Maestría en

# CIENCIA DE DATOS Y MAQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

**Trabajo previo a la obtención de título de Magister en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial.**


**AUTORES:**

Defaz Toapanta Verónica Elizabeth

Ortiz Velasco Lourdes Virginia

Reyes Pérez Joseph Emmanuel

Núñez Jácome Roberth Steven

**TUTORES:**

Alejandro Cortes

Paulina Vizcaíno


**TEMA**

## Predicción de consumo de energía eléctrica en hogares utilizando modelos de aprendizaje automático.


Quito - Ecuador


REINVENTEMOS EL FUTURO

**Certificación de autoría**

Nosotros, Defaz Toapanta Verónica Elizabeth, Núñez Jácome Roberth Steven, Ortiz Velasco Lourdes Virginia, Reyes Pérez Joseph Emmanuel**,** 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.

---

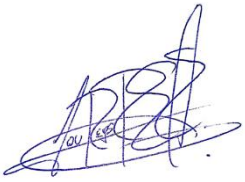
**Firma del graduando**
**Defaz Toapanta Verónica Elizabeth**

---

**Firma del graduando**
**Núñez Jácome Roberth Steven**

---

**Firma del graduando**
**Ortiz Velasco Lourdes Virginia**



---

**Firma del graduando**
**Reyes Pérez Joseph Emmanuel**

**Autorización de Derechos de Propiedad Intelectual**

Nosotros, **Defaz Toapanta Verónica Elizabeth, Núñez Jácome Roberth Steven, Ortiz Velasco Lourdes Virginia, Reyes Pérez Joseph Emmanuel** en calidad de autores del trabajo de investigación titulado ***Titulo del trabajo de investigación Predicción de consumo de energía eléctrica en hogares utilizando modelos de aprendizaje automático.*** , 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, julio 2025

---

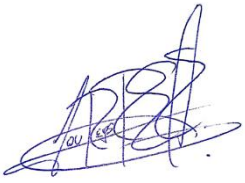
**Firma del graduando**
**Defaz Toapanta Verónica Elizabeth**

---

**Firma del graduando**
**Núñez Jácome Roberth Steven**

---

**Firma del graduando**
**Ortiz Velasco Lourdes Virginia**

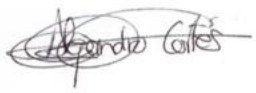


---

**Firma del graduando**
**Reyes Pérez Joseph Emmanuel**

**Aprobación de dirección y coordinación del programa**

Nosotros, **Andrés Cortes e Iván Reyes,** declaramos que los graduandos: **Defaz Toapanta Verónica Elizabeth, Núñez Jácome Roberth Steven, Ortiz Velasco Lourdes Virginia, Reyes Pérez Joseph Emmanuel** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



---

Alejandro Cortes
Director/a de la
Maestría en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

---

Iván Reyes
Coordinador/a de la
Maestría en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

# DEDICATORIA

Con fe, gratitud y alegría cerramos este ciclo, un año de esfuerzo y sacrificio, por tal motivo nos honra dedicar este proyecto a todas las personas que estuvieron a nuestro lado y de alguna forma nos incentivaron a seguir adelante y culminar con éxito.

Verónica, Lourdes, Joseph Steven.

# AGRADECIMIENTOS

” Lo mejor del trabajo en equipo es que siempre tienes alguien a tu lado”, Margaret Carty. Expresamos nuestro agradecimiento a todos los Docentes que fueron parte de este proceso, quienes fortalecieron nuestros conocimientos con sus enseñanzas y experiencias. Además, nos sentimos agradecidos con nosotros mismos por la dedicación, compromiso y el aporte profesional que cada uno brindó en el desarrollo del proyecto.

Verónica, Lourdes, Joseph Steven.

# RESUMEN

En la actualidad la demanda del servicio energético es alta, envista que actualmente existe diversos entornos que es indispensable dicho servicio y factores que afectan el consumo energético. Según el Balance Energético Nacional del 2019, el consumo eléctrico en el Ecuador, se ha incremento del 39.4% por habitante. El problema radica en que las empresas de energía eléctrica no saben como predecir el consumo, para determinar estrategias y no exista sobre carga.

Ante esta situación, se ha planteado el proyecto de investigación Predicción de consumo de energía eléctrica en hogares utilizando modelos de aprendizaje automático, para cumplir con el objetivo de determinar un modelo de aprendizaje automático para predecir el consumo energético en hogares del Ecuador en base a datos históricos. Se trabajó con un dataset de consumo de energía del año 2023, obtenida de la base de datos de la Agencia de Regulación y Control de Electricidad (ARCONEL), Se aplicó la metodología CRISP-DM, para estructurar por fases el proyecto, y determinar de mejor manera el proceso de datos y la selección de los modelos. En este caso se realizó el entrenamiento de tres modelos Regresión Lineal, Randon Forest y XGBoost. Se obtuvo como resultados que el modelo que más predomino y se obtuvo una predicción más exacta es el Random Forest que alcanzando un $R^2$ superior a 0.99. Se puede concluir que la variable predominante fue el numero de clientes, lo que significa que su incremento es un causal del incremento disminución del consumo de energía en los hogares.

**Palabras Claves:** Consumo de energía, Machine learning, Regresión Lineal, Randon Forest y XGBoost, variable predominante, $R^2$

## ABSTRACT

At present, the demand for energy service is high, since currently there are various environments that such a service is indispensable and factors that affect energy consumption. According to the 2019 National Energy Balance, electricity consumption in Ecuador has increased by 39.4% per inhabitant. The problem is that electric power companies do not know how to predict the consumption, to determine strategies and there is no overload.

Faced with this situation, the research project Prediction of Electricity consumption in homes using learning models to meet the objective of determining a learning model automatic to predict energy consumption in homes in Ecuador based on Historical data.

We worked with a dataset of energy consumption for the year 2023, obtained from the database of the Electricity Regulation and Control Agency (ARCONEL), The CRISP-DM methodology was applied to structure the project, and better determine the data processing and model selection.

The results were that the most predominant model and a The most accurate prediction is the Random Forest that reaching an $R^2$ greater than 0.99. Herself can conclude that the predominant variable was the number of customers, which means that its increase is a cause of the increase decrease in the consumption of energy in homes.


**Keywords:** Energy consumption, Machine learning, Linear Regression, Randon Forest and XGBoost, predominant variable, $R^2$

# TABLA DE CONTENIDOS (Índice)

## LISTA DE TABLAS (Índice de tablas)

# LISTA DE FIGURAS (Índice de figuras)

# Predicción de consumo de energía eléctrica en hogares utilizando modelos de aprendizaje automático.

## CAPITULO 1: INTRODUCCIÓN

### 1.1.Definición del proyecto

La energía eléctrica es un recurso esencial en la actividad diaria de los hogares, permitiendo el funcionamiento de los dispositivos y servicios dentro de los domicilios, así como la iluminación, calefacción y electrodomésticos, como refrigeradores, lavadoras y sistemas de diversión, la electricidad desempeña un papel importante para el confort y la calidad de vida de los usuarios.

El consumo de energía mundial en 2023 ha sido el 2.2%, superando la media de 2010-2019 en el 1.5% anual, este aumento fue impulsado por los países del BRICS (Brasil, Rusia, India, China y Sudáfrica), principalmente China con el 6.6%, le sigue India con el 5.1% y Brasil con el 3.3%, en cambio Rusia con 0.3% y Sudáfrica con -1.2% su crecimiento fue nulo o negativo, por lo tanto el consumo de energía cayó en los países de la OCDE **(**Organización para la Cooperación y el Desarrollo Económicos**)** con una disminución valiosa en la Unión Europea de -4.2%, especialmente en Alemania con el -9.3%, Japón con el 3.5% y Corea del Sur con -2.8%, mientras que en EEUU se conservó estable debido a un consumo menor de electricidad (Enerdata, 2024)

En 2024, el consumo eléctrico en los hogares españoles aumentó un 5,1%**,** mientras que el gasto en electricidad creció hasta un 10,5%**.** Además, un 35,6% de los hogares confirmó haber contratado un suministro eléctrico con una tarifa única, con el mismo precio de electricidad durante todas las horas del día, esta información procede de una encuesta

ejecutada en el segundo trimestre de 2024, en la que participaron 5.307 hogares y 9.016 usuarios**.** (Comisión Nacional de los Mercados y la Competencia, 2024)

De acuerdo con el Balance Energético Nacional del 2019, el consumo eléctrico por habitante en el Ecuador ha mantenido un crecimiento entre 2009 y 2019 con un incremento del 39.4% avanzando de 1.088 kWh por habitante a 1.517 kWh por habitante. El Instituto de Investigación Geológico y Energético, entre 2018 y 2019 el consumo se registró un incremento del 2%, al subir de 1.488 kWh por habitante a 1.517 kWh por habitante. (Instituto de Investigación Geológico y Energético , 2020)

El consumo de energía eléctrica en los hogares es un componente esencial, con el crecimiento de la demanda eléctrica y la necesidad de mejorar su uso, la predicción de consumo de energía eléctrica en hogares se ha transformado en una herramienta importante para fomentar la eficiencia, reducir precios y minimizar el impacto ambiental.

Según Meneses, et al (2021), el aprendizaje automático (Machine Learning), permite desarrollar modelos idóneos de predecir el consumo energético basándose en datos históricos, y patrones de consumo, aplicando algoritmos como Regresión Lineal, Random Forest, XGBoost son usualmente utilizados para esta actividad.

El proyecto de investigación se enfoca en aplicar los modelos de Machine Learning que permita predecir el consumo de energía en los hogares y anticiparse a cambios en la demanda, para optimizar el uso de energía y fomentar estrategias de consumo más razonables y a su vez determinar qué tipos de tecnologías pudiesen minimizar el consumo energético para la transformación eficiente y responsable de la electricidad.

## 1.2. Justificación e importancia del trabajo de investigación

En los últimos años se ha incrementado el consumo energético en los hogares ecuatorianos debido a factores como: el crecimiento de la población, urbanizaciones y el uso de varios electrodomésticos en los hogares. La planificación energética representa un desafío, ya que no existe predicciones exactas que permitan anticipar la demanda y optimizar el uso de los recursos eléctricos. Este proyecto se enfoca en el entrenamiento de modelos de aprendizaje automático para predecir el consumo energético en hogares del Ecuador, aplicando herramientas analíticas con énfasis en la inteligencia artificial para mejorar la administración de la demanda eléctrica.

Actualmente, en el Ecuador el tema de ciencia de datos es muy incierta, tienen diferentes conceptualizaciones de lo que es la inteligencia artificial, peor aún, está en proceso de subdesarrollo la analítica de datos, no dan la importancia de la que los datos o información es un activo muy valioso que con una buena predicción se puede tener grandes beneficios. La planificación energética depende de evaluaciones generales que no consideran patrones específicos de consumo, generando ineficiencias en la distribución y aumento de riesgo de sobrecargas en la red eléctrica. Además, los hogares carecen de información personalizada para optimizar su consumo y por ende reducir costos.

