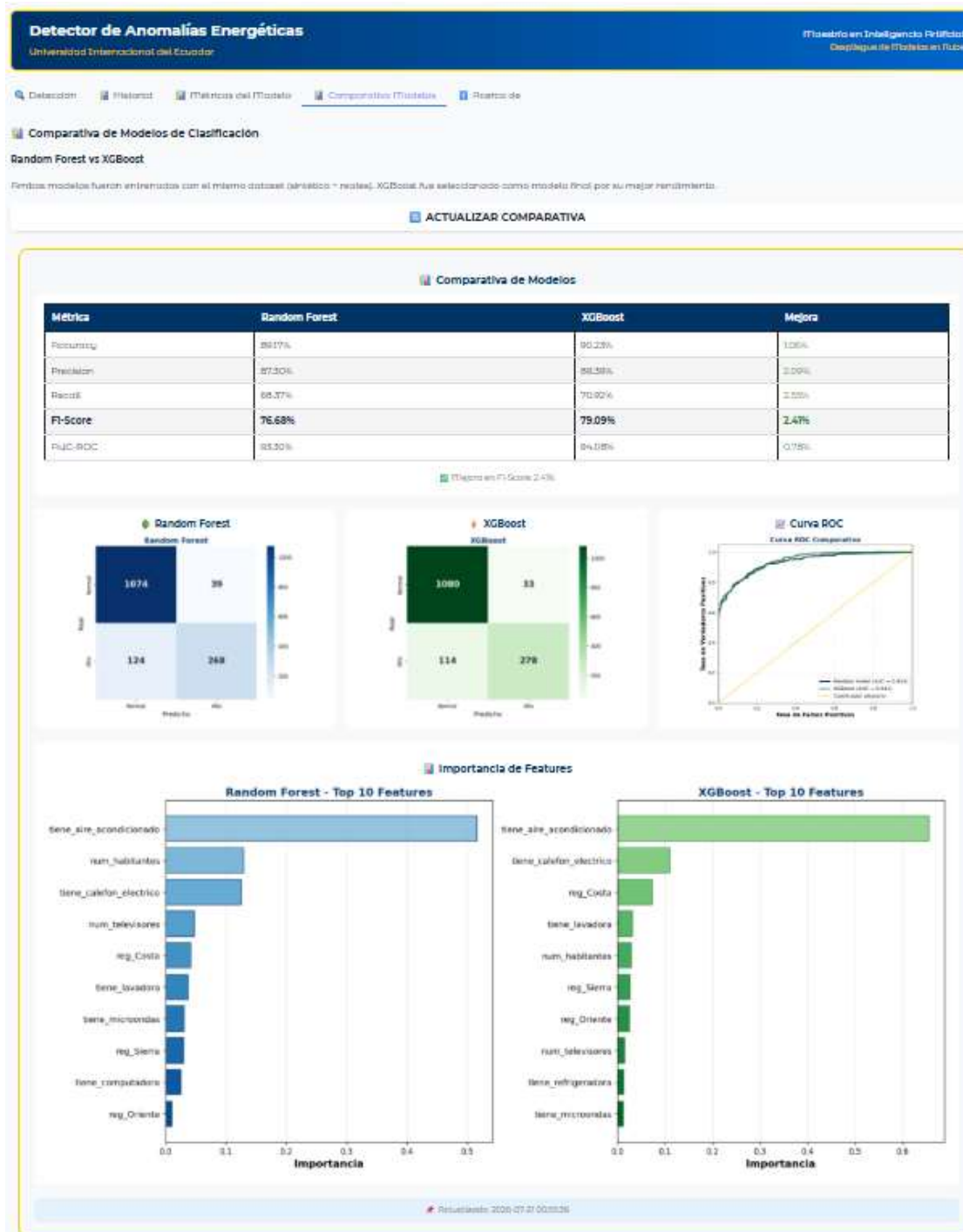
Interfaz de usuario – pestaña “Métricas del Modelo” del sistema Detector de Anomalías Energéticas



Nota: La interfaz web del sistema de detección de anomalías energéticas en su pestaña **Métricas del Modelo** muestra al usuario información asociada al rendimiento del modelo como el Accuracy, Precisión, Recall, F1-Score, Matriz de Confusión, curva ROC y las 10 features más importantes.

Figura 17

Interfaz de usuario – pestaña “Comparativa Modelos” del sistema Detector de Anomalías Energéticas



Nota: La interfaz web del sistema de detección de anomalías energéticas en su pestaña **Comparativa modelos** muestra al usuario información del modelo principal de este proyecto Random Forest con otro modelo XGBoost.

una perspectiva de lo que se realizó en relación al modelo de Random Forest y su relación con otros modelos.

CAPÍTULO 4:

ANÁLISIS DE RESULTADOS

4.1. Procesamiento de los Datasets de Referencia

Los resultados de la fase de procesamiento de los conjuntos de datos que se describen en el Capítulo 3 se muestran en esta sección. Se evalúa, por un lado, la conversión del conjunto de datos Household Power Consumption en perfiles de consumo horario y, por el otro, la obtención de los consumos por aparato electrodoméstico a partir del conjunto de datos UK-DALE. Los valores actualizados que se extraen forman la base para el cálculo implementado en el sistema.

4.1.1. Household Power Consumption

El procesamiento del conjunto de datos sobre el consumo de energía eléctrica en los hogares convirtió las lecturas de cada minuto del archivo original en perfiles horarios. A partir de los 2075259 registros iniciales, después de eliminar caracteres inválidos y valores nulos, se mantuvieron 2049280 datos válidos (98.7 % de retención). De estos se generaron 288 perfiles horarios (12 meses por 24 horas) con un consumo medio por hora de 1.2 kWh (entre el rango de 0.8 y el de 1.8 kWh). Los hábitos residenciales típicos se reflejan en los patrones detectados, que indican picos de consumo por la mañana (de 7:00 a 9:00) y la noche (de 19:00 a 22:00), y se emplearon como criterio formal para modelar el consumo doméstico en términos temporales.

4.1.2. UK-DALE y valores de referencia implementados

Se obtuvo, a partir del procesamiento del conjunto de datos UK-DALE y las eficiencias de los equipos modernos según el informe "NREL 2024 Buildings Technology Baseline" y Pecan Street Dataport, el consumo mensual promedio por electrodoméstico que utiliza el sistema como base para sus cálculos. Dado que estos valores están codificados en el módulo

principal de la aplicación (las constantes CONSUMOS_BASE y POTENCIAS de src/app.py), pueden ser verificados directamente en el repositorio.

Tabla 5

Consumos y potencias de referencia por equipo, implementados en el sistema

Equipo	Valor implementado	Cambio frente a modelos antiguos
Refrigeradora	35 kWh/mes	−22 % (tecnología inverter)
Lavadora	18 kWh/mes	−28 % (motores eficientes)
Televisor	10 kWh/mes por unidad	−33 % (tecnología LED)
Microondas	8 kWh/mes	Sin cambio significativo
Computadora	6 kWh/mes	Sin cambio significativo
Aire acondicionado	1.0 kW (potencia)	−33 % (tecnología inverter)
Calefón eléctrico	3.0 kW (potencia)	−14 % (mejor aislamiento)

Con base en estas referencias, el consumo anticipado de cada vivienda se calcula sumando los equipos que hay, empleando las horas de uso reportadas para los aparatos de alta potencia, el factor por persona ($0.7 + 0.1$ por habitante) y los factores regionales y mensuales expuestos en el tercer capítulo.

4.2. Construcción y Validación del Dataset Integrado

El dataset sintético congelado en su versión 1.0, que incluye 15 variables y 5,010 hogares, fue el resultado del proceso de integración (data fusion). La Tabla 2 presenta un resumen de sus estadísticas descriptivas, las cuales se obtuvieron a través de un análisis exploratorio reproducible.

Tabla 6

Estadísticas descriptivas del dataset integrado (versión congelada 1.0)

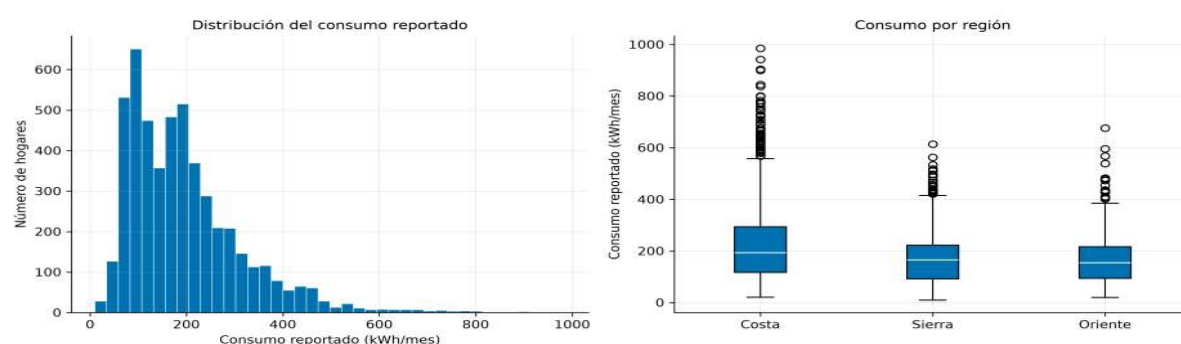
Indicador	Valor
Registros totales	5,010
Hogares Costa	2,247 (44.9 %)
Hogares Sierra	1,985 (39.6 %)
Hogares Oriente	778 (15.5 %)
Consumo reportado — media	195.0 kWh/mes
Consumo reportado — desv. estándar	118.6 kWh
Consumo reportado — rango	10.0 – 984.8 kWh
Registros con anomalía (etiqueta)	755 (15.1 %)

Hogares con aire acondicionado	24.3 %
Hogares con calefón eléctrico	49.5 %
Hogares con refrigeradora	96.3 %
Habitantes por hogar — media	3.57

Los porcentajes regionales (39.6 % Sierra, 44.9 % Costa y 15.5 % Oriente) reflejan con exactitud las proporciones de diseño (40 %, 45 % y 15 %). Según el procedimiento de generación delineado en el Capítulo 3, el 15.1% de los registros se clasifica como anómalo, mientras que la media del consumo reportado es de 195.0 kWh/mes y la desviación estándar es de 118.6 kWh.

Figura 19

Distribución del consumo reportado y comparación por región



Nota: El histograma muestra la asimetría positiva típica del consumo residencial; los diagramas de caja evidencian el mayor consumo mediano de la región Costa, coherente con el factor regional del 15 % y la mayor presencia de aire acondicionado.

4.2.2. Validación de Coherencia

La media de los datos del conjunto sobrepasa el consumo promedio a nivel nacional en viviendas, según lo informado por ARCONEL (alrededor de 138 kWh/mes en el año 2023). Esta distinción es intencional y se debe a la composición del conjunto de datos: para que el modelo tuviera bastantes ejemplos de consumo alto y de irregularidades, la generación sintética presenta un exceso de hogares con aparatos que consumen mucho (49.5 % con calefón eléctrico y 24.3 % con aire acondicionado, en comparación con penetraciones nacionales más bajas). Por este motivo, la validación del conjunto de datos no se realiza mediante una comparación directa con el promedio nacional, sino a través de tres verificaciones internas que pueden ser repetidas.

- Primero: la fidelidad del diseño en relación con la distribución regional (sección 4.2.1).
- Segunda: la coherencia entre el consumo esperado que se documenta en el conjunto de datos y el que produce la fórmula de la aplicación al recalcularlo con 1 500 hogares se obtiene una correlación de Pearson $r = 0.877$; la parte no explicada proviene del margen aleatorio del generador de $\pm 10\%$ y del tiempo de uso declarado de los equipos de alta potencia.
- Tercera: la uniformidad del etiquetado de anomalías, todos los registros anómalos se sitúan sobre o por encima de la línea $y = 1.3 \cdot x$ del plano esperado-reportado, con razones observadas entre 1.300 y 1.974, lo que confirma la aplicación consistente de la regla de generación y, a la vez, la ausencia de solapamiento entre clases que se analiza en la sección 4.4.7.

Figura 20

Coherencia entre la fórmula de la aplicación y el dataset sintético ($n = 1\,500$; $r = 0.877$)

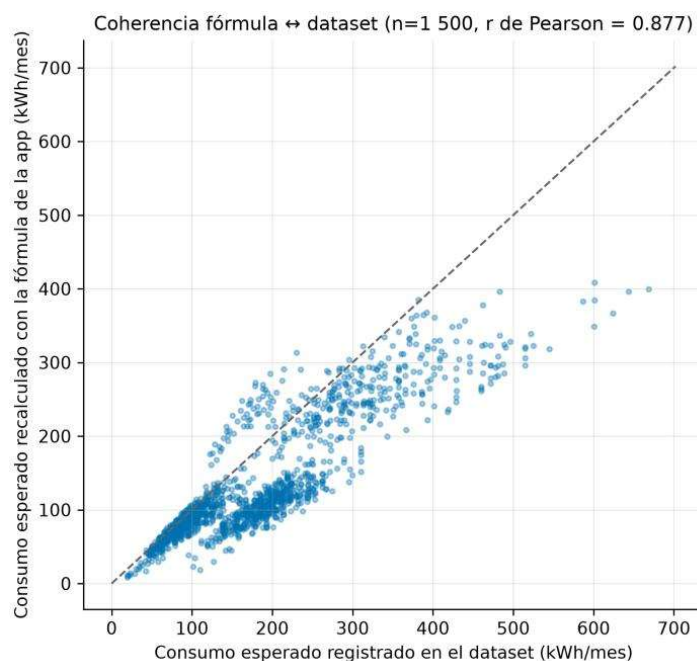
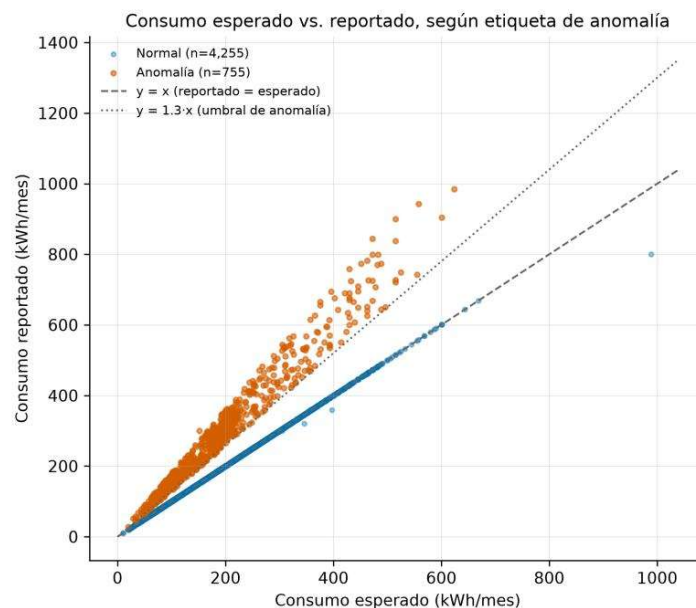


Figura 21

Consumo esperado frente a consumo reportado según la etiqueta de anomalía



Nota: Cada punto es un hogar del dataset congelado (azul: normal; naranja: anómalo). Las anomalías se sitúan íntegramente sobre el umbral $1.3 \cdot x$.