Este estudio aborda la necesidad de contar con un sistema de predicción que permita mejorar la planificación energética, anticipando picos de consumo y distribuyendo los recursos de manera eficiente. Además, con la predicción a tiempo se puede optimizar el uso de la electricidad en los hogares, otorgando recomendaciones basadas en datos. Lo que induce al cuidado del medio ambiente con un consumo más eficiente y sostenible.

Los beneficios del proyecto estarían sujetos para el sector energético con predicciones más precisas que permitan anticipar y evitar sobrecargas en la red. La optimización de estrategias de generación y distribución de electricidad. Con esto se reduciría los costos operativos al mejorar la eficiencia en la planificación de la demanda.

Los consumidores de los hogares, tendrían acceso a información detallada sobre su consumo energético, otorgando recomendaciones personalizadas para optimizar el uso de electrodomésticos y reducir costos. Induciendo a ser consientes sobre eficiencia energética y reducción del impacto ambiental.

También, tendrá un aporte en el aspecto investigativo al determinar un modelo el mismo puede servir como referencia para futuros estudios relacionados al tema, con información detalla. El aprendizaje automático, identificará patrones de consumo energético en diferentes regiones del Ecuador, evaluar factores como el clima, determinar qué electrodomésticos tienen mayor impacto en el consumo total, finalmente, se presentaría proyecciones de consumo futuro con base en datos históricos y con información verídica.

La energía representa un recurso fundamental, y su uso eficiente influye directamente en la calidad de vida de las personas y en la economía del país. Aunque los modelos de aprendizaje automático han demostrado ser herramientas eficaces para predecir tendencias y potenciar recursos en diferentes sectores, su aplicación en la gestión del consumo energético en Ecuador aún es limitada. La eficiencia energética se erige como un pilar crucial en la lucha contra el cambio climático, y este proyecto tiene como objetivo contribuir al desarrollo de estrategias sostenibles para el uso de la electricidad. Así, no solo se avanzará en el conocimiento del consumo energético, sino que también se logrará un impacto práctico al mejorar la eficiencia y sostenibilidad en el uso de la electricidad en Ecuador. A través de la implementación de modelos de aprendizaje automático, se generarán predicciones precisas y recomendaciones valiosas para una variedad de actores, desde empresas eléctricas hasta consumidores individuales. De este modo, la investigación no solo llenará un vacío teórico, sino que también ofrecerá soluciones concretas para la optimización de los recursos energéticos en el país.

## 1.3.Alcance

El consumo energético en los hogares es un aspecto relevante en la gestión de recursos en Ecuador. En la actualidad existe una creciente demanda del consumo energético, el mismo que ha ido presentando diversos problemas en la eficiencia del servicio, lo que exige a este sector buscar alternativas para mejorar el servicio eléctrico ante la predicción que les permita tomar decisiones adecuadas. Sin embargo, la falta de estudios que analicen el consumo energético en el país limita la capacidad de optimizar políticas y estrategias de ahorro energético. La tecnología conjuntamente con la inteligencia artificial ha desarrollado herramientas predictivas en base de modelos de aprendizaje automático machine learning, dichos modelos serán analizados, evaluados en este proyecto.

## 1.4.Objetivos

### *1.4.1. Objetivo general*

Determinar un modelo de aprendizaje automático para predecir el consumo energético en hogares del Ecuador en base a datos históricos.

### *1.4.2. Objetivo especifico*

- Identificar fuentes de datos para seleccionar el data set.
- Realizar el preprocesamiento de datos en el conjunto de datos históricos del consumo energético.
- Analizar el rendimiento de los diferentes modelos de aprendizaje automático para la predicción del consumo energético en los hogares, mediante métricas de evaluación.

# CAPITULO 2: REVISIÓN DE LITERATURA

## 2.1. Estado del Arte

Chicaiza et al., Previsión del consumo eléctrico en el cantón Salcedo mediante técnicas de aprendizaje automático, Revista ODIGOS, 5(1), Ecuador, febrero-mayo 2024, pp. 9-24

La finalidad de esta investigación fue analizar el desempeño de las técnicas de Random Forest (RF) y XGBoost en la predicción del consumo eléctrico en el cantón Salcedo Ecuador, utilizando información obtenida desde enero de 2017 hasta diciembre de 2022. La perspectiva principal estuvo en los usuarios residenciales de seis parroquias del cantón Salcedo.

Para ello aplicaron la metodología de la Estructura de Descomposición del Trabajo (EDT), una herramienta que ayuda la gestión de proyectos de manera ordenada, clara, controlada promoviendo una comunicación eficaz, En este contexto, se llevaron a cabo las fases de recopilación, y preprocesamiento de datos, así como el entrenamiento, ajuste y evaluación de modelos.

Recopilaron reportes mensuales pertenecientes a clientes regulados, y crearon una matriz multidimensional, con 432 muestras y 6 características: año, mes, parroquia, número de clientes, energía consumida y facturación.

Los algoritmos de Random Forest y XGBoost, demostraron un impacto significativo en la precisión y robustez de las predicciones. Tras entrenar y evaluar ambos modelos, determinaron que Random Forest presentó un mejor desempeño en términos de RMSE y MAPE en comparación con XGBoost para todas las parroquias evaluadas en el año 2022. Esta conclusión es relevante para la organización y planificación del consumo de energía eléctrica, ayudando a prevenir posibles situaciones de emergencia.

Galvis Plata, Jorman Hernando (2022), Modelo predictivo de consumos de energía eléctrica aplicando redes neuronales artificiales, San José De Cúcuta, Colombia: Universidad Francisco De Paula Santander.

El objetivo de estudió fue la predicción de consumo de energía eléctrica utilizando redes neuronales artificiales (RNA), en donde representó la aplicación de estas redes para modelar una variable importante en el análisis de eficiencia energética. Su propuesta se basó en aprendizaje automático, redes retroalimentadas de dos capas y aprendizaje profundo.

Las redes neuronales son una plataforma tecnológica eficaz para el modelo predictivo basándose en la información recopilada para generar resultados con un pequeño margen de error, con el propósito de aproximar las predicciones a los datos reales.

Los datos históricos de consumo eléctrico utilizados en esta investigación fueron recopilados por medio del Servicio Nacional de Aprendizaje (SENA) en la ciudad de Cúcuta. El elemento de estudio fue el edificio Administrativo del barrio Pescadero. El muestreo de datos se realizó con un analizador trifásico entre el 15 de febrero y el 16 de marzo del 2022, calculando hasta 40 características, en un tiempo de 60 segundos, con un total de 41877 datos por elementos.

Durante el proceso de entrenamiento, compararon diferentes modelos de redes neuronales artificiales. Los modelos seleccionados fueron:

1. Modelo LSTM 250-200, que hace referencia a una red profunda con una arquitectura LSTM (Long Short-Term Memory) que posee dos capas ocultas. Los números 250 y 200 señalan la cantidad de neuronas en cada una de estas capas. Este modelo presentó los mejores comportamientos.
2. Modelo Feed-Forward, con capas ocultas de hasta 100 neuronas, mostró un mejoramiento en la precisión a medida que aumentaba el número de neuronas. No

obstante, su tiempo de entrenamiento superó los 30 minutos, lo que implicó una mayor demanda de recursos computacionales para lograr una precisión aceptable.

De acuerdo con los resultados conseguidos, concluyeron que el modelo de regresión neuronal artificial es el más adecuado para predecir el consumo total de energía eléctrica del edificio administrativos-SENA. La regresión lineal máxima alcanzada fue de 0.99194 empleando redes neuronales artificiales. La arquitectura LSTM con 250 neuronas para la capa oculta y 200 para épocas, logró un tiempo de entrenamiento de 3 minutos y 27 segundos, resultando en una progresión lineal de 0.98286 y 0.98713 lo que demostró ser la más estable y próxima a 1.

Torres Gómez, Alfredo Gerardo (2022), Análisis del consumo de la energía eléctrica en la ciudadela La Rioja Etapa Almería con uso de Machine Learning, Guayaquil, Ecuador: Universidad Politécnica Salesiana Sede Guayaquil.

Investigaron el consumo de la energía eléctrica en los hogares de la ciudadela la Rioja utilizando procedimientos de Machine Learning. En un estudio realizado del 15 de junio al 15 de julio de 2021 en la ciudadela La Rioja, Etapa Almería, se tomó como muestra a 49 encuestados, lo que representa el 28% de un total de 170 residentes, mediante una encuesta de 12 preguntas, se recopiló información importante sobre el consumo energético en los hogares. Esta indagación fue utilizada para aplicar técnicas de Machine Learning con el propósito de predecir el consumo de energía eléctrica y evaluar el rendimiento de equipos eléctricos y electrodomésticos.

El consumo de energía eléctrica es indispensable en los hogares y depende de componentes como el número de personas que habitan en la vivienda, la cantidad de equipos electrónicos y electrodomésticos, así como el uso que le dan. Entre los elementos asociados al consumo de energía eléctrica que se incluyeron en el estudio fueron: números de personas en la familia, calidad de servicio, número de electrodomésticos, facturación de planilla.

Los principales resultados obtenidos indican que el 36% de los residentes afirmaron haber incrementado en el valor de la planilla eléctrica del año 2020 al 2021 y el 59.2% manifestó desconocer el proceso mediante el cual se obtienen los datos de consumo reflejados en las planillas de luz. Se identificó en promedio en los hogares que cuentan con cinco equipos electrónicos y electrodomésticos. Finalmente, el 57.1% encuestados confirmaron que, a pesar de no contar con suministro eléctrico en ciertos momentos, ese tiempo igualmente es cobrado.

El desconocimiento sobre la obtención de los datos mostrados en las planillas eléctricas es una preocupación frecuente. Los datos de consumo y valores a pagar en las planillas están basados en mediciones directas mensuales relacionado a periodos de lectura entre 28 y 33 días. A pesar de este método, los residentes realizan el pago mensual de sus consumos, aunque se enfrentan a deficiencias en la calidad de servicio. Además, se identificó que la mayoría de los hogares cuentan con una cantidad justificable de equipos eléctricos y electrodomésticos, los cuales influyen en su consumo energético.

Yajure-Ramírez, César. Uso de algoritmos de aprendizaje automático para analizar datos de energía eléctrica facturada. Caso: Chile 2015 – 2021, Revista I+D Tecnológico, 18 (2), Venezuela, 2022, pp. 17-31

El propósito de esta investigación fue utilizar algoritmos de aprendizaje automático para analizar los datos de energía eléctrica facturada, específicamente los datos de energía eléctrica facturada mensualmente a los clientes regulados de Chile, entre 2015 y 2021, aplicando modelos de aprendizaje supervisado y no supervisado. Utilizaron varios algoritmos como son K-Means (K-Medias), K-NN (K-Vecinos más cercanos), PCA (Análisis de Componentes Principales).

La recolección de información se realizó mediante el análisis exploratorio de los datos y la modelación de los mismos. Estos datos fueron extraídos desde la plataforma Energía

Abierta de la Comisión Nacional de Energía de Chile, correspondiendo al consumo eléctrico mensual de usuarios regulados entre 2015 y 2021. El conjunto de los datos contiene 338,652 registros y 10 campos como año, mes, región, comuna, tipo de cliente y tipo de tarifa. También se registraron el número de clientes abastecidos y tres medidas de consumo eléctrico en kWh., cada fila representa un lote de energía retirado de una subestación por una empresa distribuidora donde el lote abastece a un grupo de clientes con la misma tarifa, región y comuna. Estos datos son útiles para análisis de eficiencia energética, segmentación de usuarios y modelado predictivo.

Los resultados obtenidos son los siguientes:

1. **La limpieza y preparación de los datos:** verificaron y corrigieron el formato de los datos para asegurar que fueran correctos, sin datos faltantes ni duplicados. Detectaron 25 datos faltantes a 13 filas lo que representa el 0.004% del total de filas, y estos fueron eliminados. Además, comprobaron y eliminaron filas duplicadas dejando un total de 338,638 filas.

2. **Análisis exploratorio de los datos**: determinaron las tarifas para clientes residenciales y no residenciales. Observaron que el consumo promedio mensual de energía eléctrica fue mayor durante los meses de junio, julio y agosto, alcanzando su punto máximo de facturación en julio. En contraste, los meses con menor facturación se registraron entre noviembre y marzo, siendo febrero el mes con el consumo y facturación más bajos durante el período de estudio. Determinaron que el consumo de los clientes no residenciales se mantuvo constante, mientras que la variación mensual en el consumo de energía se presentó en los clientes residenciales.

3. **Aplicación de algoritmos de aprendizaje automático**: utilizaron diferentes técnicas, como el algoritmo de agrupamiento K-Means para detectar patrones dentro del conjunto de datos, el algoritmo de predicción K-NN, para clasificar nuevos datos

incorporados al conjunto y el análisis de componentes principales para identificar las variables del conjunto de datos original.

Concluyeron que el 96% de los clientes son de tipo residencial, quienes consumieron más del 50% de la energía total facturada, lo que resultó en un consumo unitario más bajo en comparación con los clientes no residenciales. Al aplicar el algoritmo K-Means se determinó que el valor óptimo de clústers es diez, y estos se agrupan según el tipo de cliente. Posteriormente, al aplicar el algoritmo K-NN, lograron desarrollar un modelo capaz de predecir el tipo de cliente, con una precisión del 98%. Finalmente, el análisis de componentes principales reveló que las variables que mejor explican los datos son el año, el mes y el tipo de cliente, siendo este último la variable más significativa.

## 2.2. Marco Teórico

### *2.2.1. Demanda de energía eléctrica*

La demanda eléctrica es la suma de los consumos individuales de los diferentes clientes pertenecientes al sistema de distribución, los cuales, varían según diversos factores que caracterizan dicho consumo. Lozada, et al (2022).

En el futuro la electricidad sustituirá otras fuentes de energía y se convertirá en la principal de uso en los hogares, empresas y el transporte. La electricidad se está convirtiendo en un aspecto importante de nuestra vida diaria. Es un requisito básico en la sociedad actual, el consumo de energía se incrementa rápidamente.

### 1.2.2. *Consumo de energía eléctrica*

El consumo de electricidad se ha convertido en un tema de inmensa importancia. El creciente interés en el área se ha desencadenado en gran medida por la creciente demanda de

energía en todo el mundo alimentada principalmente al aumentar las actividades económicas, especialmente en los países emergentes. (Mejía Vásquez & Gonzalez Chavez, 2019)

El sector eléctrico depende de la previsión del suministro de energía, es una base principal para la toma de decisiones para definir el diseño y la operación de los sistemas de energía. En función de la demanda, las empresas eléctricas utilizan técnicas que pueden aplicarse a corto, medio y largo plazo. Este dinamismo del consumo energético, hace que las técnicas de predicción comunes son insuficientes, por lo que, requiere el uso de estrategias más avanzadas como interactuar con la Inteligencia artificial aplicando modelos de predicción. El consumo de energía sería factible predecir de forma correctamente por el gran impacto que tiene en muchos aspectos operativos.

### 1.2.3. *Predicción de la demanda de la energía eléctrica*

Es importante el predecir el consumo de electricidad para una gestión efectiva en su generación y suministro. Predecir el consumo de electricidad es crucial en la planificación, análisis y operación de sistemas de energía para asegurar un suministro ininterrumpido, confiable, seguro y económico de electricidad. Las empresas eléctricas tienen que utilizar modelos de predicción de carga para garantizar que la energía suministrada satisfaga la carga de sus clientes más la energía perdida en el sistema (Mejía Vásquez & Gonzalez Chavez, 2019)

Alcanzar una predicción precisa del consumo energético es un desafío, especialmente en la categoría residencial, donde el consumo mensual cambia de forma drástica. La predicción efectiva de la demanda de consumo de energía es una forma de evitar el desperdicio de energía. Es difícil predecir el consumo de energía debido a los diferentes factores como el clima, uso de electrodomésticos, aspectos geográficos, etc. Los modelos de machine learning pueden enfocarse a obtener con un alto porcentaje de exactitud la

predicción para mejorar el consumo energético en los hogares y optimizar el uso de recursos eléctricos, la predicción que el presente proyecto se enfoca es un análisis del consumo para realizar las predicciones por parroquias.

### *1.2.4. Fundamentos del aprendizaje automático Machine Learnig(ML)*

Según Forero & Negre (2024), Machine Learning (ML), es una rama de la inteligencia artificial (IA), conocido como aprendizaje automático o aprendizaje de máquina, que ha tenido un impacto de relevancia en los últimos años. En la última década, los científicos enfocan su atención a las herramientas enriquecidas con tecnología inteligente, ya que tienen el potencial de revolucionar los procesos de diversos sectores.

El ML nace por la década de los cincuenta como una herramienta para emular, computacionalmente elementos del proceso cognitivo humano por medio del reconocimiento de patrones y procesos de toma de decisiones. Los algoritmos de ML pueden ser clasificados en: supervisados, no supervisados y por refuerzos. La principal diferencia que se caracteriza entre estos tipos de algoritmos es la presencia o ausencia de una variable etiquetadas Pedrero,et. al (2021)

#### 2.2.4.1. Tipos de aprendizaje automático

**Aprendizaje supervisado**

El aprendizaje supervisado es un método para desarrollar un modelo de aprendizaje automático mediante un conjunto de datos etiquetados. En este proceso, en cada punto de datos del conjunto se asocia con un resultado previsto conocido. Los modelos se entrenan para predecir los resultados (Khaoula, Amine, Mostafa, & Deifalla, 2024)

Amézquita & Eslava (2022), de igual forma dan a conocer que las técnicas de aprendizaje automático supervisado, implica en entrenar un modelo que operar con un conjunto de características y predice una etiqueta en base a un conjunto de datos que incluye valores de etiqueta ya distinguidos. Con el proceso de entrenamiento ajusta las características a las etiquetas conocidas que permite definir una función general que se puede aplicar a nuevas características en las cuales no se conoce las etiquetas. Así, tenemos la función, donde $y$ representa la etiqueta que queremos predecir y $x$ representa las características que el modelo utiliza para predecir.

$$y = f(x)$$

Donde:

**x**: Vector de característica o variable independiente

**y**: Etiqueta o variable dependiente que se va a predecir

**f(x)**: Función que estima la relación entre x e y

La mayoría de los casos, $x$ es viene hacer un vector que consigna múltiples valores de características; para ser precisos, las funciones podrían expresarse así:

$$y = f([x1, cx2, x3 \dots])$$

El entrenamiento de un modelo se podría decir que es un enfoque en encontrar una función que f(x) que generalice correctamente y ejecute cálculos con los valores de $x$ genere el resultado $y$. Es decir, que no debe funcionar bien solo con los datos de entrenamiento, sino que pueda predecir los resultados de nuevos datos. Este proceso se puede dar al aplicar un algoritmo de aprendizaje automático que ajuste de alguna forma los valores como los modelos de regresión lineal, árboles de decisión, máquinas de soporte vectorial (SVM) y

redes neuronales, cada uno con ventajas y limitaciones dependiendo del tipo de problema a resolver.

**Aprendizaje no supervisado**

Otro aprendizaje es el no supervisado, Yazici, et al., (2023), dice que este tipo de aprendizaje se base en tareas que permite descubrir patrones y estructuras que se encuentran ocultas en los datos que carecen de etiquetas o salidas desconocidas. A su vez, a diferencia del aprendizaje supervisado, determina las similitudes o estructuras para explorar datos y aprender de muestra en diferentes grupos según las similitudes entre ellos. Sin embargo, el rendimiento es subjetivo y específico del dominio en el aprendizaje no supervisado que permitirá solventar los problemas o tareas como:

- agrupamiento o clustering: Segmenta los datos en grupo basado en similitudes
- reducción de dimensionalidad: Simplifica los datos manteniendo en lo posible la mayor parte de la información. y
- detección de anomalía: Identifica los puntos de los datos que desvían de forma significativa el comportamiento esperado.

También, Zhou (2022), señala que busca patrones no detectados previamente en un conjunto de datos sin etiquetas predefinidas, como por ejemplo la indagación de datos en el mercadeo, reconocimiento de patrones en una imagen, análisis de redes sociales, etc. El método de agrupamiento es de forma jerárquica, k-medias, agrupamiento en base a la densidad de aplicaciones con ruido (DBSCAN).

En el aprendizaje no supervisado tiene otro enfoque, donde los modelos se entrenan a sí mismos, es decir, no se necesita demasiada información sobre el resultado. El entrenamiento se proporciona con un conjunto de datos que no está clasificado, etiquetado ni

categorizado. En ese caso, los modelos de ML no están obligados a trabajar en ningún conjunto de datos bajo supervisión. Estos métodos facilitan el reconocimiento de características a partir de datos no etiquetados. Los métodos más comunes de aprendizaje no supervisado son K-medias, mezcla gaussiana y análisis de componentes principales (PCA) (Amor, Tayyab, Petru, & Neethu, 2023)

**Aprendizaje por refuerzo**

El aprendizaje por refuerzo (Reinforcement Learning, RL) presentan otro paradigma, aprendizaje por refuerzo no es generalizar situaciones no vistas durante el entrenamiento; en esencia, un agente aprende a interactuar a través de un proceso de prueba y error incorporando componentes para mejorar su desempeño (Montenegro, Menchaca, & Menchaca, 2023).

Los elementos clave del aprendizaje por refuerzo son:

- **Agente**: Toma decisiones en el entorno
- **Entorno**: El sistema con el que interactúa el agente
- **Acciones**: Las posibles decisiones que el agente puede tomar
- **Recompensas o Castigos**: Retroalimentación que guía el aprendizaje
- **Política**: Estrategia que define las acciones del agente en cada estado

El fundamento del aprendizaje por refuerzo es el comportamiento que se basa en estímulos de Pavlov y el desarrollo de la teoría del control ecuaciones de Bellman. Es un aprendizaje que modela la dinámica de problemas de toma de decisiones secuenciales por medio de procesos de decisión Markov (Markov Decision Process, MDP).