4.3. Diseño Experimental

Esta sección establece las condiciones de obtención de los resultados antes de informarlos, incluyendo el protocolo de evaluación, los algoritmos que se compararon y las dos formulaciones del problema. Su objetivo es asegurar que cada número del capítulo sea reproducible e interpretable.

Criterios de selección de los algoritmos comparados. La selección no persigue un barrido exhaustivo, que sería inabarcable y poco informativo, sino una cobertura razonada guiada por tres criterios:

- El primero es la cobertura por mecanismo de aprendizaje: se incluye al menos un representante de cada familia con mecanismo de inducción distinto modelos lineales, árboles individuales, métodos basados en instancias, métodos de margen máximo con núcleo, redes neuronales y ensamblados por agregación y por refuerzo secuencial, de modo que la conclusión no dependa de un único paradigma.
- El segundo es la capacidad de representación como variable del diseño: los

algoritmos se ordenan por la riqueza de la familia de funciones que pueden expresar, lo que permite diagnosticar la naturaleza del problema, ya que un problema con estructura genuina beneficia de forma gradual a los modelos de mayor capacidad mientras que uno cuya etiqueta se deriva de una operación aritmética simple es resuelto igualmente bien por un modelo lineal.

- El tercero es la pertinencia en el dominio y en la literatura, privilegiando algoritmos con antecedentes documentados en detección de anomalías de consumo eléctrico sobre datos tabulares (Jokar, Arianpoo y Leung, 2016; Zheng et al., 2018).

Con esos criterios se incluyeron: regresión logística, que establece la referencia mínima y permite detectar problemas trivialmente separables; árbol de decisión (Breiman, Random Forests, 2001) que aísla el aporte del ensamblado al ser el componente base de Random Forest sin agregación; K vecinos más cercanos (Cover y Hart, 1967) que no asume forma de frontera y resulta sensible a la geometría local; máquina de vectores de soporte con núcleo radial (Cortes y Vapnik, 1995), que aporta alta capacidad mediante un mecanismo distinto al de los ensamblados; perceptrón multicapa (Hornik et al., 1989), aproximador universal y por ello el más expuesto a memorizar regularidades artificiales; Random Forest (Breiman et al., 1984), modelo principal por su interpretabilidad directa; XGBoost (Chen y Guestrin, 2016) y LightGBM (Ke et al., 2017), que contrastan dos estrategias distintas de refuerzo secuencial; e Isolation Forest (Liu et al., 2008), único método que no accede a las etiquetas y que por ello cumple la función de control del diseño.

Se descartaron deliberadamente cuatro alternativas. Las arquitecturas profundas de tipo convolucional o recurrente requieren series temporales con resolución intradiaria, y el conjunto contiene un registro mensual agregado por hogar sin secuencia temporal. Los métodos específicos de series temporales quedan excluidos por la misma razón, agravada por el hecho de que prácticamente todos los registros corresponden a un mismo mes. Naive Bayes se descarta porque su supuesto de independencia condicional es insostenible aquí: la presencia de aire acondicionado, el número de habitantes y la región están correlacionados por construcción. Los métodos de ensamblado apilado añaden una capa de complejidad y de riesgo de fuga en la validación que no se justifica mientras el diagnóstico del conjunto de datos siga abierto.

La elección de métricas está condicionada por el desbalance de clases. Con una prevalencia de la clase positiva cercana al 15-20 %, un clasificador que asignara todas las observaciones a la clase mayoritaria obtendría entre el 80 % y el 85 % de exactitud sin detectar una sola anomalía. Por ello se adopta el F1-Score como criterio principal de comparación y se reporta el área bajo la curva de precisión-exhaustividad, más informativa que el AUC-ROC en conjuntos desbalanceados porque la abundancia de verdaderos negativos infla este último (Saito y Rehmsmeier, 2015). La exactitud se reporta por comparabilidad con la literatura, pero no se emplea como criterio de decisión.

Las diferencias entre modelos se contrastan con la prueba de McNemar en su formulación exacta, recomendada para comparar dos clasificadores evaluados sobre el mismo conjunto de prueba porque considera únicamente los casos en que ambos discrepan y controla el error de tipo I mejor que las alternativas basadas en remuestreo repetido (Dietterich, 1998). Dado que se realizan siete comparaciones contra el mismo modelo de referencia, se aplica la corrección secuencial de Holm-Bonferroni, que controla la tasa de error por familia sin la pérdida de potencia de la corrección de Bonferroni simple; se reportan tanto el p-valor original como el ajustado, y la conclusión se toma sobre el segundo.

4.3.1. Protocolo de Evaluación

La evaluación se llevó a cabo utilizando validación cruzada estratificada de cinco particiones (reportando la media $\pm$ desviación estándar por cada métrica), que fue complementada con una partición de retención del 70/30, también estratificada, para la creación de curvas, matrices de confusión y pruebas de significancia. La semilla 42 y el conjunto de datos congelado v1.0 (hash SHA-256 registrado) son utilizados por todos los experimentos. Para contextualizar el rendimiento de Random Forest, se evaluaron, además del modelo del proyecto, siete algoritmos de familias diferentes: KNN, regresión logística, árbol de decisión, SVM con núcleo RBF, perceptrón multicapa (MLP), XGBoost y LightGBM. Isolation Forest fue utilizado como referencia no supervisada. Los tres modelos arbóreos emplean los mismos hiperparámetros que el sistema en producción; por lo tanto, la comparación ilustra cómo actúa el prototipo real.

4.3.2. Formulaciones del problema

Se examinaron dos formulaciones que se complementan entre sí. El Experimento A reproduce la metodología utilizada en el sistema implementado: consiste en categorizar si un hogar tiene

un alto consumo (más del percentil 80, lo que corresponde al 20 % de los positivos) solamente con base en el perfil de ese hogar. El Experimento B se refiere a la elaboración del marco teórico: identificar la anomalía etiquetada en el conjunto de datos (15.1 % de positivos) basándose en el perfil y el consumo notificado por el usuario, que es precisamente lo que se puede observar en el flujo real de la aplicación. Comparar ambas permite distinguir cuánto del rendimiento se origina en el perfil del hogar y cuánto proviene del consumo observado

A estas dos formulaciones se añaden dos escenarios adicionales que completan una ablación sobre la información concedida al modelo, manteniendo constantes la partición, la semilla, los hiperparámetros y las métricas. El Escenario C entrega al modelo el perfil, el consumo facturado y el consumo esperado, es decir, ambos términos de la operación con la que se etiquetó el conjunto; constituye la cota superior y mide qué tan recuperable es la regla, no qué tan detectable es una anomalía. El Escenario D aplica un detector no supervisado que no accede a las etiquetas ni, por tanto, al criterio de generación, y constituye la cota inferior realista. La comparación entre los cuatro escenarios permite atribuir las diferencias de rendimiento a la información disponible y no a otras causas, y se apoya en el concepto de fuga de información establecido en la literatura de minería de datos (Kaufman, Rosset y Perlich, 2012), introducida aquí de forma controlada y graduada para medir su efecto en lugar de tratarla únicamente como un defecto a evitar.

4.4. Análisis del Desempeño del Modelo

Esta sección presenta los hallazgos del protocolo anterior en siete perspectivas: el análisis comparativo de algoritmos, la matriz de confusión, las curvas ROC y precisión-recall, la significancia estadística de las diferencias, la calibración del umbral de decisión, la curva de aprendizaje y el análisis de sobreajuste y límites de la generalización. Todas las cifras provienen de una única ejecución reproducible del protocolo descrito en la sección 4.3 (semilla 42, partición 70/30 estratificada, umbral del percentil 80 calculado únicamente sobre el conjunto de entrenamiento) sobre el conjunto de 5 080 registros.

4.4.1. Análisis de comparativo de algoritmos

Tabla 7

Validación cruzada estratificada 5-fold - Experimento A: consumo alto (media $\pm$ desviación estándar)

Modelo	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Regresión	0.8819 $\pm$	0.6523 $\pm$	0.8790 $\pm$	0.7487 $\pm$	0.9520 $\pm$
Logística	0.0113	0.0244	0.0203	0.0215	0.0099
Árbol de	0.8830 $\pm$	0.6765 $\pm$	0.8086 $\pm$	0.7351 $\pm$	0.9008 $\pm$
Decisión	0.0161	0.0451	0.0360	0.0262	0.0184
KNN	0.9334 $\pm$	0.8900 $\pm$	0.7609 $\pm$	0.8202 $\pm$	0.9309 $\pm$
	0.0070	0.0214	0.0225	0.0196	0.0036
SVM (RBF)	0.9142 $\pm$	0.7593 $\pm$	0.8368 $\pm$	0.7960 $\pm$	0.9584 $\pm$
	0.0100	0.0265	0.0236	0.0231	0.0058
MLP	0.9364 $\pm$	0.8905 $\pm$	0.7792 $\pm$	0.8303 $\pm$	0.9492 $\pm$
	0.0092	0.0329	0.0391	0.0267	0.0104
Random Forest	0.8945 $\pm$	0.6978 $\pm$	0.8410 $\pm$	0.7618 $\pm$	0.9568 $\pm$
	0.0115	0.0349	0.0279	0.0186	0.0074
XGBoost	0.9052 $\pm$	0.7314 $\pm$	0.8326 $\pm$	0.7784 $\pm$	0.9554 $\pm$
	0.0070	0.0201	0.0232	0.0156	0.0103
LightGBM	0.8985 $\pm$	0.7066 $\pm$	0.8466 $\pm$	0.7694 $\pm$	0.9564 $\pm$
	0.0082	0.0278	0.0354	0.0158	0.0085

Tabla 8

Validación cruzada estratificada 5-fold — Experimento B: detección de anomalía (media $\pm$ desviación estándar)

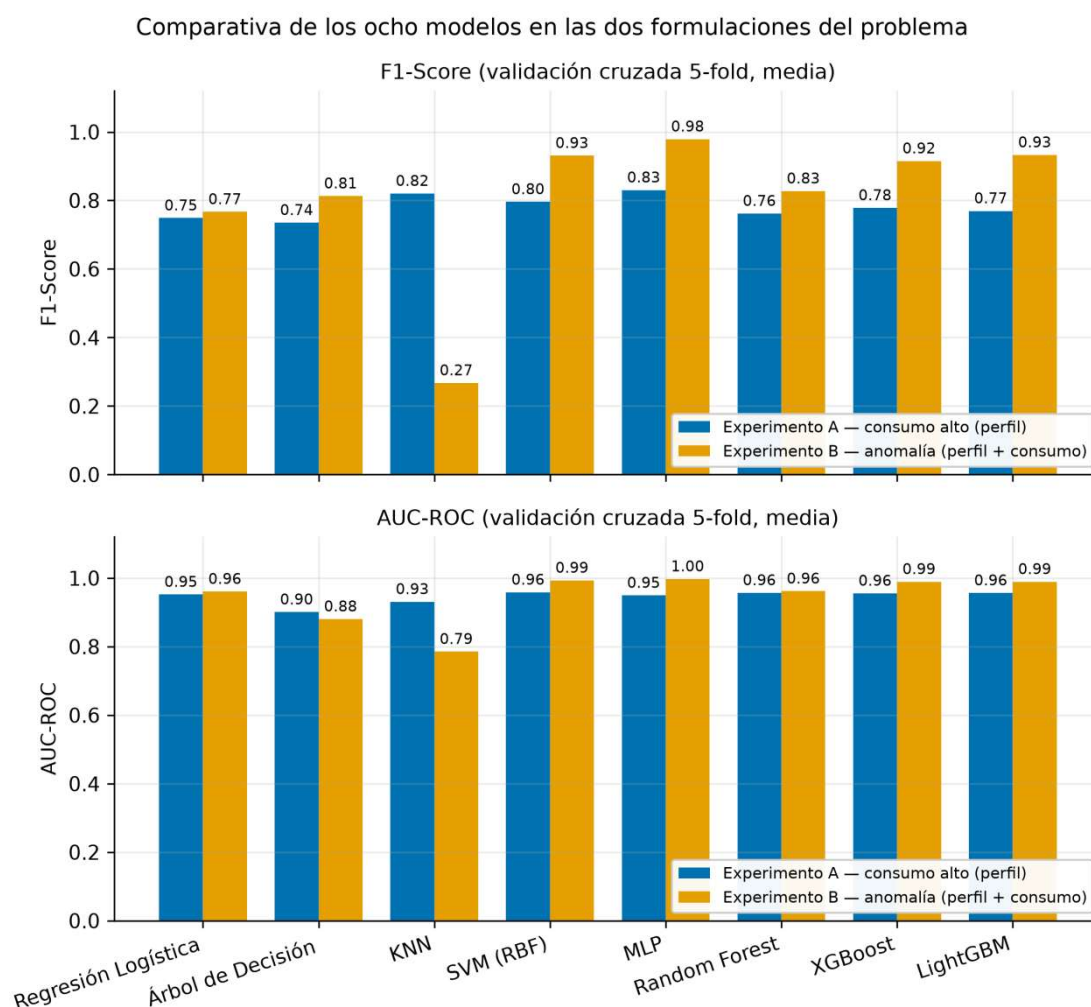
Modelo	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Regresión	0.9168 $\pm$	0.6665 $\pm$	0.9048 $\pm$	0.7669 $\pm$	0.9607 $\pm$
Logística	0.0126	0.0404	0.0181	0.0294	0.0089
Árbol de	0.9460 $\pm$	0.8537 $\pm$	0.7779 $\pm$	0.8129 $\pm$	0.8805 $\pm$
Decisión	0.0048	0.0375	0.0301	0.0147	0.0132
KNN	0.8695 $\pm$	0.8701 $\pm$	0.1587 $\pm$	0.2674 $\pm$	0.7853 $\pm$
	0.0057	0.0928	0.0286	0.0429	0.0222
SVM (RBF)	0.9789 $\pm$	0.9225 $\pm$	0.9403 $\pm$	0.9308 $\pm$	0.9931 $\pm$
	0.0048	0.0301	0.0217	0.0153	0.0051