El aprendizaje de refuerzo es una técnica de aprendizaje que tiene a una tendencia de señales positivas recompensa y señales negativas castigo, en donde alguien está presente no

para guiar, sino que esta para penar las acciones incorrectas, así, como para recompensar las acciones correctas. Por lo tanto, el aprendizaje de refuerzo constituye cuatro elementos: crítico, entorno, recompensa/castigo y acción. El modelo se esfuerza por que las acciones obtengan la mayor recompensa posible. Dichos aprendizajes están orientados a la recompensa los diseños de juegos como: ajedrez, Atari y GO, aplicaciones de control, teoría de la información, navegación y robótica (Vaish, Dwivedi, Tewari, & Tripathi, 2021).

El aprendizaje por refuerzo está inspirado en el aprendizaje de los humanos y animales en base a experiencias, similar a la enseñanza de los niños acciones buenas tienes una recompensa caso contrario un castigo. Sumando a ello el marco de Procesos de Decisión de Markov (MDP), con el cual modela problemas donde las decisiones se toman de forma secuencial y los resultados depende de la acción y el estado actual.

ML en base a sus tres aprendizajes supervisado, no supervisado y por refuerzo incorpora enfoques adicionales que permiten solucionar diversos tipos de problemas o tareas según su entorno de los datos. El entender los principios, ventajas y limitaciones es esencial para seleccionar el modelo adecuado en la analítica de datos con inteligencia artificial. Lo que conlleva, a la correcta aplicación de estos algoritmos para mejorar la toma de decisiones, que a su vez permitirá descubrir patrones ocultos en grandes volúmenes de información. En nuestro caso para realizar la predicción del consumo de energía en los hogares aplicaremos el aprendizaje supervisado por contar con datos etiquetados.

#### 2.2.4.2. Fases del proceso de modelado

De acuerdo con Bonilla y Molina, et.al., (2024), al aplicar la metodología Cross-Industry Standard Process for Data Mining (CRISP-DM) es un método estructurado de seis fases que constituye un compartimiento de tareas estructuradas, que está diseñado para la

ejecución de proyectos de minería de datos, aprendizaje automático e inteligencia de negocios. CRISP-DM extrae los conocimientos relevantes e identifica patrones no evidentes en conjuntos de datos, que apoya así la toma de decisiones.

De manera preliminar, Brzozowska et al. (2023) afirman que, para obtener los resultados esperados, es indispensable organizar un proceso para recolección de datos, análisis, difusión de resultados y evaluación de la implementación del modelo planteado. El CRISP - DM es una metodología de análisis de datos que se orienta a la minería de datos a través de seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de estos, Fase de Modelado, Evaluación e Implementación. Es de gran relevancia implementar en los proyectos este modelo para alcanzar los objetivos. Generan estrategias y mejoras en cada evaluación de la planificación del modelo a través de la información obtenida, sin considerar los datos que se evalúen.

Según González (2023), para desarrollar los proyectos de Machine Learning, se puede aplicar la metodología de CRISP-DM (Cross-Industry Standard Process for Data Mining), una estructura aceptada por su enfoque flexible y adaptable al ciclo de vida un análisis de datos. De acuerdo a este contexto, propone un modelo de siete fases para ejecutar de forma sistemática el proceso de desarrollo en ML:

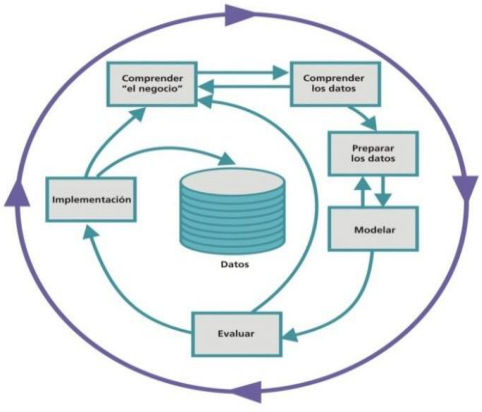
- **Comprensión del Problema:** Comprender a profundidad el problema a resolver, con la colaboración de expertos en la materia, dicho conocimiento es clave para contextualizar los datos y las posibles soluciones.
- **Comprensión de los datos:** Desde el inicio se debe establecer los criterios específicos para evaluar la eficacia y el rendimiento del modelo, realizar un análisis de la posible solución para el problema, permite identificar las limitaciones y

oportunidades de mejora conforme a las técnicas de análisis de datos.

- **Preparación de los Datos:** En esta fase, se estaría relacionado con el proceso ETL (Extract, Transform, Load), aquí se desarrollará todas las actividades para preparar los datos y construir el conjunto de datos dataset, definiendo calidad, integridad y los ajustes necesarios para el entrenamiento de modelos.
- **Construcción del Modelo:** En este punto es la selección del tipo de aprendizaje de Machine Learning, se entrenan los algoritmos o modelos que se adapten al problema y a los datos disponibles.
- **Evaluación:** Análisis exhaustivo de los errores que presenta el modelo, para identificar ajustes, precisión y robustez.
- **Integración del Modelo en un Sistema:** Finalmente, es aconsejable implementar en un sistema funcional, para ser usado en entornos de producción para la toma de decisiones.

**Figura 1**

*Metodología CRISP-DM*




*Nota:* Fases del proceso de CRISP-DM (IBM, 2021)

Este tipo de metodología permite abordar los proyectos de Machine Learning de manera organizada y secuencial cumpliendo con el objetivo de tener calidad de los datos y la efectividad en los modelos.

### 2.2.4.3. Modelos de Aprendizaje Supervisado

Los modelos de aprendizaje supervisado como se ha detallado en el contexto anterior son las técnicas utilizadas para solventar problemas de predicción y clasificación, conforme a un conjunto de datos etiquetados. A continuación, se detallan los principales modelos, en el cual se resaltan sus principales características, y su vinculación con el entorno energético.

#### Regresión

Partiendo del concepto que la regresión está dentro del aprendizaje automático y es una técnica que permite entrenar modelos entre la relación de las variables independientes y variable dependiente, con el propósito de realizar predicciones, los tipos de modelos más utilizados son:

#### Regresión lineal simple

De acuerdo a Mohsin Abdulazeez & Hussen (2021), el modelo se enfoca en establecer una relación lineal entre las variable independiente y variable dependiente, según la ecuación:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

donde $y$ representa la variable dependiente, $x$ la independiente, $\beta_0$ y $\beta_1$ los coeficientes del modelo, y $\varepsilon$ el término de error.

#### Regresión Lineal Múltiple (MLR)

La regresión múltiple desarrolla un modelo con varios predictores, analiza los diferentes factores que influyen de forma simultánea en la variable objetivo. Su función sería:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + e$$

este tipo de modelo es más utilizado en un proyecto de predicción energética, en vista que las diversas variables como: condiciones climáticas, hábitos de los habitantes y características hogar son causales impactar el consumo de energía (Ortega & Cardenas, 2022)

La regresión múltiple sirve para entender situaciones complejas, donde influyen más de dos variables. Con este modelo se evalúa los efectos de las variables, que influyen en la variable objetivo mediante una ecuación lineal (Arellano & Peña, 2020)

**Regresión Polinomial**

La regresión polinomial es aplicada cuando la relación entre variables no es lineal, introduce términos de mayor grado con solo una variable regresora o predictora X para capturar comportamientos curvilíneos:

$$y = \beta^0 + \beta^1 x + \beta^2 x^2 + \cdots + \beta_h x^h + \varepsilon$$

$$i = 1, \ldots \ldots n$$

Donde:

n es el número de datos disponibles a ser utilizados para hallar el modelo.

h es el grado del polinomio

Cabe señalar que la regresión polinomial es un caso especial de MLR, porque involucra la transformación de las variables originales con el objetivo de adaptarlas a un modelo lineal en parámetros.

Adicionalmente, se puede aplicar la regresión logística en una clasificación binaria, la cual, permite estimar la probabilidad de ocurrencia de un evento, como: el encendido/apagado de equipos en función de patrones de consumo energético.

## Clasificación

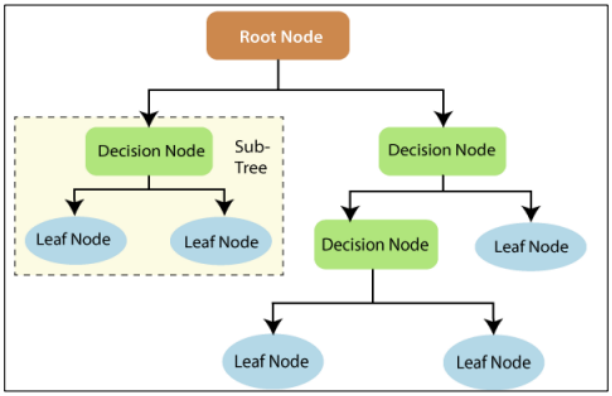
Los algoritmos de clasificación es una subcategoría del aprendizaje supervisado es predecir las clases categóricas a nuevas observaciones, por ejemplo, analizar la información del correo spam, que sería una clasificación binaria (es spam=1; no es spam=0; basándose en un conjunto de entrenamiento previamente etiquetado. Dentro de este grupo destacan los árboles de decisión y los modelos basados en Random Forest (Meneses Díez, Crespo García, & Monzo Sanchez, 2021).

## Árboles de decisión y Random Forest

Según Jijo & Abdulazeez (2021) , los árboles de decisión es una técnica muy utilizada que se basa en estructuras jerárquicas que dividen el espacio de datos en función de los atributos más relevantes, capaces de manejar grandes volúmenes de información. Estos algoritmos se pueden utilizar para clasificar información en función de conjuntos de entrenamiento y etiquetas de clase. La Figura 2 ilustra la estructura de un árbol de decisión.

**Figura 2**

*Árbol de decisiones.*



*Nota:* Estructura de un árbol de decisión (Jijo & Abdulazeez, 2021)

**Random Forest**

El random Forest analizado por Espinoza (2020) consiste en la generación de múltiples árboles de decisión entrenados de manera aleatoria y los resultados obtenidos se combinan a fin de obtener un modelo único y robusto. 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.

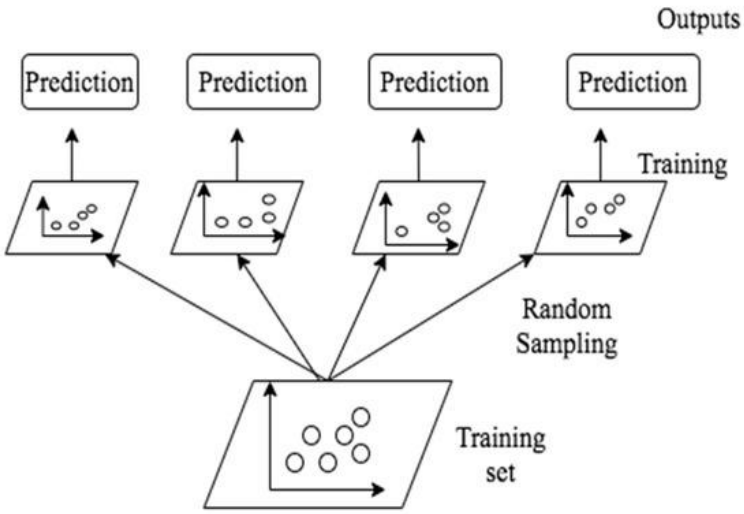
Además, presenta las siguientes ventajas:

- Se obtiene alta precisión y robustez, incluso con grandes volúmenes de datos.
- Tiene la capacidad de manejar diversas variables sin ser necesario su eliminación.
- Identifica las variables importantes.
- Tolerancia a datos faltantes y resistencia al sobreajuste.