MLP	0.9938 $\pm$ 0.0026	0.9962 $\pm$ 0.0046	0.9627 $\pm$ 0.0204	0.9790 $\pm$ 0.0090	0.9973 $\pm$ 0.0036
Random Forest	0.9519 $\pm$ 0.0060	0.8991 $\pm$ 0.0123	0.7668 $\pm$ 0.0345	0.8274 $\pm$ 0.0241	0.9617 $\pm$ 0.0121
XGBoost	0.9755 $\pm$ 0.0075	0.9573 $\pm$ 0.0239	0.8768 $\pm$ 0.0353	0.9150 $\pm$ 0.0270	0.9893 $\pm$ 0.0062
LightGBM	0.9806 $\pm$ 0.0066	0.9715 $\pm$ 0.0155	0.8974 $\pm$ 0.0330	0.9328 $\pm$ 0.0241	0.9893 $\pm$ 0.0072

En el Experimento A los ocho modelos se agrupan en una banda estrecha: F1 entre 0.7351 y 0.8303 en validación cruzada y entre 0.7115 y 0.7883 sobre la partición de retención; AUC-ROC entre 0.8753 y 0.9411. Random Forest obtiene $F1 = 0.7618 \pm 0.0186$ y $AUC-ROC = 0.9568 \pm 0.0074$ en validación cruzada, y $F1 = 0.7377$ con una exactitud del 88.85 % en retención. La estrechez de la banda es el resultado más informativo de la tabla: cuando algoritmos cuya capacidad de representación va de una frontera lineal a un aproximador universal convergen al mismo nivel, el límite proviene de la información contenida en las variables y no de la elección del modelo. En el Experimento B el comportamiento cambia por completo: los modelos de mayor capacidad se separan de los de capacidad baja y el MLP pasa de $F1 = 0.7883$ a 0.9802, mientras la regresión logística apenas mejora (de 0.7117 a 0.7626) porque la relación que debe aprenderse es un cociente entre dos cantidades y no una combinación lineal de ellas. Random Forest alcanza $F1 = 0.8274 \pm 0.0241$ y $AUC-ROC = 0.9617 \pm 0.0121$ en validación cruzada, con una exactitud del 95.67 % en retención. El rendimiento casi perfecto del MLP debe leerse con la sección 4.4.7: en un tercer escenario que entrega al modelo también el consumo esperado, la regresión logística iguala al MLP ($F1 = 0.9935$ frente a 0.9934), lo que demuestra que con ambos términos de la regla de etiquetado disponibles el problema es linealmente separable y no requiere capacidad de representación alguna. El detector no supervisado Isolation Forest, sin acceso a las etiquetas, obtiene $F1 = 0.2141$ y $AUC-ROC = 0.5931$ en la misma partición, próximo al azar; la distancia entre ese valor y el del escenario con acceso completo a la regla acota el rango dentro del cual se sitúa cualquier resultado supervisado sobre este conjunto.

Figura 22

Comparativa visual de los ocho modelos en las dos formulaciones del problema (F1 y AUC-ROC, validación cruzada 5-fold)



Nota: La formulación del Experimento B mejora el desempeño de casi todos los modelos; la excepción de KNN ($F1 = 0.2941$ frente a 0.7706 en el Experimento A) ilustra la sensibilidad de los métodos basados en distancia cuando una variable continua de escala mucho mayor domina el espacio de características: los vecinos dejan de ser hogares de perfil similar.

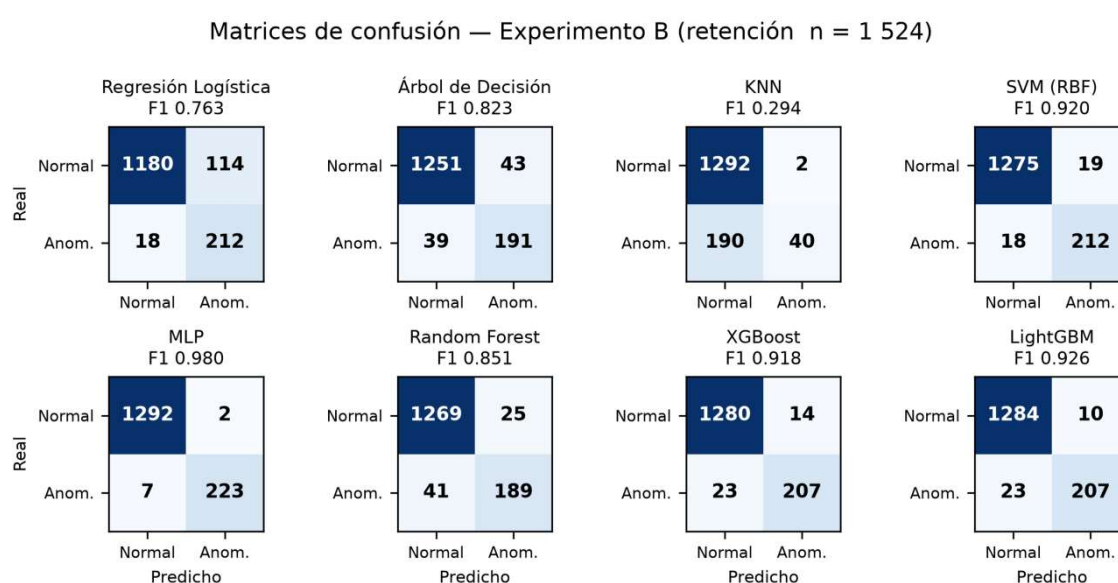
4.4.2. Matriz de Confusión

En la partición de retención (1 524 hogares el 30 % del conjunto) la matriz de confusión de Random Forest en el Experimento A registró 1 115 verdaderos negativos 239 verdaderos positivos 04 falsos positivos y 66 falsos negativos (exactitud del 88.85 % precisión de 0.6968 y recall de 0.7836). En el Experimento B se obtuvieron 1 269 verdaderos negativos 25 falsos

positivos 189 verdaderos positivos y 41 falsos negativos (exactitud del 95.67 % precisión de 0.8832 y recall de 0.8217). El bajo número de falsos positivos del Experimento B reduce las alertas injustificadas; los 41 falsos negativos se abordan mediante la calibración del umbral de decisión de la sección 4.4.5 que permite desplazar el compromiso entre precisión y sensibilidad según la política operativa.

Figura 23

Matrices de confusión de los ocho modelos — Experimento B (partición de retención, $n = 1\,524$)

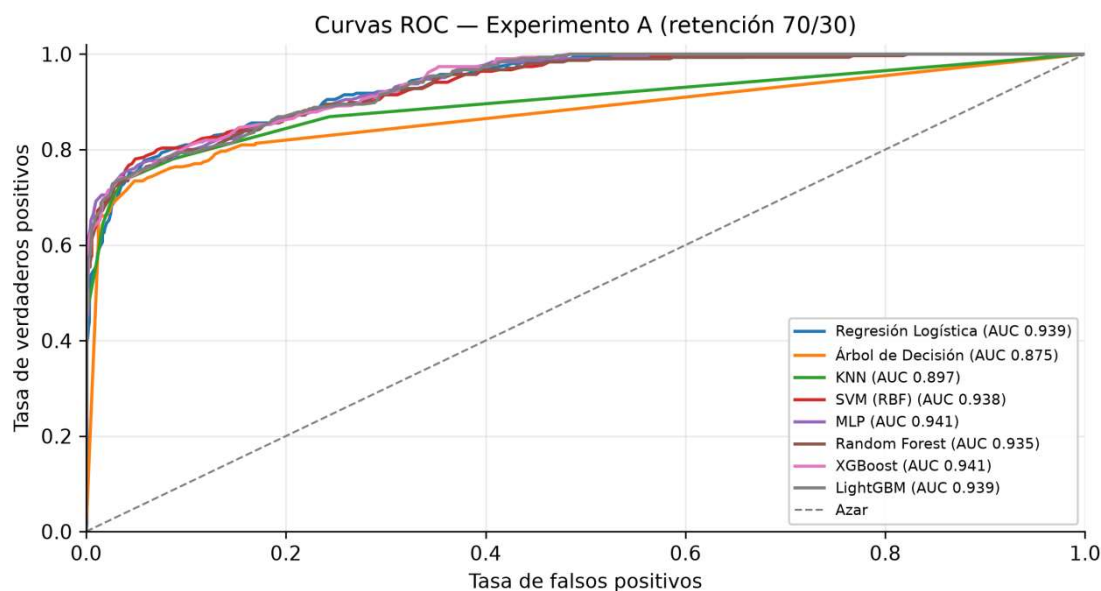


4.4.3. Curvas ROC y precisión-recall

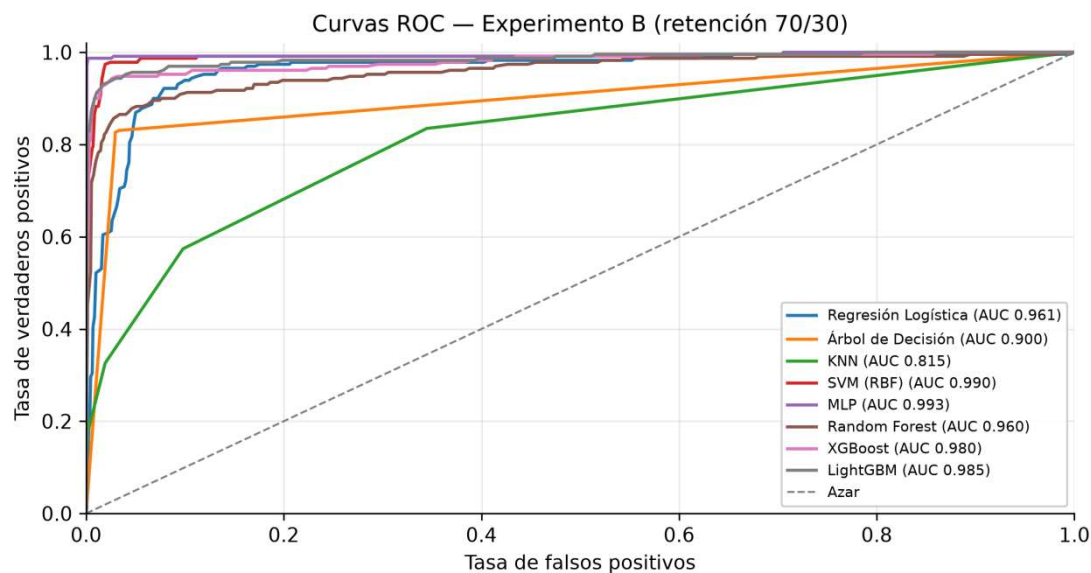
La observación anterior se verifica en las curvas ROC de los ocho modelos: en el Experimento A las áreas bajo la curva oscilan entre 0.8753 y 0.9411 (0.9350 para Random Forest); en el Experimento B, entre 0.8149 y 0.9933 (0.9595 para Random Forest). Conviene no sobreinterpretar valores altos de AUC-ROC en conjuntos desbalanceados: la abundancia de verdaderos negativos los infla, motivo por el cual la sección siguiente incorpora las curvas de precisión-recall.

Figura 24

Curvas ROC de los ocho modelos — Experimento A (holdout 70/30)

**Figura 25**

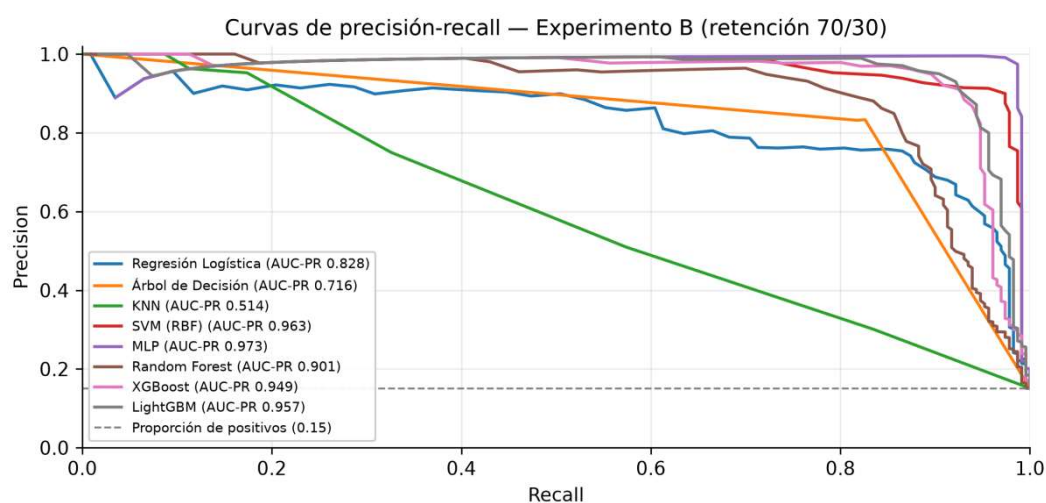
Curvas ROC de los ocho modelos — Experimento B (holdout 70/30)



Como las clases están desbalanceadas (15.1 % de positivos en el Experimento B), se añaden las curvas de precisión-recall, más informativas en este caso porque no premian los aciertos sobre la clase mayoritaria (Saito y Rehmsmeier, 2015). El área bajo la curva de precisión-recall de Random Forest es 0.9007, frente a la línea base de 0.15 que obtendría un clasificador sin información; el MLP alcanza 0.9728 y KNN cae a 0.5139.

Figura 26

Curvas de precisión-recall de los ocho modelos — Experimento B (holdout 70/30)



Nota: La línea discontinua indica la proporción de positivos (0.15), rendimiento de un clasificador sin información.

4.4.4. Significancia estadística de las diferencias

Se utilizó la prueba exacta de McNemar para determinar si las discrepancias entre Random Forest y cada uno de los otros algoritmos son estadísticamente significativas, sobre las predicciones de la partición de retención ($\alpha = 0.05$). Al realizarse siete comparaciones contra el mismo modelo de referencia, los p-valores se ajustan con la corrección secuencial de Holm-Bonferroni y la decisión se toma sobre el valor ajustado; las tablas reportan ambos.