Sin embargo, presenta desventajas como la visualización de los resultados existe dificultad para interpretar, la posibilidad de sobreajuste ciertos grupo de datos en presencia de ruido y no puede predecir más allá del rango de valores del conjunto de datos usado para entrenar el modelo. Este modelo ha tenido éxito en la detección de fraudes, clasificación de clientes, diagnóstico médico, predicción de mercados financieros y análisis energético.

***Figura 3.***

*Generación de árboles de decisión.*

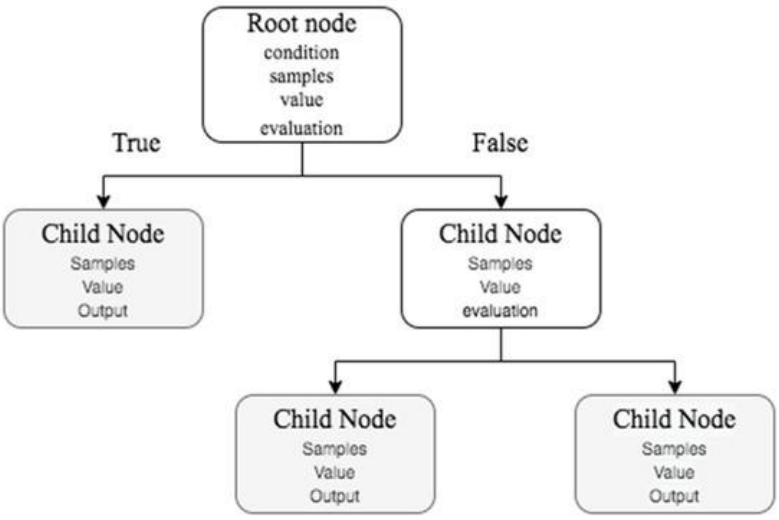



*Nota:* Random Forest, (Salinas, Garciá, Riveros, Gonzalez, & Goanzalesz, 2024)

Cabe destacar lo manifestado por Salinas et al., (2024) que el Random Forest emplea múltiples árboles de decisión para construir modelos predictivos de alta precisión, lo que facilita la identificación de factores principales de los estudios o análisis de datos. En base a la estructura de árbol para modelar las relaciones entre las características y los resultados. Inicia con un nodo raíz (profundidad 0) y los algoritmos empiezan a establecer condiciones o suposiciones sobre los datos; si las suposiciones son verdaderas, se pasa al nodo hijo izquierdo de la raíz (profundidad 1, izquierda), Como lo muestra la figura 4.

**Figura 4**

*Identificación de nodos, Random forest*




*Nota:* Random Forest, nodos estableciendo condiciones Salinas (2024)

**XGBoost**

XGBoost es un modelo de aprendizaje automático muy eficaz y popular para actividades de aprendizaje supervisado, particularmente en las áreas de clasificación y regresión. Se destaca por su gran efectividad, eficiencia y resistencia (Luo & Chen, 2025), de acuerdo, Chen y Guestrin, (2016) . XGBoost es un algoritmo de gradient boosting que integra de forma secuencial modelos débiles, como los árboles de decisión, utilizando métodos de optimización para alcanzar una precisión predictiva elevada y abordar relaciones complejas en los datos.

Como ejemplo del rendimiento de este algoritmo podemos destacar el concurso de competencias de machine learning Kaggle. En el cual, presentaron 17 soluciones utilizaron XGBoost. De las cuales ocho usaron exclusivamente XGBoost para entrenar el modelo, mientras que la mayoría le combinaron XGBoost con redes neuronales en ensembles.

Ante lo contextualizado, los modelos de aprendizaje supervisado, demuestran resultados de vanguardia en una alta gama de problemas. Algunos ejemplos de los problemas

abordados en estas soluciones predicción de ventas en tiendas; clasificación de eventos de física de alta energía; clasificación de texto web; predicción de comportamiento del cliente, etc., su aplicación en la vida diaria se puede evidenciar que los modelos regresión lineal, random forest, SVM han sido aplicados en proyectos de energía eléctrica y XGBoost, su impacto llama la atención de ser modelado para observar que resultados se obtiene en la predicción del consumo de energía en los hogares y por contar con un dataset etiquetado.

#### 2.2.4.4. Métricas para la Evaluación de modelos

Según Meneses (2021), existen diversas métricas para evaluar los algoritmos de Machine Learning, por ejemplo, la precisión, que relaciona el número de casos positivos correctamente clasificados y el total de casos positivos predichos. También existe la métrica de accuracy o exactitud, que se obtiene a partir de la división entre la suma de clasificaciones correctas y el total de los casos clasificados Otra métrica importante es la exhaustividad o Recall, que es la relación entre el número de casos positivos correctamente clasificados y el total de casos positivos reales.

De acuerdo a Alipujiang, (2021, pág. 5), el $R^2$ es la razón de la variable predicha que es explicada por un modelo de regresión. Lo que significa, que es la razón de la variable explicada de la variable total. $R^2$ es el cuadrado de la correlación entre la variable real y la variable predicha. Por lo tanto, $R^2$ varía de 0 a 1. Un valor de 0 indica que el modelo de regresión no explica ninguna de las variables predichas, lo que significa que no hay correlación entre las dos variables. Un valor de 1 indica que el modelo de regresión explica todas las variables predichas, lo que significa que la correlación entre las dos variables es perfecta.

La MAE es una métrica que mide la magnitud promedio de los errores absolutos entre el valor predicho y el valor real. La MAE se suele denominar desviación absoluta media (MAD). El rango de la MAE es (0, + ∞ + ∞ Cuanto menor sea el valor de MAE, mayor será la precisión del modelo de predicción. La ventaja de MAE es que su unidad es la misma que la de los datos originales y es fácil de calcular y comprender. MAE se utiliza a menudo como una función de pérdida simétrica.

El RMSE es la distancia promedio medida verticalmente desde el valor real hasta el valor predicho correspondiente en la línea de ajuste. En pocas palabras, es la raíz cuadrada del MSE. De la misma manera que MSE, el rango de RMSE es (0, + ∞ + ∞) Cuanto menor sea el valor RMSE, mayor será la precisión del modelo de predicción.

#### 2.2.4.5. Comparación de modelos aplicados para la predicción del consumo de energía

Existen proyectos referentes al consumo de energía que han aplicado diversos modelos como:

El proyecto “Modelos de aprendizaje automático para la predicción del consumo energético en función de las cargas de refrigeración y calefacción en edificios inteligentes basados en el Internet de las cosas”, los autores Ghasemkhani, et al., (2022) estudiantes de la Universidad Dokuz Eylul de Turquía, los modelos de aprendizaje automático que construyeron utilizando la red neuronal de tres capas (TNN) y algoritmos de máxima relevancia y mínima redundancia (MRMR). Esto permitió identificar cada característica relacionada con los edificios, a su vez términos de asimetría para determinar si distribuciones son simétricas o asimétricas. Los resultados de este estudio muestran que las propiedades de compacidad relativa y área de acristalamiento tienen el mayor impacto en el consumo de

energía en los edificios, mientras que la orientación y la distribución del área de acristalamiento están menos correlacionadas con las variables de salida. Además, el mejor error absoluto medio (MAE) se calculó en 0,28993 para la predicción de la carga de calefacción (kWh/m²) y 0,53527 para la predicción de la carga de refrigeración (kWh/m²), respectivamente. Los resultados experimentales mostraron que el método aplicado superó a los métodos más avanzados en el mismo conjunto de datos.

El estudio propuesto por Rizwan & Khan.Anam (2024), estudiantes de la Universidad Kyung Hee sobre "Mejorar la predicción del consumo energético en hogares inteligentes: un enfoque de aprendizaje por transferencia federada consciente de la convergencia", en el mismo

Sugiere un nuevo enfoque para predecir el consumo energético, en base al Aprendizaje Federado (AF) con el objetivo de entrenar un modelo global, garantizando la privacidad de los datos locales y transfiriendo conocimiento de edificios con abundante información a edificios con poca información. Para evitar la negatividad algunos nodos participantes con bajo rendimiento debido a datos ruidosos o limitados, proponen una estrategia de selección de clientes para seleccionar a los mejores participantes posibles para el modelo global. El modelo propuesto evalúa y realiza los análisis exhaustivos de los patrones de consumo energético. Validan el rendimiento comparando su Error Absoluto Medio (MAE), Error Cuadrático Medio (MSE) y valores de $R^2$ con los de los modelos tradicionales. Los hallazgos indican que el modelo propuesto basado en FL con participación selectiva del cliente supera a sus métodos homólogos en cuanto a precisión predictiva y robustez.

En la India se desarrolló el Proyecto "ELECTRICITY CONSUMPTION PREDICTION", por los autores Vijendar. Et al., (2023), para predecir el consumo de

energía, manifiestan que ahora es posible anticipar adecuadamente el consumo de energía utilizando datos previos gracias a las mejoras en las técnicas de aprendizaje automático. Investigaron diversas técnicas de aprendizaje automático, incluyendo regresión lineal, K vecinos más cercanos, XGBOOST, bosque aleatorio y redes neuronales artificiales (RNA), para pronosticar el consumo de energía. Utilizaron diversas medidas de evaluación, incluyendo el Error Absoluto Medio (MAE), el Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación ($R^2$), para evaluar el rendimiento de los modelos. El modelo K Neighbours (KNN) superó a todos los demás en rendimiento, con una precisión del 90,92 % en la predicción.

También, podemos mencionar, los estudios realizados en América, referente al tema de la predicción del consumo de energía, así tenemos, el estudio de la ” Predicción de Consumo y Demanda de Electricidad Mediante Redes Neuronales Artificiales y Algoritmo Iterativo”, en donde Montero et al, (2024), entrenan modelos con redes neuronales artificiales (RNA) mediante aprendizaje automático basado en un algoritmo iterativo (ML). El ajuste del modelo fue del 94.5%, mejorando los resultados logrados mediante modelación paramétrica y de regresión múltiple. Los resultados de la investigación reflejan la utilidad del aprendizaje iterativo automático basado en la determinación de modelos predictivos, lo cual favorece los procesos de toma decisiones en cuanto a operación de las redes eléctricas y la planificación energética.