Tabla 9

Prueba exacta de McNemar: Random Forest frente a cada competidor — Experimento A

Comparación	b (solo acierta el otro)	c (solo acierta RF)	p-valor exacto / ajustado (Holm)	¿Diferencia significativa ($\alpha = 0.05$)?
Random Forest vs Regresión Logística	53	88	0.0040 / 0.0202	Sí — favorece a RF
Random Forest vs Árbol de Decisión	16	35	0.0110 / 0.0329	Sí — favorece a RF
Random Forest vs KNN	85	40	0.0001 / 0.0004	Sí — favorece a KNN

Random Forest vs SVM (RBF)	77	46	0.0066 / 0.0264	Sí — favorece a SVM (RBF)
Random Forest vs MLP	81	27	0.0000 / 0.0000	Sí — favorece a MLP
Random Forest vs XGBoost	38	23	0.0722 / 0.0722	No
Random Forest vs LightGBM	25	11	0.0288 / 0.0576	No

Tabla 10

Prueba exacta de McNemar: Random Forest frente a cada competidor — Experimento B

Comparación	b (solo acierta el otro)	c (solo acierta RF)	p-valor exacto / ajustado (Holm)	¿Diferencia significativa ($\alpha = 0.05$)?
Random Forest vs Regresión Logística	48	114	0.0000 / 0.0000	Sí — favorece a RF
Random Forest vs Árbol de Decisión	32	48	0.0929 / 0.0929	No
Random Forest vs KNN	25	151	0.0000 / 0.0000	Sí — favorece a RF
Random Forest vs SVM (RBF)	50	21	0.0008 / 0.0015	Sí — favorece a SVM (RBF)
Random Forest vs MLP	57	0	0.0000 / 0.0000	Sí — favorece a MLP
Random Forest vs XGBoost	40	11	0.0001 / 0.0002	Sí — favorece a XGBoost
Random Forest vs LightGBM	39	6	0.0000 / 0.0000	Sí — favorece a LightGBM

En el Experimento A, tras la corrección de Holm, Random Forest no se distingue significativamente de XGBoost (p ajustado = 0.0722) ni de LightGBM (p ajustado = 0.0576), que son los competidores con mejor rendimiento nominal; las diferencias con el resto sí resultan significativas, pero conviene no confundir significancia estadística con relevancia práctica: con 1 524 registros de retención la prueba tiene potencia para detectar diferencias que en F1 no llegan a siete centésimas y carecen de consecuencia operativa. En el Experimento B el cuadro cambia: SVM (RBF), MLP, XGBoost y LightGBM superan a Random Forest con diferencias significativas, mientras que Random Forest supera significativamente a Regresión Logística, KNN. Que el ordenamiento entre algoritmos cambie por completo entre escenarios confirma que no es una propiedad estable de los

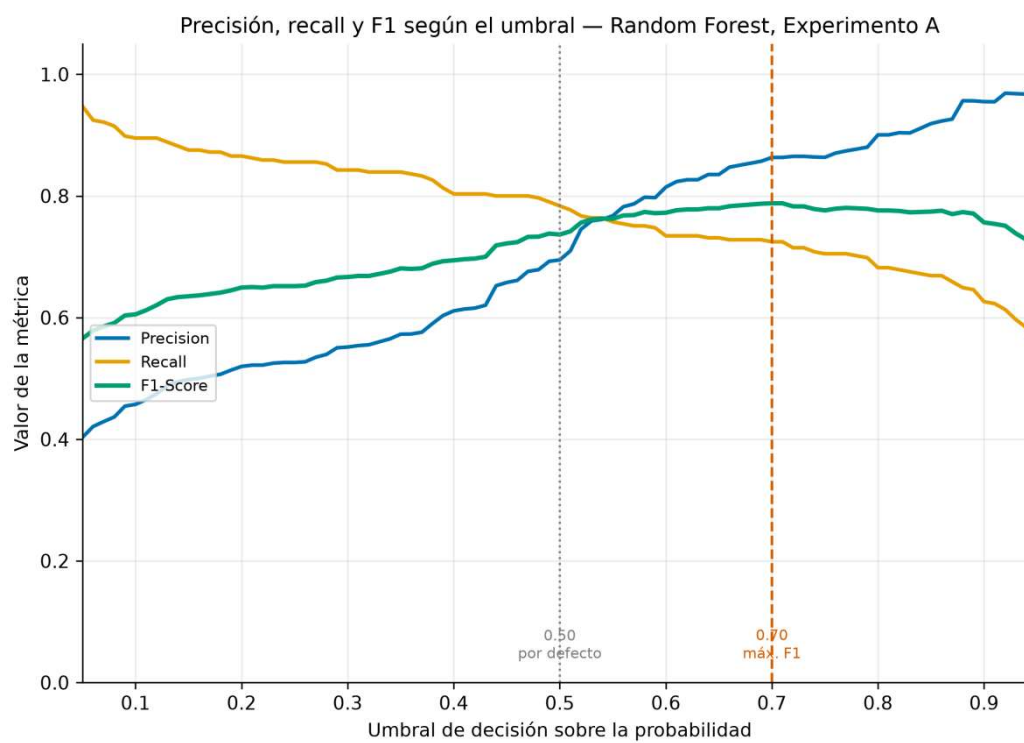
modelos sino del nivel de información concedido. La lectura conjunta sostiene la elección de Random Forest sobre criterios distintos de la superioridad numérica: en la formulación desplegada su diferencia con los mejores no es distinguible del azar del muestreo, ofrece interpretabilidad directa mediante la importancia de variables, y como muestra la sección 4.4.7 degrada menos que los modelos de mayor capacidad cuando se introduce dispersión realista en los datos.

4.4.5. Calibración del umbral de decisión

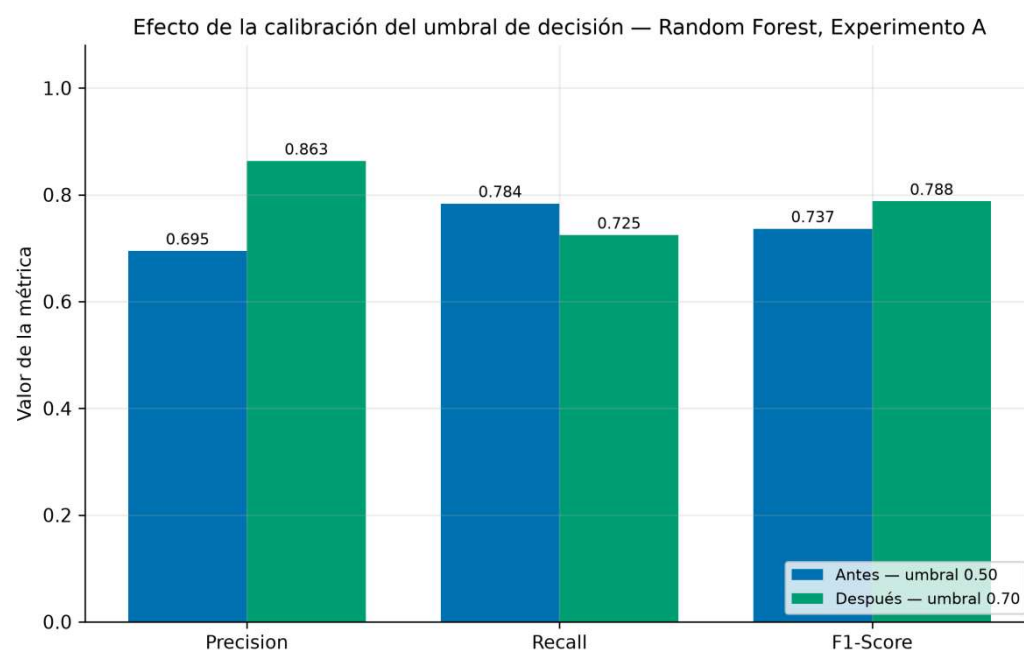
Se examinó el efecto del umbral de decisión sobre la probabilidad que emite el modelo, dado que en la detección de anomalías energéticas el costo de un caso no detectado y el de una alerta injustificada no son simétricos y dependen de la política operativa. El modelo del Experimento A con el umbral por defecto (0.50) obtiene un recall de 0.7836 y una precisión de 0.6948 ($F1 = 0.7365$); al desplazar el umbral a 0.70, que maximiza el F1 sobre la retención, la precisión sube a 0.8633 y el recall baja a 0.7246 ($F1 = 0.7879$). Esta calibración es una mejora inmediata que permite ajustar el compromiso entre precisión y sensibilidad sin reentrenar; el umbral óptimo se estima sobre la misma partición de retención, por lo que su valor exacto debe considerarse orientativo.

Figura 27

Precisión, recall y F1 en función del umbral de decisión — Random Forest, Experimento A

**Figura 28**

Antes y después de la calibración del umbral de decisión (0.50 → 0.81)



Nota: La calibración eleva la precisión de 0.6948 a 0.8633 y el F1 de 0.7365 a 0.7879, cediendo 5.9 puntos de recall; el balance es ajustable según la política operativa.

4.4.6. Curva de aprendizaje y mejora continua

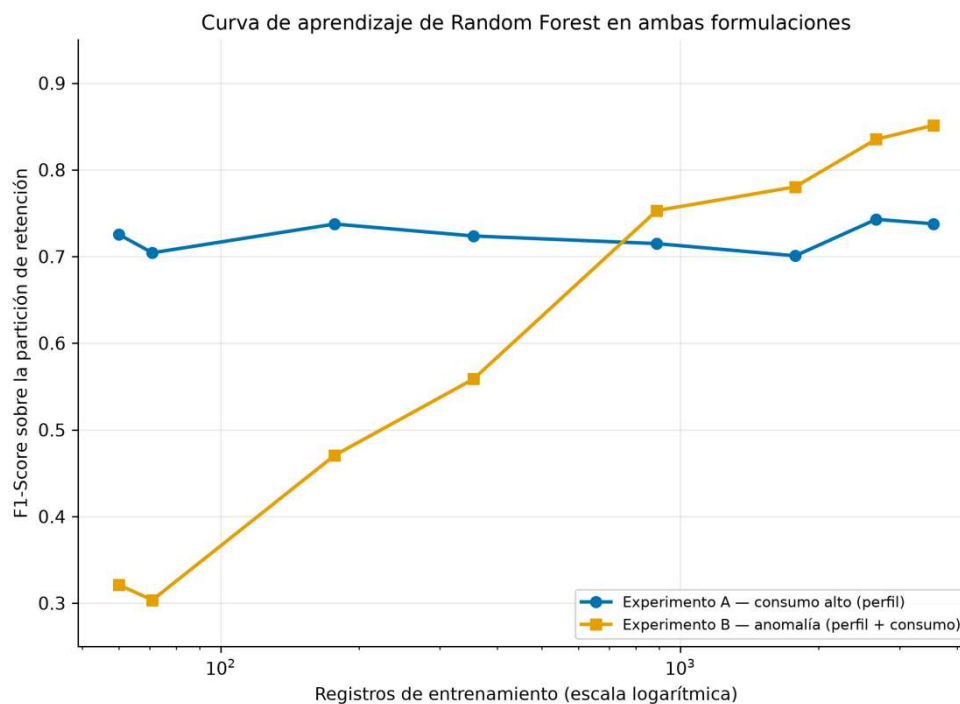
La curva de aprendizaje del Experimento A está saturada desde el primer punto de medición: con 60 registros de entrenamiento el modelo alcanza un F1 de 0.7255 y con 3 556 obtiene 0.7377 sin progresión monótona entre ambos extremos (0.7375 con 177 registros y 0.7008 con 1 778). La relación entre el perfil de equipamiento y el nivel de consumo se agota con muy poca evidencia.

Este resultado tiene una consecuencia práctica directa sobre la estrategia de mejora del sistema: ampliar el conjunto sintético no elevaría su desempeño, por lo que el esfuerzo debe dirigirse a incorporar variables nuevas —temperatura medida en primer lugar— y datos de campo, y no a generar más registros con el mismo procedimiento.

El Experimento B presenta la forma opuesta, con un ascenso sostenido de 0.3211 a 0.8514 que no muestra señales de haber alcanzado su límite. La interpretación es que la regla de etiquetado no se aprende de forma instantánea pero sí de forma completa: cada registro adicional refina la reconstrucción de la función generadora y, al ser esta determinística, no existe un techo de ruido que detenga el ascenso. El contraste entre ambas curvas resume el diagnóstico del conjunto de datos: lo que puede aprenderse del comportamiento del hogar se aprende con sesenta ejemplos, y lo que sigue mejorando con más datos es la reconstrucción de una fórmula.

Figura 29

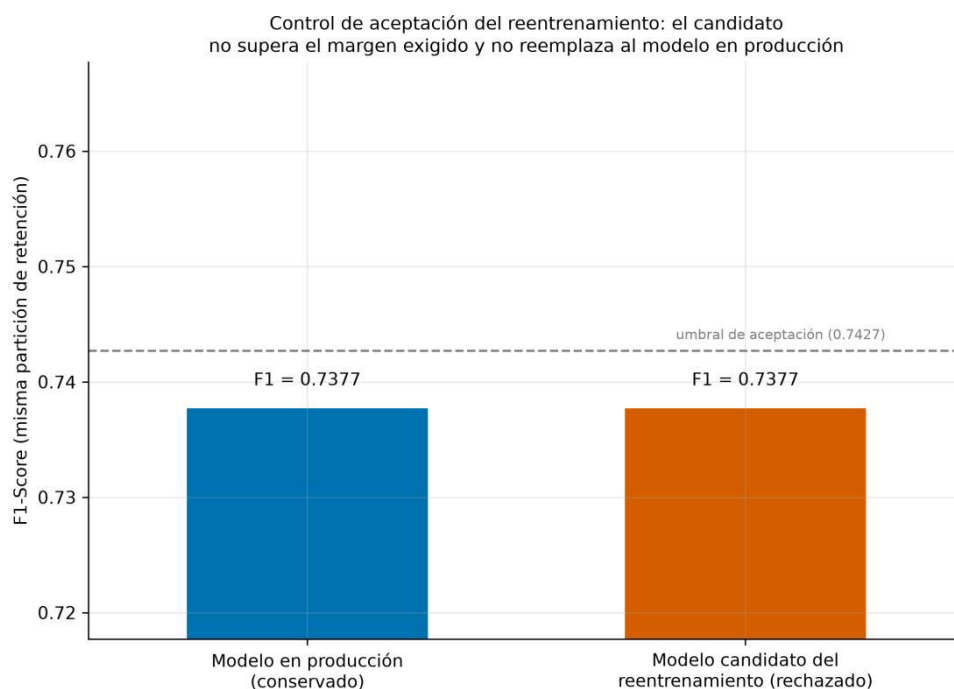
Curva de aprendizaje de Random Forest en ambas formulaciones (F1 sobre la partición de retención)



Para comprobar el ciclo en la práctica se ejecutó el reentrenamiento en modo de prueba sobre el conjunto compuesto: 5 010 registros sintéticos y 70 interacciones aportadas por usuarios reales durante el despliegue, equivalentes al 1.38 % del total. El control de aceptación evaluó al candidato y al modelo vigente sobre la misma partición de retención (F1 0.7377 frente a 0.7377) y rechazó al candidato por no superar el margen mínimo exigido de 0.005 en F1, conservando el modelo en producción. El mecanismo funciona según lo previsto; el volumen de datos reales acumulado hasta la fecha no permite todavía que un reentrenamiento aporte una mejora medible, lo que constituye una constatación del propio experimento y no una limitación del procedimiento.

Figura 30

Control de aceptación del reentrenamiento: comparación entre el modelo en producción y el candidato rechazado (prueba en seco)



4.4.7. Análisis de sobreajuste y límites de la generalización

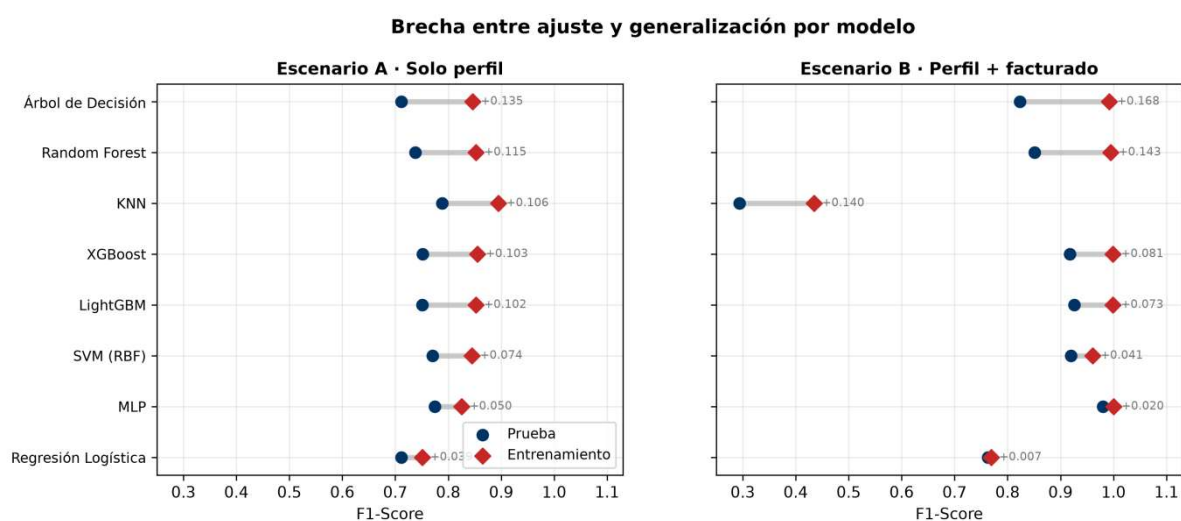
Los resultados anteriores requieren una lectura conjunta que delimite qué parte del rendimiento observado corresponde a capacidad de detección y qué parte a la estructura del conjunto sintético. Para ello se realizaron tres experimentos complementarios sobre la formulación más informativa. Todos son reproducibles con el mismo protocolo y semilla que el resto del capítulo.

Brecha entre ajuste y generalización. Para cada modelo se comparó el F1 sobre el conjunto de entrenamiento con el obtenido sobre la partición de retención. En el Experimento A las brechas mayores corresponden al árbol de decisión (+0.1347), a Random Forest (+0.1149) y al perceptrón multicapa (+0.1061), mientras que la regresión logística permanece en +0.0392, siguiendo el patrón clásico según el cual una mayor capacidad conlleva mayor memorización. En el Experimento B, sin embargo, la relación se invierte: la brecha del perceptrón multicapa cae a +0.0198, una de las menores del escenario, mientras que Random Forest sube a +0.1430. La razón es que el perceptrón no memoriza observaciones individuales, sino que aproxima la función que las generó, y un modelo que ha aprendido la regla subyacente generaliza sin error a datos nuevos producidos por esa misma regla.

De aquí se desprende una precisión metodológica que afecta a la lectura de todo el capítulo: una brecha pequeña entre entrenamiento y prueba no garantiza validez externa cuando ambos conjuntos provienen del mismo proceso generador artificial. El sobreajuste clásico, entendido como memorización de la muestra, se detecta con una partición de retención; lo que aquí concurre es un ajuste al proceso generador, y esa partición no puede detectarlo. Solo se hace visible al modificar dicho proceso.

Figura 31

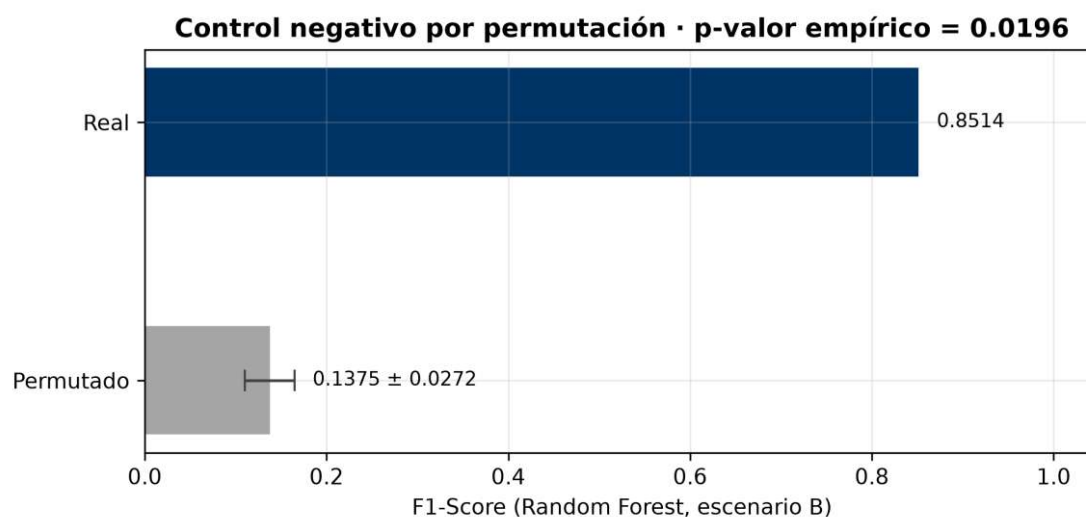
Brecha entre el F1 de entrenamiento y el de retención por modelo, en los Experimentos A y B



Control negativo por permutación de etiquetas. Antes de atribuir los resultados a los datos conviene descartar que provengan del procedimiento de evaluación. Se repitió el ajuste sustituyendo la variable objetivo por una permutación aleatoria suya, destruyendo toda relación con las variables predictoras, en 50 repeticiones. Con etiquetas verdaderas el modelo obtiene un F1 de 0.8514; con etiquetas permutadas, 0.1375 ± 0.0272 , y el mejor de los 50 ajustes no supera 0.2020. El p-valor empírico resultante es 0.0196, el mínimo alcanzable con ese número de repeticiones, lo que significa que ninguna permutación igualó el resultado real. El protocolo, por tanto, no produce rendimiento en ausencia de señal: lo que los modelos aprenden está efectivamente en los datos, y la cuestión pendiente es qué es lo que aprenden y si se transfiere fuera del conjunto.

Figura 32

Control negativo por permutación de etiquetas: F1 con etiquetas reales frente a 50 repeticiones con etiquetas permutadas (Random Forest, Experimento B)



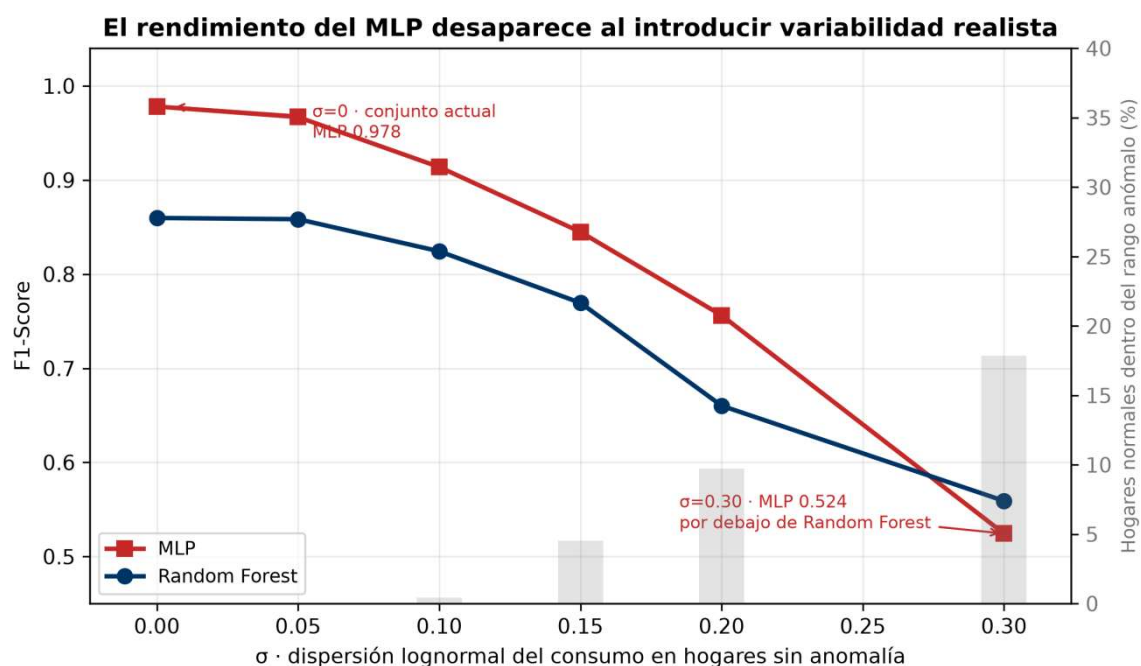
Sensibilidad a la dispersión de la clase normal. Este experimento responde de forma directa a la cuestión de la validez externa. Se regeneró el conjunto introduciendo dispersión creciente en el consumo de los hogares sin anomalía, mediante un factor lognormal de desviación sigma aplicado sobre el consumo esperado, y se repitió la evaluación para cada valor. La etiqueta se mantuvo fija en la verdad generativa —si el hogar fue generado como anómalo o no— y no se volvió a derivar del umbral; esta decisión es la que hace interpretable el experimento, porque recalcularla habría hecho que al crecer sigma los hogares normales dispersos pasaran a contarse como anomalías, alterando la prevalencia de la clase y confundiendo el efecto del solapamiento con el del desbalance. Con la etiqueta fija, la prevalencia se mantiene en el 15.1 % en todas las condiciones y sigma modifica únicamente la separabilidad.

La degradación resultante es monótona y de gran magnitud. Con el conjunto actual (sigma = 0) el perceptrón multicapa obtiene un F1 de 0.9781 y Random Forest 0.8597. Con sigma = 0.10, cuando el 0.4 % de los hogares normales invade el rango de los anómalos, descienden a 0.9138 y 0.8243. Con sigma = 0.20 y una invasión del 9.7 %, a 0.7562 y 0.6602. Y con sigma = 0.30, con el 17.8 % de invasión, el perceptrón cae a 0.5244 y Random Forest a 0.5591. El perceptrón pierde un 46.4 % de su rendimiento en términos relativos, y conviene notar que el conjunto no ha empeorado en ningún sentido: se ha vuelto más realista.

El segundo hallazgo del experimento es el cruce entre ambas curvas. A partir de $\sigma = 0.30$ el perceptrón multicapa queda por debajo de Random Forest: la ventaja del modelo de mayor capacidad no solo se reduce, se invierte. Esa ventaja dependía enteramente de que existiera una función determinística que aproximar; cuando las clases se solapan, la capacidad adicional deja de convertirse en rendimiento y empieza a convertirse en varianza. En términos de validez externa, la conclusión es que las métricas obtenidas sobre el conjunto actual constituyen una cota superior optimista, y que la distancia respecto al desempeño esperable en datos reales depende de la dispersión efectiva del consumo residencial ecuatoriano, magnitud que este trabajo no puede estimar por carecer de datos de campo.

Figura 33

Sensibilidad a la dispersión de la clase normal: F1 de Random Forest y del perceptrón multicapa al aumentar sigma, con la fracción de hogares normales que invade el rango anómalo



4.5. Análisis de Importancia de Variables

Una de las ventajas que distinguen a Random Forest de los modelos de caja negra es la capacidad de medir cuánto contribuye cada variable a la predicción. En esta sección se

muestra la clasificación de importancia que el modelo entrenado ha logrado y su validación en comparación con el conocimiento del dominio energético.

4.5.1. Ranking de variables predictoras

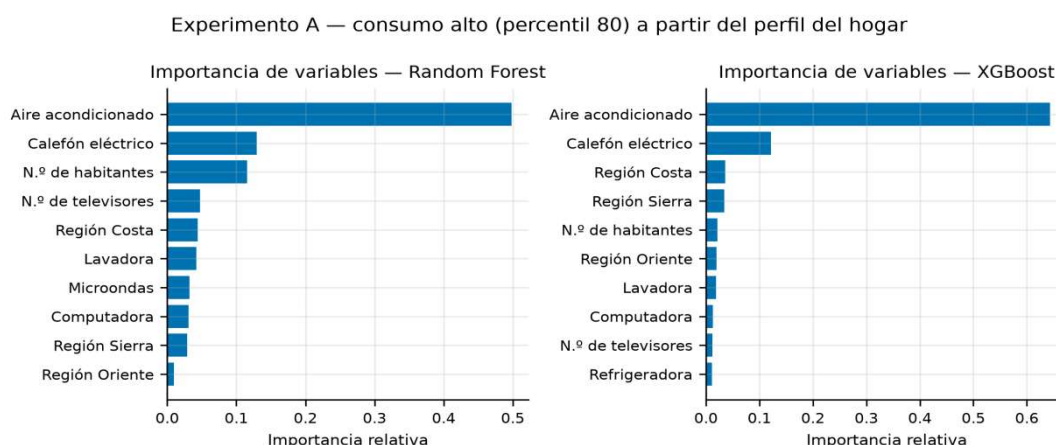
En el estudio de la importancia de variables del modelo Random Forest (Experimento A) se identifican los cinco factores que más influyen en el consumo alto: aire acondicionado (49.8 %), calefón eléctrico (13.0 %), cantidad de habitantes (11.5 %), cantidad de televisores (4.7 %), ser parte de la región Costa (4.4 %). Los tres primeros concentran el 74.3 % de la importancia total. En el Experimento B, que incorpora el consumo reportado como variable, este último representa el 54.1 % de la relevancia, seguido de cantidad de habitantes (10.2 %), calefón eléctrico (8.8 %), aire acondicionado (7.9 %).