En Chile Yajure, (2022) realizó el estudio “Use of machine learning algorithms to analyze billed electricity data. Case: Chile 2015 – 2021”, en el cual analizó los datos de energía eléctrica facturada mensual de los clientes regulados de Chile, con el fin de detectar patrones y predecir la categoría a la que pertenecen. Utilizó los algoritmos K-Means para la detección de patrones, K-NN para la predicción de la categoría de los clientes, y PCA para

determinar las variables más significativas dentro del conjunto de datos. el algoritmo K-NN, se logró obtener un modelo para predecir el tipo de cliente de los datos, con una exactitud del 98%.

Finalmente, en Ecuador en el "ANÁLISIS DE PREDICCIÓN Y AGRUPACIÓN DE CONSUMO ELÉCTRICO EN LA PROVINCIA DE IMBABURA – ECUADOR PARA LA OPTIMIZACIÓN DE RECURSOS ENERGÉTICOS", Rosero (2021), realizó el entrenamiento con modelos de regresión, lineal, random forest, SVM. De lo cual concluyó que los modelos de regresión no lineal permiten entrenar modelos con variables categóricas, las cuales son utilizadas para definir ciertos factores de las características del usuario. En consecuencia, las máquinas de soporte de decisión y los bosques aleatorios resultaron ser los más adecuados para presentar las tendencias en el consumo eléctrico en la provincia de Imbabura-Ecuador, organizados por cantones y municipalidades

De acuerdo a las comparaciones de modelos, tiene sus ventajas en las predicciones del consumo de energía, se puede evidenciar el uso del KNN y RNA alcanzo una exactitud cerca del 100%, el presente proyecto se basara en  de regresión, lineal, random forest, SVM y XGBoox, por contar con datos etiquetados, y nuestro objetivo es predecir el consumo energético en hogares por parroquias, e identificar cual es el modelo más óptimo, el dataset contiene variables que induce al entrenamiento por aprendizaje supervisado lo cual podemos fragmentar los datos en conjunto de entrenamiento y conjunto de prueba para ejecutar un escalado y de esta manera convertir a las variables numéricas, se evaluara los modelos conforme a las métricas de evaluación como Error Absoluto Medio (MAE), el Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación ($R^2$).

# CAPITULO 3: DESARROLLO

## 3.1. Metodología

En el proyecto se aplica la metodología de CRISP-DM (Cross-Industry Standard Process for Data Mining), es una de las metodologías fuertemente utilizada para el análisis de datos en Machine Learning. CRISP-DM siendo una metodología fácil de manejar, el cual se adapta para diferentes tipos de investigaciones, está orientada en diferentes fases como la comprensión del problema o negocio, la comprensión de datos, la preparación de datos, el modelado, la evaluación del modelo, y la implementación o visualización (Espinosa-Zúñiga J. J., 2020), estas etapas son empleadas para la "Predicción de consumo de energía eléctrica en hogares utilizando modelos de aprendizaje automático", donde ayuda a estructurar, analizar de forma correcta el proceso de trabajo con el fin de asegurar un estudio sistemático.

## 3.2 Recolección y descripción de los datos

La investigación se enfoca en un conjunto de datos históricos, se fundamenta en el consumo energético en hogares del Ecuador, esta información se ha recopilado de la Agencia de Regulación y Control de Electricidad (ARCONEL), es una empresa responsable en administrar y gestionar las funciones de servicio público de Energía Eléctrica y Alumbrado Público General, según www.gob.ec/arconel, para obtener la base de datos se ingresa al enlace https://arconel.gob.ec/ , luego estadística y base de datos, aquí se obtiene el conjunto de datos con el que se trabaja es "facturación de clientes regulados".

El conjunto de datos se adquiere con el formato de archivo de excel .xlsx, su información es del año 2023, con un alcance territorial de las regiones Sierra, Costa, Oriente e Insular. El dataset está compuesto por datos generales, ubicación geográfica, número y consumo de clientes, y valores económicos facturados y subsidiados.

La base de datos está estructura con 17 campos y 60353 registros, los cuales son los siguientes:

**Tabla 1**

*Descripción de los campos de la base de datos*

| Campos | Descripción |
|---|---|
| **Anio** | Año de registro de consumo eléctrico |
| **Mes** | Mes de registro de consumo eléctrico |
| **Empresa** | Institución distribuidora de consumo eléctrico |
| **Grupo Consumo** | Tipo de consumo (comercial, residencial, industrial, alumbrado público y otros) |
| **Provincia** | Lugar territorial |
| **Canton** | Lugar territorial |
| **Parroquia** | Lugar territorial |
| **Numero Clientes** | Total de clientes por parroquia |
| **Energia Facturada (kWh)** | Total de consumo eléctrico por parroquia |
| **Facturacion Servicio Electrico (USD)** | Precio total del consumo eléctrico |
| **Recaudacion (USD)** | Recaudación total del consumo eléctrico |
| **Impuesto Bomberos (USD)** | Gravamen al cuerpo de bomberos |
| **Basura (USD)** | Basura |
| **Tercera Edad (USD)** | Valor subsidiado de la Tercera edad |
| **Tarifa Dignidad (USD)** | Valor subsidiado de la tarifa de dignidad |
| **Energia Subsidiada PEC (kWh)** | Energía subsidiada para el programa de eficiencia energética (PEC) |
| **Valor Subsidiado PEC USD** | Valor subsidiado PEC |

*Nota*: Nombres de los campos obtenidos del “dataset_2023_consumo_electrico.xlsx”

Se realiza un estudio inicial al conjunto de datos para determinar si tiene datos nulos o faltantes como resultado se obtiene el 0%, así como también se observa que no existe duplicados en el dataset.

### 3.3 Preparación y transformación de datos

Se organiza el conjunto de datos con la que se va a realizar una serie de procesos para ello se toma en consideración la columna “Grupo Consumo”, en donde presenta varios tipos de consumo como comercial, residencial, industrial, alumbrado público y otros, pero como el estudio de la investigación se establece en la predicción del consumo energético en hogares del Ecuador, para ello se procede hacer un filtrado de todos los datos con las condiciones de

que el grupo de consumo sea igual a residencial, y la energía facturada sea mayor a cero, con esto se genera otra base de datos (residencial) que está compuesta por 17 columnas y 16385 filas, este dataset presenta sus características con sus respectivos tipos de datos.

**Tabla 2**

*Dataset generada residencial*

| # | Características | Tipo de dato |
|---|---|---|
| 0 | Anio | int64 |
| 1 | Mes | object |
| 2 | Empresa | object |
| 3 | Grupo Consumo | object |
| 4 | Provincia | object |
| 5 | Canton | object |
| 6 | Parroquia | object |
| 7 | Numero Clientes | int64 |
| 8 | Energia Facturada (kWh) | int64 |
| 9 | Facturacion Servicio Electrico (USD) | float64 |
| 10 | Recaudacion (USD) | float64 |
| 11 | Impuesto Bomberos (USD) | float64 |
| 12 | Basura (USD) | float64 |
| 13 | Tercera Edad (USD) | float64 |
| 14 | Tarifa Dignidad (USD) | float64 |
| 15 | Energia Subsidiada PEC (kWh) | float64 |
| 16 | Valor Subsidiado PEC USD | float64 |

*Nota*: Características de la nueva dataset generada residencial

Podemos observar que existe diferentes tipos de datos como: 8 float64, 3 int64 y 6 object, entonces tenemos 11 características numéricas y 6 características categóricas.

Se modifica el conjunto de datos aumentando una columna llamada mes_num con su tipo de dato numérico (int64) para almacenar los números de los meses del 1 al 12, con el propósito de hacerle más fácil el filtrado de la Energía Facturada (KWh) por Provincia y el Número de clientes por Provincia cada mes.

Se realiza un amplio análisis exploratorio del conjunto de datos para comprender como está relacionado las diferentes características como:

La relación entre número de clientes y consumo residencial, muestra que su concentración máxima del número de clientes va desde 0 a 50000 con su consumo

residencia(KWh) que va desde 0 a 10000000 se observa 16211 registros, le sigue el número de clientes de 50000 a 100000 con su consumo residencia(KWh) de 10000000 a 20000000 se encuentra 29 registros y la concentración mínima de registros es 12, va desde 250000 a 300000 de número de clientes con su consumo residencia(KWh) de 60000000 a 90000000, esto expone una gran influencia en lugares de parroquias pequeñas o medianas.

La relación del consumo de energía eléctrica (KWh) por provincia, indica que Guayas es la provincia con máximo consumo de energía eléctrica (KWh), le continua Pichincha y la provincia con menor consumo energético es Galápagos. Podríamos decir que el mayor consumo energético está en el Guayas y Pichincha, ya que tiene mayor población e industrialización y Galápagos menor población, por eso su consumo energético es menor.

La relación de los 20 top de cantones, y parroquias con el consumo energético, tenemos que el máximo consumo energético, es el cantón Guayaquil y su parroquia Ximena, el mínimo consumo energético, es el cantón Babahoyo y la parroquia Tundayme perteneciente al cantón El Pangui, provincia de Zamora Chinchipe, ya que es una zona poco habitada.

La relación del consumo de energía eléctrica (KWh) por mes, no existe mucha diferencia entre los meses el consumo energético, pero el mes donde más consumen energía es mayo y diciembre es el mes que menos consumen, es decir que con esta información el consumo energético es estable.

La relación del número de clientes por provincia, el que más sobresale es la provincia de Pichicha y le continua Guayas, son dos provincias con mayor número de clientes y con menor número de clientes es Galápagos.

Este análisis exploratorio inicial nos da un panorama amplio de la relación de las variables tanto positivo como negativo con el consumo energético, el desempeño de las características en el lugar territorial como provincia, cantón y parroquia, también el distinguir

los datos separados o agrupados en diferentes rangos y por último se confirma que Guayas y Pichincha influyen en el número de clientes y en el consumo energético, así como menos relevante está Galápagos.

Se procede a seleccionar de la base de datos residencial las variables que se dispone a establecer el conjunto de datos con el que se va a trabajar en forma definitiva estas variables son Mes, Provincia, Cantón, Parroquia, Numero Clientes, Energia Facturada (kWh).

Las variables Mes, Provincia, Cantón, Parroquia son características categóricas, a estas variables se les aplica LabelEncoder en cada columna para normalizar las etiquetas, lo que significa es transformar el texto a números enteros. (Pineda Pertuz, 2022, pág. 77)

Para determinar la correlación que existe entre las variables Mes, Provincia, Cantón, Parroquia, Numero Clientes, con la variable objetivo Energia Facturada (kWh), se aplica el método de la matriz de correlación, esta matriz se utiliza para analizar variables numéricas, donde ayuda a establecer las relaciones positivas, negativas o que no existe una relación, señala la forma que se vinculan entre sí, y se obtiene diferentes combinaciones posibles de valores de una tabla para hallar patrones. (aichallenge, 2021). De acuerdo a nuestro análisis con las variables el Numero Clientes con la Energia Facturada (kWh) tiene una relación positiva del 95%, por lo tanto, la característica Numero Clientes es esencial para predecir la característica Energia Facturada (kWh), esto significa que mayor número de clientes corresponde a mayor consumo energético.