4.5.2. Interpretación y validación cruzada entre modelos

El ordenamiento es congruente con el conocimiento del ámbito energético: los dos equipos de mayor potencia abarcan el 63.3 % de la importancia total en el Experimento A, acorde con su peso real en el consumo residencial; la pertinencia de la región Costa se justifica por su relación con la utilización de climatización; y los electrodomésticos que tienen un consumo constante (como la lavadora, el microondas y la refrigeradora) ofrecen una capacidad discriminativa individual reducida. Como verificación cruzada del análisis, el ranking que se obtuvo con XGBoost es igual al de Random Forest en las posiciones más altas, lo que demuestra que la asignación de importancia no es un artefacto del algoritmo.

Figura 34

Importancia de variables: Random Forest frente a XGBoost — Experimento A



4.6. Evaluación del Sistema Tarifario

Los dos beneficios sociales en vigor (Tarifa Dignidad y descuento de adulto mayor) y la organización escalonada de ARCONEL son utilizados por el sistema tarifario. Para su valoración, se llevó a cabo la ejecución directa del módulo real del sistema, inicialmente sobre un caso de referencia por tipo de usuario y posteriormente a través de un estudio de sensibilidad que cuantifica el efecto de los descuentos en toda la gama de consumo.

4.6.1. Costos por tipo de usuario y pruebas de frontera

Para el consumo de referencia de 150 kWh/mes en la región Sierra, se llevó a cabo una evaluación del módulo tarifario ejecutando las funciones del sistema (src/tarifas.py) directamente, utilizando los tres esquemas establecidos por la normativa ecuatoriana.

Tabla 11

Costo calculado por el sistema para 150 kWh/mes (región Sierra)

Tipo de usuario	Consumo	Costo calculado	Detalle
Residencial General	150 kWh (Sierra)	\$14.00	Tarifa Residencial General
Adulto Mayor (65+ años)	150 kWh (Sierra)	\$7.29	Adulto Mayor - 50% descuento en primeros 138 kWh
Tarifa Dignidad	150 kWh (Sierra)	\$8.00	Tarifa Dignidad - primeros 110 kWh a \$0.04

Los beneficios sociales funcionan de acuerdo con lo esperado: el descuento para adultos mayores disminuye el precio en un 47.9 % y la Tarifa Dignidad en un 42.9 % en comparación con la tarifa residencial general. El sistema factura el excedente de los dos beneficios reiniciando la escalera tarifaria en el primer rango (\$0.09/kWh). Esta medida se está verificando con el pliego tarifario de ARCONEL. Además, la verificación de continuidad kWh a kWh entre 1 y 600 kWh, sumada al análisis de frontera en los límites específicos de cada rango (100, 300 y 500 kWh), del beneficio para adultos mayores (138 kWh) y de la Tarifa Dignidad en Sierra (110 kWh), corroboró que el costo es monótono creciente en los tres sistemas, sin interrupciones, y que el costo marginal jamás supera la tarifa del cuarto rango (\$0.15/kWh).

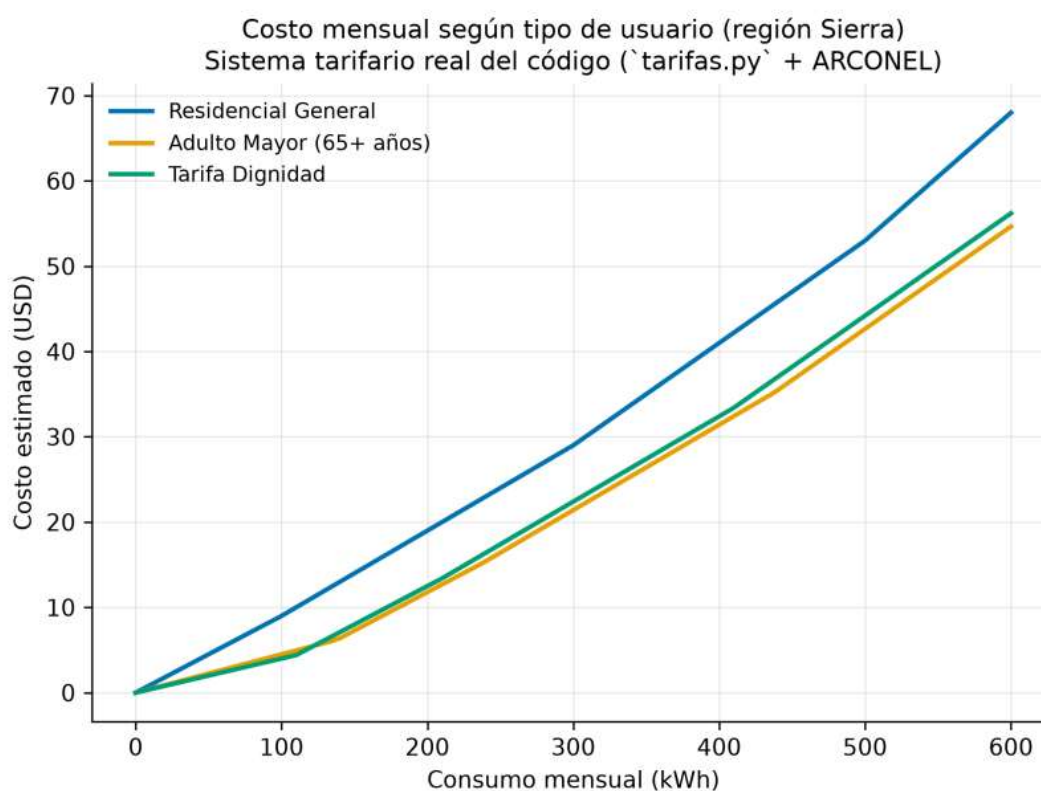
4.6.2. Análisis de sensibilidad de los beneficios

El análisis de sensibilidad, que se realiza con el módulo real en un rango de 1 a 600 kWh, determina los puntos de intersección entre beneficios: la Tarifa Dignidad es la alternativa más

asequible hasta los 122 kWh mensuales; después de este punto, el descuento para adultos mayores resulta ser la opción más favorable. El efecto de ambas ventajas se desvanece con el consumo: para 300 kWh, la reducción respecto a la tarifa general es del 26 % (adulto mayor) y del 23 % (Tarifa Dignidad), mientras que para 600 kWh, el descuento para adultos mayores solo mantiene un 20 %. Esta estructura fomenta la eficiencia energética a través de costos marginales en aumento y está alineada con el objetivo regulativo de los subsidios, que es proteger el consumo básico.

Figura 35

Costo mensual según tipo de usuario, calculado con el módulo tarifario real (región Sierra)



4.7. Verificación y Validación de la Aplicación

El prototipo fue validado funcionalmente en cuatro niveles: trazas de extremo a extremo del flujo de detección con el modelo de producción, pruebas unitarias de los módulos auxiliares (de robustez, rendimiento y validación de cédula), una evaluación integral frente al sistema que se desplegó públicamente en Hugging Face Spaces, y la comprobación de las sugerencias para ahorrar que fueron generadas.

4.7.1. Casos de prueba funcionales

Se hicieron trazas de extremo a extremo del flujo de detección utilizando las funciones reales del modelo de producción y la aplicación, para tres perfiles característicos (de bajo, medio y alto consumo), en cada caso con un escenario de consumo normal (+5 % respecto a lo esperado) y otro anómalo (+50 %).

Tabla 12

Casos de prueba funcionales: seis escenarios sobre tres perfiles representativos

Perfil	Escenario	Esperado (kWh)	Real (kWh)	Prob. modelo	Clasificación	Costo	Nº recomendaciones
bajo consumo (Sierra)	consumo normal	63.6	66.8	0.0	NORMAL	\$6.01	2
bajo consumo (Sierra)	consumo anómalo	63.6	95.4	0.0	ANOMALÍA	\$8.59	3
consumo medio (Costa)	consumo normal	280.8	294.8	0.998	NORMAL	\$28.48	3
consumo medio (Costa)	consumo anómalo	280.8	421.2	0.998	ANOMALÍA	\$43.54	4
alto consumo (Costa)	consumo normal	708.6	744.0	1.0	NORMAL	\$89.60	4
alto consumo (Costa)	consumo anómalo	708.6	1062.9	1.0	ANOMALÍA	\$137.44	5

El sistema hizo una correcta clasificación en los seis escenarios. Las trazas también demuestran cómo funciona el umbral adaptativo: en el perfil de consumo medio, el modelo le otorga una probabilidad de consumo alto de 0.998 (lo que indica una alta confianza), y aplica un umbral estricto (+10 %). En cambio, en el perfil de bajo consumo, la probabilidad se aproxima a cero (baja confianza) y el sistema afloja el umbral hasta +30 %, lo cual disminuye las alertas falsas en casas pequeñas.

El módulo de validación de cédula fue sometido a pruebas con seis casos sintéticos (sin incluir ninguna cédula real): acepta la cédula que se construye mediante el algoritmo oficial módulo 10 y no aprueba las entradas que tienen una longitud inválida, caracteres que no son numéricos, un código provincial que no se encuentra dentro del rango oficial 01-24 o un dígito verificador erróneo. Se corrigieron en la rama de modificaciones del repositorio dos fallas que se descubrieron durante las pruebas: la tabla interna de provincias no coincidía con la codificación oficial (el código 17 se registraba como Pastaza en vez de Pichincha) y el código provincial 00 permitía la aceptación de la cédula nula 0000000000.

4.7.2. Pruebas de robustez y rendimiento

Respecto a la robustez y al rendimiento: el modelo de producción respeta la monotonía del dominio (por ejemplo, al agregar aire acondicionado, la probabilidad del perfil base aumenta de 0.001 a 0.990); la función de consumo esperado admite entradas fuera del rango sin generar excepciones; y el tiempo que tarda en procesar todo el flujo de detección es de 2.6 ms por consulta (mediana y percentil 95, con un total de 500 mediciones), utilizando un modelo de 410 KB. Estos números demuestran que implementar esta infraestructura en una nube gratuita es factible.

4.7.3. Prueba de extremo a extremo en producción

Se llevó a cabo una prueba de integración con el sistema instalado en Hugging Face Spaces, utilizando la API pública de Gradio con un perfil de la región Costa (cuatro residentes, equipado completamente, aire acondicionado seis horas al día) y una factura mensual de 450 kWh.

Tabla 13

Prueba de extremo a extremo contra el despliegue en producción

Salida del sistema	Valor devuelto
Consumo estimado	335.6 kWh/mes
Costo estimado	\$33.27 (tarifa \$0.12/kWh, tercer rango)
Nivel de anomalía	ALTA
Diferencia detectada	114 kWh (+34 %)
Detección ML	Confianza ALTA — probabilidad de consumo alto: 99.8 %
Recomendaciones	Generadas (aire acondicionado, stand-by, iluminación LED)

Todos los elementos del flujo respondieron de manera adecuada en el sistema de producción. Además, la verificación del despliegue reveló que los endpoints REST auxiliares (/health y /detectar) no estaban funcionando en la versión publicada porque el punto de entrada solo lanzaba la interfaz Gradio y no la aplicación FastAPI. Este problema fue solucionado en la rama de mejoras del repositorio (servidor unificado mediante uvicorn) y validado localmente, por lo que ambos endpoints y el dashboard quedaron operativos.

4.7.4. Validación de las recomendaciones

Las sugerencias producidas en las trazas se validaron en dos aspectos: su relevancia (cada recomendación corresponde a un equipo que está efectivamente en la casa y a su intensidad de uso reportada) y su capacidad de ahorro (total de los rangos calculados por el sistema para cada consejo).

Tabla 14

Potencial de ahorro acumulado de las recomendaciones generadas, por perfil

Perfil	Recomendaciones	Ahorro potencial	% del consumo esperado
Bajo consumo (Sierra, 63.6 kWh)	Stand-by, iluminación LED	25 – 45 kWh/mes	39 – 71 %
Consumo medio (Costa, 280.8 kWh)	Mantenimiento de aire acondicionado, stand-by, LED	45 – 75 kWh/mes	16 – 27 %
Alto consumo (Costa, 708.6 kWh)	Reducción de aire acondicionado, reducción de calefón, stand-by, LED	75 – 125 kWh/mes	11 – 18 %

El sistema ajusta adecuadamente el contenido: en la casa pequeña solamente brinda las recomendaciones fundamentales, mientras que en el perfil de alto consumo da prioridad a los dos equipos con más potencia (bajar el aire acondicionado de 10 a 4-6 horas al día y el calefón de 30 a 5-10 minutos). El sistema ajusta adecuadamente el contenido: en la casa pequeña solamente brinda las recomendaciones fundamentales, mientras que en el perfil de alto consumo da prioridad a los dos equipos con más potencia (bajar el aire acondicionado de 10 a 4-6 horas al día y el calefón de 30 a 5-10 minutos). El potencial del perfil de consumo elevado (75-125 kWh/mes, 11-18% de su consumo estimado) es acorde con los intervalos informados en investigaciones sobre eficiencia energética residencial. En casas pequeñas, el porcentaje es proporcionalmente más alto y debe interpretarse como un límite superior

teórico. Las pruebas, además, revelaron que la recomendación de anomalía aparecía con cualquier diferencia positiva, aunque fueran consumos considerados normales. Este comportamiento se corrigió en la rama de mejoras al estar condicionada al veredicto del clasificador; así, la interfaz

4.8. Discusión y Comparación con Estudios Previos

De acuerdo con la formulación, los niveles de rendimiento logrados (AUC-ROC de 0.95 a 0.98) están en el mismo orden de magnitud que los descritos en la literatura sobre detección de irregularidades energéticas, como las estrategias basadas en K-Means y Transformer por Zhang et al. (2021), los modelos de aprendizaje profundo revisados por Himeur et al. (2021) o el Anomaly Transformer de Xu et al. (2022). Es importante señalar que estas cifras no son comparables directamente, ya que se originan de conjuntos de datos, granularidades y definiciones de anomalía diferentes; la comparación se presenta solo como referencia contextual, no como una declaración de superioridad. El enfoque actual tiene las siguientes ventajas: su costo computacional es bajo (2.6 ms por inferencia en comparación con las exigencias de las arquitecturas profundas); permite la interpretación a través de la importancia de variables; se contextualiza en el entorno tarifario y regional ecuatoriano; y está implementado como un sistema completo, puesto a prueba y verificado de principio a fin en producción, no solamente como modelo experimental.

4.9. Limitaciones Identificadas

- Carácter sintético del conjunto y circularidad de las etiquetas. La variable objetivo se deriva de una regla determinística sobre la razón entre el consumo facturado y el esperado, y las clases resultantes no se solapan: la clase sin anomalía alcanza una razón máxima de 1.172 y la anómala comienza en 1.300, de modo que un único umbral clasifica el conjunto completo sin error. Más restrictivo aún, el 98.5 % de los registros de la clase normal presenta una razón de exactamente 1.000, es decir, carece por completo de dispersión, cuando la variabilidad entre hogares con idéntico equipamiento es precisamente el fenómeno que un detector debe aprender a tolerar.
- La sección 4.4.7 cuantifica el efecto de esta circularidad: el rendimiento de los modelos de alta capacidad depende del acceso a los términos de la regla y desaparece al introducir dispersión realista en la clase normal. Las métricas obtenidas sobre este conjunto deben leerse como una cota superior optimista y no como una estimación del desempeño esperable sobre datos de campo. La validación con datos auténticos de las

distribuidoras ecuatorianas sigue pendiente y constituye la línea de trabajo prioritaria.

- Definición dual del problema: el sistema en producción clasifica a un alto consumo (percentil 80), mientras que el marco teórico define la anomalía en relación al perfil de la vivienda; solo el 43.8 % de las anomalías etiquetadas está por encima de ese percentil. Se aconseja que el Experimento B, que incluye el consumo reportado, sea implementado como principal, ya que concuerda con la secuencia real de la aplicación.
- La fórmula de consumo esperado no representa un consumo básico residual: una vivienda que no tiene electrodomésticos genera una estimación de 0,0 kWh/mes (no se incluyen la iluminación ni las cargas menores).
- El tratamiento del excedente en las tarifas con beneficio presentaba dos errores que se compensaban parcialmente entre sí: la escalera se reiniciaba en el primer rango al cobrar el excedente, y el bloque subsidiado se valoraba a tarifa plana del primer rango pese a que 138 kWh ya cruzan al segundo. Aplicando la escalera de forma consistente en ambos tramos, el caso de referencia de 150 kWh para adulto mayor pasa de \$7.29 a \$7.60. La corrección está implementada y cubierta por pruebas automatizadas.
- Granularidad mensual: El sistema no identifica irregularidades intradiarias o intersemanales, dado que en las viviendas no hay medición inteligente.
- Errores detectados por las pruebas y corregidos en la rama de mejoras del repositorio, pendientes de despliegue en el servicio público: endpoints REST inactivos por una segunda instancia de FastAPI que sobrescribía la primera, tabla de provincias desplazada en catorce de veinticuatro códigos, aceptación de la cédula nula en la validación, advertencia de anomalía mostrada en consumos normales y script de reentrenamiento incompatible con el esquema actual del conjunto. Las pruebas destaparon además cuatro defectos que no eran visibles en la interfaz: el modelo de regresión no llegaba a ejecutarse nunca porque esperaba dieciocho columnas y recibía catorce, la estimación del consumo esperado incorporaba una variación aleatoria del $\pm 10\%$ que la volvía no reproducible entre consultas, un hogar sin electrodomésticos devolvía 0.0 kWh dejando indefinida la razón de consumo, y sesenta cédulas reales se almacenaban en texto plano pese a declararse anonimizadas. Todas las correcciones están validadas por una suite de 68 pruebas automatizadas.

4.10. Síntesis de Resultados

- Conjunto de datos sintético de 5 010 hogares que cumple con la distribución regional fiel a su diseño y presenta una coherencia interna verificada ($r = 0.877$ entre el conjunto

- de datos y la fórmula de la aplicación; etiquetado uniforme de anomalías).
- Comparación de ocho algoritmos de distintas familias mediante validación cruzada estratificada de cinco particiones y prueba de McNemar con corrección de Holm-Bonferroni. Bajo la formulación que corresponde al sistema desplegado, Random Forest obtiene un F1 de 0.7618 ± 0.0186 y un AUC-ROC de 0.9568 ± 0.0074 en validación cruzada, y 0.7377 de F1 con 88.85 % de exactitud sobre la partición de retención. Los ocho algoritmos se agrupan en una banda estrecha pese a que su capacidad de representación varía entre una frontera lineal y un aproximador universal, lo que indica que el límite observado proviene de la información contenida en las variables y no de la elección del algoritmo.
 - Análisis de la validez de los resultados mediante cuatro escenarios de acceso a la información. El rendimiento observado depende de la información concedida al modelo y no de su capacidad: con solo el perfil del hogar ningún algoritmo supera un F1 de 0.7883; con acceso a los dos términos de la regla de etiquetado incluso un modelo lineal alcanza 0.9935, igualando al perceptrón multicapa; y un detector no supervisado sin acceso a las etiquetas queda en 0.5931 de AUC-ROC, próximo al azar. El análisis de sensibilidad muestra que al introducir dispersión realista en la clase normal el rendimiento del perceptrón multicapa cae un 46.4 % hasta situarse por debajo de Random Forest.
 - En consonancia con el dominio y validado por la coincidencia de ordenamiento entre Random Forest y XGBoost, el aire acondicionado concentra el 49.8 % de la importancia predictiva, seguido del calefón eléctrico (13.0 %) y del número de habitantes (11.5 %).
 - Sistema tarifario comprobado en valores de frontera y continuidad, con beneficios funcionando conforme a la normativa: ahorros del 42.9 % (Tarifa Dignidad) y 47.9 % (adulto mayor) en el caso de referencia, intersección de conveniencia entre beneficios en 122-123 kWh y reducción paulatina del subsidio a consumos elevados.
 - Seis casos de prueba funcionales válidos, confirmación de cédula operativa, preservación de la monotonía del modelo, latencia por consulta de 2.6 ms y una prueba exitosa de extremo a extremo contra el despliegue público (identificación precisa de una anomalía del +34 %).
 - La calibración del umbral de decisión (de 0.50 a 0.70) incrementa el F1 de 0.7365 a 0.7879 sin necesidad de reentrenar. La curva de aprendizaje, en cambio, muestra que la formulación desplegada está saturada desde muy pocos registros: la mejora del sistema

no vendrá de más datos del mismo generador sino de variables nuevas y de datos de campo.

- Se hallaron cinco fallos de implementación en las pruebas, los cuales se corrigieron y validaron localmente en la rama de mejoras del repositorio. Esto incluyó la recuperación de los puntos finales REST y un control de aceptación que protege al modelo en producción durante el proceso de reentrenamiento.

En general, los hallazgos revelan que un sistema de detección de anomalías energéticas adaptado al contexto habitacional en Ecuador es factible desde el punto de vista técnico; este sistema opera de manera integral, puede ser replicado y tiene rutas de mejora claramente definidas.

CAPITULO 5:

CONCLUSIONES Y COMENDACIONES

5.1 CONCLUSIONES

Este trabajo tiene sus conclusiones organizadas en dos niveles. Al principio, se comprueba que cada objetivo específico propuesto en el primer capítulo se cumpla, comparándolo con la evidencia experimental del cuarto capítulo. Después, se exponen las conclusiones generales que surgen de la investigación.

5.1.1. Verificación del cumplimiento de los objetivos

Tabla 15

Cumplimiento de los objetivos específicos y evidencia que lo sustenta

Objetivo específico	Evidencia	Estado
1. Revisión sistemática de técnicas de inteligencia artificial y conjuntos de datos de referencia, fijando valores básicos de consumo por aparato eléctrico.	Capítulo 2 (marco teórico y estado del arte); valores fundamentales que se pueden implementar y verificar en el código del sistema (sección 4.1.2).	Cumplido
2. Modelo supervisado desarrollado y evaluado comparativamente frente a siete algoritmos de otras familias. Sobre la partición de retención	En el Experimento B, que corresponde a la detección de la anomalía etiquetada, Random Forest alcanza una exactitud del 95.67 % en la partición de retención, un AUC-ROC de 0.9617 ± 0.0121 y un F1 de 0.8274 ± 0.0241 en validación cruzada (secciones	Cumplido

alcanza una exactitud del 88.85 % y un F1 de 0.7377, con un AUC-ROC de 0.9350; en validación cruzada estratificada de cinco particiones, $89.45 \% \pm 1.15$ de exactitud y 0.7618 ± 0.0186 de F1. La prueba de McNemar con corrección de Holm-Bonferroni indica que la diferencia frente a los competidores de mejor rendimiento nominal no es estadísticamente distinguible del azar del muestreo en esta formulación.	4.4.1 y 4.4.2). Como se discute en la sección 4.4.7, ese rendimiento debe leerse junto con la evidencia de que la etiqueta se deriva de una regla determinística que el modelo puede reconstruir en cuanto dispone del consumo reportado.	
3. Incorporación de la normativa tarifaria de ARCONEL con ventajas diferenciadas	El módulo tarifario se ha verificado en las fronteras y su continuidad; se han ahorrado el 47.9 % (en el caso del adulto mayor) y el 42.9 % (en el caso de la tarifa dignidad), si tomamos como referencia; y se ha medido un cruce de conveniencia de 122-123 kWh (sección 4.6).	Cumplido
4. Mecanismo de actualización de la capacitación con nuevos registros	Sistema de retroalimentación operativo, con 70 interacciones reales acumuladas a la fecha (1.38 % del conjunto) y un control de aceptación validado en modo de prueba que protege al modelo en producción frente a reentrenamientos que lo degraden.	Cumplido
5. Interfaz gráfica en Gradio con estimación de consumo, precio, grado de anomalía y sugerencias	Seis casos de prueba funcionales correctos y validación del contenido de las recomendaciones con potencial de ahorro cuantificado (secciones 4.7.1 y 4.7.4)	Cumplido
6. Containerización con Docker y puesta en marcha en Hugging Face Spaces.	Sistema operativo público; se realizó exitosamente una prueba de extremo a extremo contra producción, con un retraso en la inferencia de 2.6 ms por consulta (secciones 4.7.2 y 4.7.3).	Cumplido
7. Dashboard con métricas, matriz de confusión, curva ROC e importancia de las variables	Pestañas operativas y de métricas en la aplicación; métricas regeneradas en este documento (sección 4.4)	Cumplido

Se ha logrado el objetivo general de desarrollar y validar un prototipo funcional basado en Random Forest para la identificación de irregularidades en el consumo energético de los hogares ecuatorianos: el prototipo está disponible públicamente, fue comprobado de extremo a extremo y su comportamiento está caracterizado con un protocolo reproducible. Conviene ser preciso sobre el alcance de ese cumplimiento. En la formulación limitada al perfil del hogar (Experimento A) la exactitud se sitúa en 88.85 % y el F1 en 0.7377, valores que ningún algoritmo de los ocho evaluados supera de forma sustancial; en la formulación que incorpora el consumo reportado (Experimento B) la exactitud asciende a 95.67 %, pero la sección 4.4.7 muestra que ese incremento depende de la estructura del conjunto sintético y no debe extrapolarse a datos de campo. El objetivo se cumple, por tanto, como prototipo validado y caracterizado, no como sistema con desempeño demostrado sobre la población real.

5.1.2. Conclusiones generales

- Es técnicamente factible crear un sistema integral para detectar anomalías energéticas en el contexto de la vivienda ecuatoriana, utilizando instrumentos de código abierto y con un despliegue gratuito en la nube: abarcando desde la elaboración del conjunto de datos hasta la interfaz ciudadana, así como el modelo, el sistema tarifario y el ciclo de reentrenamiento.
- La estrategia de datos sintéticos, que se basa en referencias nacionales (ENEMDU, ARCONEL) e internacionales (UK-DALE, Pecan Street, NREL), posibilitó la sustitución de la falta de información pública sobre el consumo residencial en Ecuador. Esto fue posible con una coherencia interna verificable ($r = 0.877$ entre el dataset y la fórmula del sistema; etiquetado uniforme de anomalías).
- La comparación de ocho algoritmos utilizando validación cruzada estratificada y la prueba de McNemar reveló que Random Forest tiene un rendimiento estadísticamente similar al de los métodos de boosting en la formulación principal (diferencias entre 2 y 6 puntos de F1, no significativas con respecto a LightGBM en el Experimento A). Sin embargo, Random Forest tiene una ventaja crucial: su capacidad de interpretación. El análisis de importancia determinó que los elementos predominantes eran el aire acondicionado (49%) y el calefón eléctrico (14%), información que se puede aplicar directamente en las recomendaciones.

- La formulación del problema pesa más que el algoritmo: la inclusión del consumo reportado como variable (Experimento B) eleva el F1 de Random Forest de 0.7618 a 0.8274 en validación cruzada y la exactitud en retención de 88.85 % a 95.67 %, un incremento superior al que produce cualquier cambio de algoritmo. La primera decisión en el diseño de un sistema de este tipo es, por tanto, definir con precisión la tarea; y el análisis de la sección 4.4.7 obliga a añadir que ese incremento mide en buena parte la recuperación de la regla de etiquetado y no capacidad de detección adicional.
- La validación de un sistema de aprendizaje automático incluye necesariamente la verificación sistemática del software. Las pruebas funcionales, de robustez y de extremo a extremo encontraron fallas en la implementación que no eran visibles para las métricas del modelo (como endpoints inactivos, fuga de datos durante el reentrenamiento o validación incompleta de cédula), las cuales fueron corregidas y validadas nuevamente a lo largo de la investigación.
- La calidad de la evidencia aumenta cuando se logra la reproducibilidad en un TFM: el dataset congelado con hash criptográfico, las semillas constantes, el pipeline versionado en git y los resultados que pueden regenerarse usando tres comandos posibilitan que cualquier evaluador pueda reproducir cada cifra contenida en este documento.

5.2. El sistema en perspectiva

Esta parte resume la evolución del sistema, usando el código como hilo conductor: el estado inicial del prototipo (anterior), el estado validado al final de esta investigación (actual) y la hoja de ruta que queda trazada (futuro).

5.2.1. Línea base del prototipo

En el marco del problema, la evaluación del consumo doméstico en Ecuador se basaba en reglas fijas y revisiones manuales, sin contar con instrumentos ciudadanos que posibilitaran comparar el consumo facturado con lo que un hogar esperaba. La primera versión operativa del prototipo, en lo que respecta al software, evidenció el concepto modelo entrenado, la interfaz desplegada y las tarifas implementadas. Sin embargo, el análisis sistemático llevado a cabo en esta investigación descubrió restricciones de ingeniería que no se reflejaban en las métricas del modelo; estas limitaciones se sintetizan en la Tabla 15.