Se efectúa una evaluación de las columnas fuertemente correlacionadas que son el Numero Clientes y la Energia Facturada (kWh), mediante la eliminación de los valores atípicos usando el método IQR (Rango interquartil), su funcionamiento es calcular la diferencia entre el tercer cuartil(Q3) que equivale al 75% y el primer cuartil(Q1) que equivale al 25% , para obtener el IQR y los limites inferior y superior, todos los puntos que están por debajo del límite inferior y por encima del límite superior son valores atípicos(outlier), los

cuales deben ser eliminados, (Nazari, 2024, pág. 63), con este proceso el conjunto de datos queda con 14043 filas y 6 columnas.

En el procedimiento de división de entrenamiento y prueba se determina las características para “X” está Mes, Provincia, Cantón, Parroquia, Numero Clientes, y para “y” la característica a predecir está Energia Facturada (kWh), con esta operación se define el conjunto de entrenamiento para entrenar y aprender el modelo y el conjunto de prueba para valorar correctamente el funcionamiento y la eficiencia del modelo entrenado para gestionar nuevos datos, el porcentaje del conjunto de entrenamiento debe ser grande para alcanzar resultados relevantes, pero no tan grande para que el modelo no se sobreajuste, (Sheikh, 2025), para ello, el 70% está para el conjunto de entrenamiento y el 30% para el conjunto de prueba, con una semilla de 7 para una valoración sólida en el desempeño del modelo.

Después de fragmentar los datos en conjunto de entrenamiento y conjunto de prueba se ejecuta un escalado para convertir a las variables numéricas en un mismo intervalo, donde se resta la media y se divide la desviación estándar, (Ahmad, 2024), por consiguiente, se normaliza la variable independiente(X) y la variable dependiente(y) del conjunto de entrenamiento, los datos están listos para ser ejecutados como datos de entrada en los diferentes modelos que presentamos en el literal 3.4.

### 3.4 Selección y Construcción del Modelo

La investigación se basa en los modelos de aprendizaje automático o Machine Learning (ML), una especialidad de la inteligencia artificial, su propósito es elaborar métodos que favorezcan a las máquinas aprender y optimizar su desempeño a partir de datos. Específicamente, el estudio está orientado en el aprendizaje supervisado el cual maneja conjuntos de datos etiquetados que depende de una variable objetivo, (Pineda Pertuz, 2022, pág. 35), en función de eso, se aplica un conjunto de datos estructurado, visualizado en forma

de tabla dentro de una base de datos. Este tipo de aprendizaje facilita definir las relaciones entre variables independientes y una variable dependiente, lo que implica a modelos de regresión, un procedimiento característico del aprendizaje automático supervisado que se efectúa para predecir valores continuos.

En esta fase se exponen varios modelos de regresión, con el objetivo de reconocer el alto desempeño predictivo a partir de los datos existentes entre los modelos seleccionados se proponen: Regresión lineal, Random Forest, XGBoost (Extreme Gradient Boosting).

Como punto inicial se plantea una función denominada gráficos de residuo con el fin de explorar el modelo mediante gráficos, para eso, se debe obtener el valor de los residuos que se calcula mediante la diferencia entre los valores reales de la variable objetivo y los predichos por el modelo, con el propósito de visualizar dos tipos de gráficos:

**El histograma de residuos**: proyecta una curva de densidad suave encima del histograma con 30 barras de intervalo para determinar la distribución continua de los datos y evaluar la eficiencia del modelo por medio de sus errores.

**Un gráfico de dispersión residuos vs predichos**: visualiza una línea paralela al eje x, donde y = 0, misma que permite verificar los errores que están centrados alrededor de cero. (Soporte de Minitab, 2025). Estos gráficos se emplean para interpretar el desarrollo de cada uno de los modelos propuestos y así comprobar la estabilidad entre los valores reales y predichos.

La **regresión lineal** entrena el modelo con los datos escalados y predice sobre el conjunto de prueba que están escalados para luego efectuar la transformación inversa del escalado en relación a los valores predichos, con el fin de regresar al escalado original, este proceso se realiza porque estamos pronosticando el consumo energético, por lo tanto, necesitamos trabajar en kilowatios hora (KWh). También se verifica el rendimiento del modelo mediante las métricas más utilizadas en regresión, las cuales se describen a

continuación: Error absoluto medio (MAE), Raíz del Error cuadrático medio (RMSE) y Coeficiente de determinación ($R^2$).

Además de los resultados de las métricas, integra la visualización del histograma de residuos, así como los gráficos de dispersión “residuos vs predichos”, y “predicción vs real”, este último gráfico se relaciona los valores reales con la predicción, donde se incluye una línea diagonal para mostrar la estimación perfecta. Mientras más cerca esté los puntos a la línea proyecta la exactitud del modelo. Por último, se usa statsmodels para describir un resumen estadístico del modelo de regresión lineal entrenado, incorporando los componentes R-squared, Adj. R-squared, Coef, std err, t, P> t, etc. (Ramérez Gil, 2023, pág. 90). En el procedimiento incluye un término constante para que la línea de regresión no acceda por el origen y se diseña un modelo de regresión lineal por mínimos cuadrados ordinarios (OLS), (fastercapital, 2025), con la finalidad de enlazar la variable objetivo con las variables independientes.

El **Random Forest** (bosque aleatorio), construye un modelo de regresión de 100 árboles de decisión diferentes dentro del bosque aleatorio para que tenga una estabilidad eficiente, con una semilla igual a 23, con el fin de generar los números aleatorios, tanto el entrenamiento como las predicciones se emplean variables escaladas, para que después sean retornadas a su escala original, pero Random Forest no requiere de datos escalados, ya que se maneja mediante divisiones dentro de los árboles, aunque es importante aplicar datos escalados para conservar el equilibrio. (Ramérez Gil, 2023, pág. 83) .

Se ejecuta la función evaluar_modelo para medir el rendimiento de modelo, donde se calcula las métricas de evaluación como Error absoluto medio (MAE), Raíz del Error cuadrático medio (RMSE) y Coeficiente de determinación ($R^2$), acompañado de la visualización del gráfico de dispersión “predicción vs real”. Además, se evalúa los residuos(errores) mediante el histograma de residuos, y el gráfico de dispersión “residuos vs

predichos", ejecutando la función graficos_residuos, así como para determinar las características más importantes, se presenta un gráfico de barras, estas características se obtienen del conjunto de variables independientes X.

El **XGBoost** frecuentemente usado para aplicaciones de regresión, trabaja desarrollando numerosos árboles, por lo mismo, se construye 100 árboles de decisión, con una semilla aleatoria igual a 23, incluyendo la opción de eliminar los mensajes en consola a lo largo del entrenamiento. Contreras Bravo, Leonardo Emiro, et al (2024, pág. 271). Igual que los modelos anteriores se procede a entrenar y predecir con datos escalados, es decir las variables independientes y la variable objetivo ya están previamente escalados, y después se revierte el escalado a sus valores originales, pero antes de efectuar el revestimiento se convierte en una matriz columna los resultados de la predicción del modelo. Así como también se ejecuta las funciones: evaluar_modelo para medir el desempeño del modelo con las métricas de evaluación y el gráfico de dispersión "predicción vs real" y graficos_residuos para examinar los errores con el histograma, y el gráfico de dispersión "residuos vs predichos". En última instancia presenta un gráfico de barras de las variables independientes mostrando la relevancia que existe entre ellas.

Se ha evaluado tres modelos de regresión estableciendo así: la regresión lineal es fácil de implementar, ayuda a identificar qué variables tiene un impacto relevante en la variable dependiente, el bosque aleatorio genera múltiples árboles de decisión, siendo un modelo más robusto, y simple de entrenar, a veces es difícil de interpretar su visualización gráfica, el XGBoost se basa en árboles de decisión, sus resultados son precisos y excelente velocidad de ejecución, sin embargo puede consumir muchos recursos computacionales en grandes base de datos, (Espinosa-Zúñiga J. , 2020), los beneficios de estos algoritmos permiten predecir en muchos campos diferentes.

### 3.5 Herramientas y Tecnologías Utilizadas

La propuesta de investigación para predecir el consumo energético en los hogares del Ecuador, se desarrolla en el lenguaje de programación Python. Python es un lenguaje que está creciendo a gran velocidad, por ello se ha vuelto el lenguaje de programación más utilizado, ya que ayuda a solucionar varios problemas, está presente en distintas áreas, siendo importante en el ámbito de la ciencia de datos y el aprendizaje automático. Python contiene una amplia gama de librerías de manipulación de datos, visualización de datos, y aprendizaje automático y profundo. En este proyecto se ha utilizado una serie de bibliotecas para realizar, el análisis exploratorio de datos, el preprocesamiento de datos, el modelado, la evaluación del modelo y la visualización, a continuación, se despliega el conjunto de herramientas más relevantes en Python.

- **Numpy**: emplea para análisis numéricos, como cálculos matemáticos, estadísticos, algebra lineal, etc. Aceptan datos multidimensionales.
- **Pandas**: aplica para la depuración, estandarización y preparación de datos, está estructurado con series y dataframe.
- **Seaborn**: basado en matplotlib, crea gráficos estadísticos atractivos.
- **Matplotlit**: crea graficos 2D y 3D, como de líneas, de barras, de dispersión e histogramas. Incio Puyen, Jenifer at. al (2021)
- **sklearn**: suministra una gran variedad de técnicas de aprendizaje supervisado y no supervisado.
- **StandardScaler**: Estandariza los datos a valores numéricos, forma parte del módulo dsklearn.preprocessing.
- **train_test_split**: divide los datos en entrenamiento y prueba, pertenece al módulo de sklearn.model_selection. (datascientest, 2022)

- **LinearRegression**: ajusta a un modelo lineal, es parte del paquete sklearn.linear_model
- **XGBRegressor**: usa en el algoritmo de regresión basado en XGBoots, pertenece al módulo xgboost.
- **RandomForestRegressor**: establece diferentes regresores de árboles de decisión, está incluido en sklearn.ensemble.
- **mean_squared_error, mean_absolute_error, r2_score**: evalúa el desempeño de los modelos, es componente de sklearn.metrics
- **Statsmodels**: suministra clases y funciones para la evaluación de diversos modelos estadísticos. (statsmodels, 2024)

Las plataformas que se maneja para redactar y ejecutar el código de Python, es mediante el soporte establecido en la nube "Google colab", el cual brinda elementos computacionales como el CPUs, también se trabaja en el entorno del Visual Studio Code, instalado en la maquina local.

### 3.6 Validación Técnica Inicial

La validación técnica cumple una función importante en el mejoramiento de los procesos, asegurando la calidad del rendimiento de los modelos, para ello tenemos algunas técnicas de validación empleadas en los algoritmos de regresión, por consiguiente, se describe a continuación:

**División del conjunto de datos:** Al conjunto de datos se le particiona en dos subconjuntos: uno para entrenamiento y el otro para prueba, cada subconjunto le corresponde un valor en porcentaje, es decir al conjunto de datos de entrenamiento se le asigna el 70% y al conjunto de datos de prueba le pertenece el 30%, para valorar el desarrollo de un modelo en datos de validación. (Gatica Oyarzún, 2024, pág. 5)

**Métricas de evaluación para algoritmos de regresión:** El usar varias métricas facilita evaluar el rendimiento para certificar un adecuado funcionamiento, con el fin de mejorar la predicción del modelo, estas métricas se despliegan a continuación:

MAE (Mean Absolute Error): Error absoluto medio, determina el promedio de la sumatoria de la diferencia absoluto entre los datos reales y los valores pronosticas. Los resultados predichos del algoritmo son mejores, cuando el valor del MAE es bajo.