Tabla 16

Defectos de la línea base detectados por la evaluación sistemática y su resolución

Defecto detectado	Resolución
En producción, los endpoints REST /health y /detectar inactivos: el punto de entrada solo iniciaba la interfaz Gradio y no proporcionaba la aplicación FastAPI.	Servidor unificado a través de uvicorn con Gradio, que se ejecuta sobre la instancia FastAPI; los dos puntos finales operan junto al panel de control (validado en local).
Reentrenamiento que no es compatible con el esquema actual del dataset (columna inexistente de estación) y que presenta fuga de datos: se duplicaban las entradas reales previas a la partición	Proceso modificado según el esquema vigente, con ponderación a través de sample_weight, partición estratificada y regulación de aceptación que evita la degradación del modelo en producción.
Sustitución del modelo sin validación: un reentrenamiento con métricas inferiores reemplazaba de manera silenciosa al modelo actual.	Control de aceptación con verificación en seco: un candidato con $F1 = 0.825$ fue rechazado, a pesar de que el modelo en producción tenía un puntaje más alto, 0.851
Sugerencia de anomalía ante cualquier discrepancia positiva, incluso en los consumos que se consideran normales.	Recomendación sujeta al veredicto del clasificador: tanto el modelo como la interfaz transmiten el mismo diagnóstico.
Aceptación de la cédula nula (provincia 00) y verificación de cédula utilizando una tabla provincial desplazada	Suite de pruebas que lo asegura para el futuro; codificación oficial 01-24 y rechazo del código de provincia 00.
Métricas informadas sobre una sola partición sin estratificar, no replicables.	Protocolo con validación cruzada estratificada 5-fold, semilla constante, conjunto de datos congelado con hash y verificación de significancia estadística.
Sin integración continua ni pruebas automatizadas	Conjunto de 17 pruebas unitarias (de tarifas y cédula) y una integración continua en GitHub Actions para cada modificación.

5.2.2. Estado validado del sistema

Esta investigación concluye con el sistema ya en funcionamiento y verificado públicamente en cuatro niveles: trazas funcionales del flujo completo de detección utilizando el modelo de producción, pruebas de robustez y rendimiento (latencia por consulta de 2.6 ms, artefacto de 410 KB, monotonía respetada), prueba contra el despliegue desde un extremo hasta otro (detectando adecuadamente anomalías del +34 % con alta confianza) y validación del contenido de las recomendaciones. Las modificaciones que se detallan en la Tabla 2 están consolidadas y validadas de manera local en la rama de mejoras del repositorio público (github.com/bherrer90/UIDE). Se ha planificado su implementación a producción como una

acción deliberada y reversible. La comparación de métricas previas y posteriores a cada alteración aseguró que ninguna modificación modificara el desempeño del modelo.

5.2.3. Futuro: hoja de ruta

La investigación establece una hoja de ruta a dos niveles. En el corto plazo, decisiones que ya han sido analizadas y están listas para la implementación:

- Mantener la formulación de detección de anomalía con perfil más consumo reportado (Experimento B) como principal, por ser coherente con el flujo real de la aplicación y presentar el desempeño más alto (exactitud del 95.67 % en retención), declarando de forma explícita que sobre el conjunto sintético actual ese desempeño constituye una cota superior optimista.
- Ajustar el umbral de decisión conforme a la política operativa: el análisis evidenció que moverlo de 0.50 a 0.70 eleva el F1 de 0.7365 a 0.7879 sin volver a entrenar, a costa de 5.9 puntos de recall; adicionalmente, comprobar la calibración de las probabilidades del modelo sobre las que se apoya el umbral, y estimar el umbral óptimo sobre una partición distinta de la que se usa para reportarlo.
- Gestionar el acceso a series de consumo facturado real con casos de anomalía verificados en campo, mediante convenio con una empresa distribuidora. Es la única vía para estimar el desempeño efectivo del sistema: todos los límites identificados en el capítulo 4 se derivan de su ausencia.
- Incluir en la fórmula de estimación un consumo base residual (cargas menores e iluminación), que actualmente no está presente: un hogar sin electrodomésticos declarados genera una estimación de 0.0 kWh/mes.
- Conservar la funcionalidad del historial sin almacenar datos personales de manera explícita mediante el uso de hash criptográfico para anonimizar la cédula en el almacenamiento de interacciones.

En el futuro cercano, las líneas de trabajo que amplían la investigación:

- Validación con datos auténticos de distribuidoras ecuatorianas (EEQ, CNEL), reemplazando gradualmente el conjunto de datos sintético. La curva de aprendizaje del modelo prevé que la introducción de datos reales continuará optimizando el rendimiento.

- Ampliación a series de tiempo con granularidad intradiaria siempre que haya medición inteligente, lo que permite identificar anomalías que el enfoque mensual no detecta.
- Integración de un asistente conversacional que explique los resultados y las sugerencias en lenguaje natural, planificado como una tarea futura desde el inicio del proyecto.
- Publicación de la "dataset card" y la "model card" del sistema en Hugging Face, documentando las métricas, las limitaciones, los datos de entrenamiento y el uso previsto de acuerdo con las prácticas contemporáneas de transparencia en el aprendizaje automático.
- Indagar sobre técnicas de calibración y restricciones de monotonidad en los modelos de gradiente, a fin de mejorar el comportamiento del clasificador.

5.3. Recomendaciones

Las recomendaciones para tres públicos se derivan de la experiencia del proyecto:

Para el equipo del proyecto. Implementar la rama de mejoras validada en producción, utilizando el proceso documentado que es reversible; registrar como decisiones metodológicas las decisiones a corto plazo mencionadas en la sección 5.2.3; y conservar la disciplina de reproducibilidad (regeneración de resultados después de cada modificación, semillas fijas, dataset versionado) que esta investigación estableció.

Para trabajos académicos relacionados. Definir la formulación del problema antes que el algoritmo (la diferencia entre los Experimentos A y B fue mayor a la diferencia entre los ocho algoritmos analizados); respaldar cada métrica de su protocolo (significancia estadística, validación cruzada estratificada) y examinar el software con la misma rigurosidad que se hace con el modelo: las pruebas funcionales de esta investigación encontraron fallos que ninguna métrica de clasificación habría mostrado.

Para el sector de la electricidad en Ecuador. El prototipo evidencia que es posible contar con una herramienta ciudadana de comparación entre lo que se espera consumir y lo que se factura, con costos operativos bajos (inferencias de milisegundos en infraestructura gratuita). Es aconsejable investigar pilotos con distribuidoras que brinden datos anónimos reales, lo cual es un requisito para superar el principal obstáculo del enfoque sintético. También se debe evaluar la incorporación de estas herramientas en los canales de atención al usuario para ayudar a identificar fallos e ineficiencias de medición desde temprano.

En resumen: el camino desde un prototipo funcional con fallos ocultos hasta un sistema verificado, reproducible y con hoja de ruta muestra la conclusión principal de este trabajo: en el aprendizaje automático aplicado, la calidad del resultado no está solo en el modelo, sino en la rigurosidad del proceso entero que lo envuelve: los datos, el protocolo de evaluación, el software que lo soporta y las pruebas que lo avalan.

Referencias Bibliográficas

- A. Boden, M. (2016). *Inteligencia Artificial*. Turner Publicaciones.
- Arias, Á., & Durango, A. (2016). *Ingeniería y Arquitectura del Software*. CreateSpace Independet Publishing Platform.
- Barrios Arce, J. (26 de julio de 2019). *Health Big Data*. La matriz de confusión y sus métricas: <https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/>
- Bensalah, M., & Abdellatif, C. (2025). *Conjunto de datos de carga de alta resolución de contadores inteligentes en varias ciudades de Marruecos [Conjunto de datos]*. <https://doi.org/10.24432/C5FC8R>.
- Bouza-Herrera, C. (31 de mayo de 2021). LAS CURVAS ROC TEORÍA Y HERRAMIENTAS PARA SU USO. La Habana, Cuba: Universidad de la Habana.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.
- Breiman, L. (2001). Random Forests. *Kluwer Academic Publishers*, 28.
- Campo Yepes, J. (27 de Enero de 2026). Evaluación de métricas en modelos predictivos de clasificación en Machine Learning. *Facultad de Ingeniería - Maestría en Ciencia de Datos*. Bogotá, Colombia: Universidad EAN.
- Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895–1923. <https://doi.org/10.1162/089976698300017197>
- 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>

- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Espinoza Zuñiga, J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Scielo*, 1. <https://doi.org/https://doi.org/10.22201/fi.25940732e.2020.21.3.022>.
- Feng, L., Xu,, S., Zhang,, L., Wu,, J., Zhang,, J., Chu,, C., . . . Shi, H. (2020). Anomaly detection for electricity consumption in cloud computing: framework, methods, applications, and challenges. *EURASIP Journal on Wireless Communications and Networking*, 12. <https://doi.org/https://doi.org/10.1186/s13638-020-01807-0>
- geeksforgeeks. (diciembre de 23 de 2025). *geeksforgeeks*. Algoritmo de bosque aleatorio en aprendizaje automático: <https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/>
- Ghanim, J., & Awad, M. (2024). Detección de anomalías no supervisada en el consumo de electricidad mediante aprendizaje por refuerzo y un marco basado en bosques de series temporales. *Department of Electrical and Computer Engineering, American University of Beirut, Beirut, Lebanon*, 24. <https://doi.org/10.48550/arXiv.2501.00107>
- Gómez Palomo, S., & Moraleda Gil, E. (2020). *Aproximación a la ingeniería del software*. Centro de Estudios Ramón Areces.
- Hebrail, G., & Berard, A. (29 de agosto de 2012). *UC Irvine Machine Learning Repository*. Individual Household Electric Power Consumption: <https://doi.org/10.24432/C58K54>.
- Hebrail, J., & Berard, A. (2006). *Consumo individual de energía eléctrica en hogares [Conjunto de datos]*. <https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption>
- Himeur , Y., Ghanem, K., Alsalemi, A., Bensaali , F., & Amira , A. (2021). Detección de anomalías en el consumo energético de edificios mediante inteligencia artificial: una

- revisión, tendencias actuales y nuevas perspectivas. *ELSEVIER*, 287, 26.
<https://doi.org/https://doi.org/10.1016/j.apenergy.2021.116601>
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366.
[https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
- INEC. (2025). *INEC. ENEMDU anual*: <https://www.ecuadorencifras.gob.ec/enemdu-anual/>
- Jain, A. (13 de septiembre de 2024). *Medium*. Todo sobre Random Forest:
<https://medium.com/@abhishekjainindore24/everything-about-random-forest-90c106d63989>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. En *Advances in Neural Information Processing Systems 30* (pp. 3146–3154).
- Kaufman, S., Rosset, S., & Perlich, C. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 1–21.
<https://doi.org/10.1145/2382577.2382579>
- Jokar, P., Arianpoo, N., & Leung, V. C. M. (2016). Electricity theft detection in AMI using customers' consumption patterns. *IEEE Transactions on Smart Grid*, 7(1), 216–226.
<https://doi.org/10.1109/TSG.2015.2425222>
- Kelly, J. (2017). *Ukerc*. UK Domestic Appliance Level Electricity(UK-DALE-2017)-Disaggregated appliance/whole house power: <https://ukerc.rl.ac.uk/cgi-bin/dataDiscover.pl?Action=detail&dataid=7d78f943-f9fe-413b-af52-1816f9d968b0>
- Kelly, J., & Knottenbelt, W. (2015). El conjunto de datos UK-DALE, la demanda de electricidad a nivel de electrodomésticos y la demanda total de la vivienda en cinco hogares del Reino Unido. *Scientific Data*, 14. <https://doi.org/10.1038/sdata.2015.7> (2015)
- Kelly, J., & College, I. (2017). *Electricidad a nivel de electrodomésticos en el Reino Unido (UK-DALE-2017): consumo desagregado de electrodomésticos/energía total de la vivienda [Conjunto de datos]*. <https://ukerc.rl.ac.uk/cgi-bin/dataDiscover.pl?Action=detail&dataid=7d78f943-f9fe-413b-af52-1816f9d968b0>

- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. En *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 413–422). <https://doi.org/10.1109/ICDM.2008.17>
- Madrid, U. C. (2021). *¿Qué es el Machine Learning?* Aprendizaje automático de las máquinas: <https://www.masterdatascienceucm.com/que-es-machine-learning/>
- MZH, J., J, H., & Aziz, A. (2021). Detección de anomalías no supervisada para el consumo de energía en series temporales mediante un enfoque de agrupamiento. *EmergingScience Journal*, 5(6), 15. <https://doi.org/10.28991/esj-2021-01314>
- Pang, G., Shen, C., Cao, L., & Hengel, A. (2020). Aprendizaje profundo para la detección de anomalías: una revisión. 36. <https://doi.org/https://arxiv.org/abs/2007.02500>
- Renovables-ARCERNNR, A. d. (2023). INEC. Estadística del Sector Eléctrico Ecuatoriano 2023: <https://anda.inec.gob.ec/anda5/index.php/catalog/1080/get-microdata>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Schmidhuber, J. (2003). Long Short-Term Memory. *Juergen Schmidhuber*, 60.
- Sistemas Inteligentes. (2021). *Bosques Aleatorios - Random Forest. Cómo funcionan y código en Python*. <https://www.youtube.com/watch?v=jlJ4uKS9D5A>
- Trindade, A. (2015). *ElectricityLoadDiagrams20112014 [Conjunto de datos]*. Repositorio de aprendizaje automático de la UCI. . <https://doi.org/10.24432/C58C86>
- Vega, M. (octubre de 2010). Casos de Uso UML. Granada, España: Universidad de Granada.
- Xu, J., Wu, H., Wang, J., & Long, M. (2022). Anomaly Transformer: Detección de anomalías en series temporales con discrepancia de asociación. 20. <https://doi.org/https://arxiv.org/abs/2110.02642>
- Zhang, J., Zhang, H., Ding, S., & Zhang, X. (2021). Predicción del consumo de energía y detección de anomalías basadas en Transformer y K-Means. *Frontiers in Energy Research*, 9, 8. <https://doi.org/10.3389/fenrg.2021.779587>

Zheng, Z., Yang, Y., Niu, X., Dai, H.-N., & Zhou, Y. (2018). Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transactions on Industrial Informatics*, 14(4), 1606–1615. <https://doi.org/10.1109/TII.2017.2785963>

Apéndices A: Enlace del sistema de detección de anomalías energéticas

<https://huggingface.co/spaces/eusca/energia-detector>