MSE (Mean Squared Error): Error cuadrático medio, obtiene el promedio de los errores al cuadrado. Es lo mismo que MAE, bajo MSE mayor exactitud en los resultados.

RMSE (Root Mean Squared Error): Raíz del Error cuadrático medio, es la raíz cuadrada del MSE.

$R^2$ (Coeficiente de determinación): Coeficiente de determinación, representa el valor de varianza de la predicción. Varia de 0 a 1. (Pineda Pertuz, 2022, pág. 123).

# CAPITULO 4: Análisis y Discusión de los Resultados

## 4.1. Introducción al Capítulo

Este capítulo presenta un análisis exhaustivo de los resultados obtenidos tras la implementación de los modelos predictivos de consumo energético. El objetivo central del estudio, como se ha establecido previamente, es desarrollar un modelo capaz de predecir con alta precisión la variable "Energia Facturada (kWh)" en el sector residencial de Ecuador, basándose en características geográficas, temporales y demográficas.

Para abordar este problema de regresión, se han entrenado y evaluado tres algoritmos de aprendizaje automático con distintas características y niveles de complejidad:

1. **Regresión Lineal:** Un modelo estadístico fundamental que sirve como línea base (*baseline*) para cuantificar el rendimiento de enfoques más avanzados.
2. **Random Forest:** Un modelo de conjunto (*ensemble*) basado en árboles de decisión, conocido por su robustez y alta capacidad predictiva en problemas complejos.
3. **XGBoost (Extreme Gradient Boosting):** Un algoritmo avanzado de *boosting* que destaca por su eficiencia y rendimiento, siendo frecuentemente el estado del arte en competiciones de ciencia de datos.

El rendimiento de cada modelo se evaluará rigurosamente utilizando un conjunto de métricas estándar en tareas de regresión:

- **Error Absoluto Medio (MAE):** Mide la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección. Se expresa en las mismas unidades que la variable objetivo (kWh), ofreciendo una interpretación directa del error de predicción.

- **Raíz del Error Cuadrático Medio (RMSE):** Similar al MAE, pero penaliza en mayor medida los errores grandes debido a la exponenciación. Es una métrica sensible a valores atípicos y también se expresa en kWh.
- **Coeficiente de Determinación ($R^2$):** Indica la proporción de la varianza en la variable dependiente (consumo energético) que es predecible a partir de las variables independientes. Un valor cercano a 1 sugiere que el modelo explica una gran parte de la variabilidad de los datos.

La estructura de este capítulo seguirá un orden lógico: en primer lugar, se analizará el rendimiento y las características de cada modelo de forma individual. Posteriormente, se llevará a cabo un análisis comparativo para identificar el modelo con el mejor desempeño predictivo. Finalmente, se discutirán los hallazgos clave, con especial atención a la importancia de las variables, para extraer conclusiones significativas sobre los factores que determinan el consumo eléctrico residencial en el contexto ecuatoriano.

## 4.2. Evaluación del Modelo de Regresión Lineal

El modelo de Regresión Lineal se implementó como un punto de partida (baseline) para establecer un umbral de rendimiento mínimo. A pesar de su simplicidad, este modelo proporciona información valiosa sobre las relaciones lineales entre las variables predictoras y el consumo energético.

### 4.2.1. Rendimiento Cuantitativo

El modelo fue evaluado en el conjunto de prueba, arrojando las siguientes métricas de rendimiento:

- **Error Absoluto Medio (MAE):** 44,309.91 kWh.
- **Raíz del Error Cuadrático Medio (RMSE):** 74,643.14 kWh.

- **Coeficiente de Determinación ($R^2$):** 0.8049.

El valor de $R^2$ de 0.8049 indica que el modelo de Regresión Lineal es capaz de explicar aproximadamente el 80.5% de la variabilidad en el consumo de energía facturado. Este es un resultado considerablemente bueno para un modelo de línea base, sugiriendo que existe una fuerte componente lineal en la relación entre las variables. Sin embargo, el MAE y el RMSE revelan que el error de predicción promedio es significativo, con desviaciones que pueden superar los 74,000 kWh, lo cual evidencia un margen de mejora considerable.

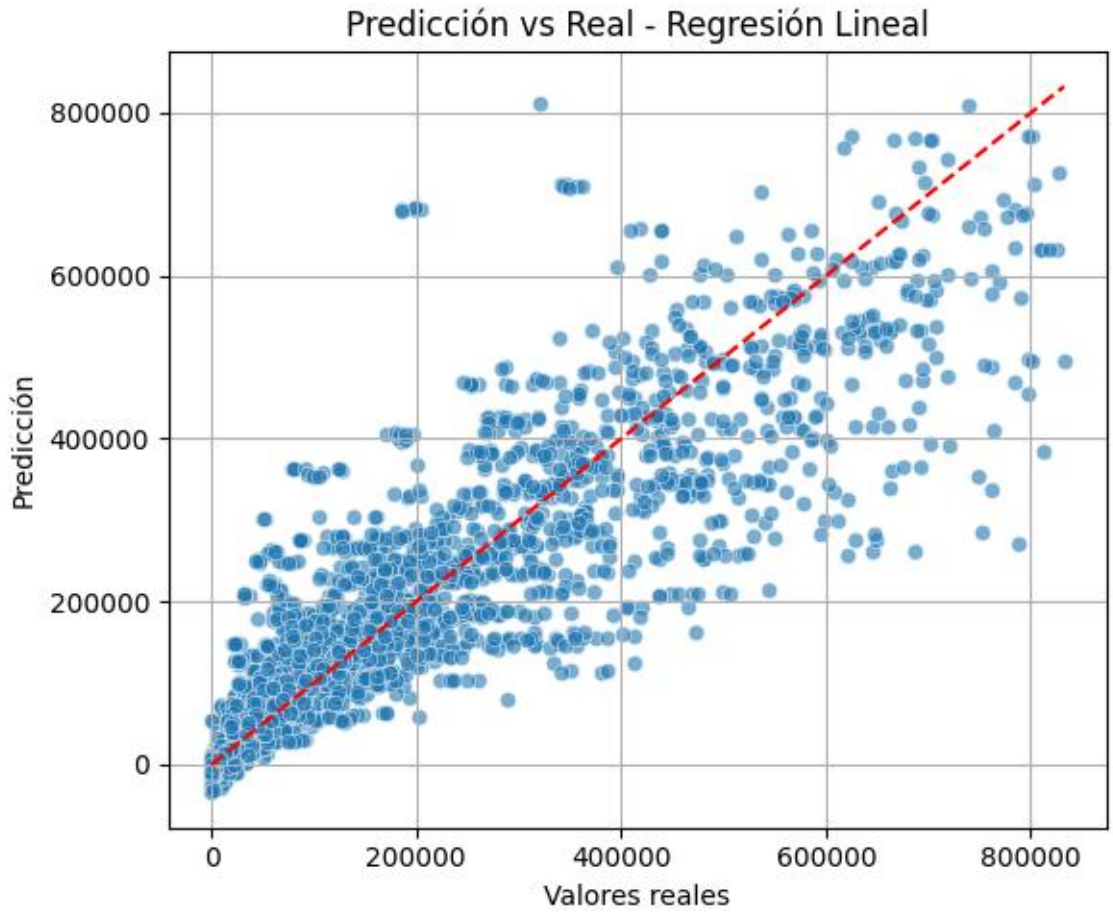
### 4.2.2. Análisis de Residuos

El análisis de los gráficos de residuos permite diagnosticar la calidad y las limitaciones del modelo:

- **Gráfico de Predicción vs. Real:** En la figura que compara los valores reales con los predichos, se observa que los puntos se agrupan en torno a la línea diagonal, lo que confirma la tendencia lineal capturada por el modelo. No obstante, la dispersión es notable, especialmente para valores altos de consumo, donde el modelo tiende a subestimar el consumo real.

**Figura 5**

*Predicción Regresión lineal*

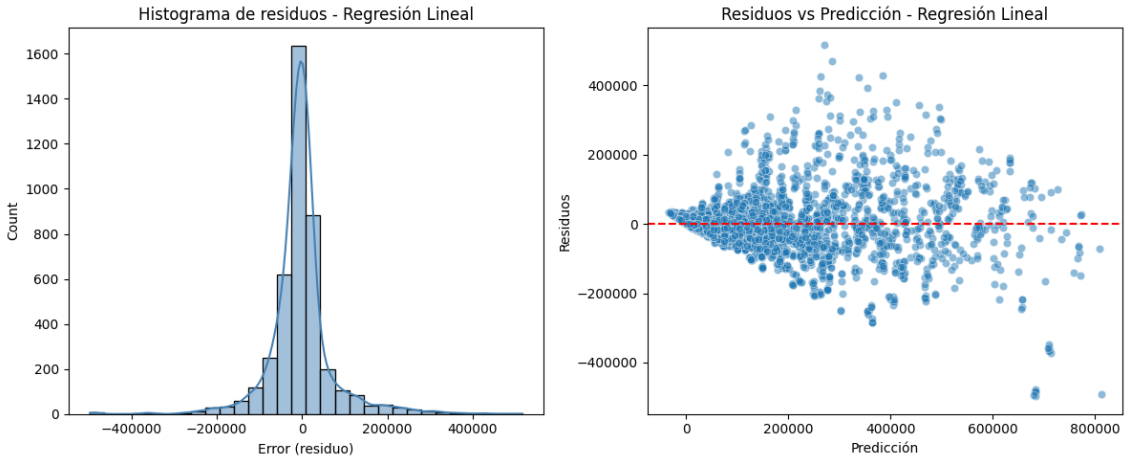



*Nota:* Comparación de los valores reales con los predichos, (Defaz, Nuñez, Ortiz, & Reyes, 2025)

- **Gráficos de Residuos:** El histograma de residuos muestra una distribución que se aproxima a una curva normal centrada en cero, lo cual es un indicio positivo. Sin embargo, el gráfico de "Residuos vs. Predicción" revela un patrón claro de heterocedasticidad: a medida que aumenta el valor de la predicción, la dispersión de los residuos también aumenta. Esto toma la forma de un cono y significa que la precisión del modelo disminuye significativamente al predecir consumos energéticos más elevados, constituyendo una de sus principales debilidades.

**Figura 6**

*Gráficos de residuos*




*Nota:* Distribución que se aproxima a una curva normal centrada en cero, (Defaz, Nuñez, Ortiz, & Reyes, 2025)

### 4.2.3. Interpretación de Coeficientes y Significancia Estadística

El resumen del modelo OLS (Ordinary Least Squares) proporciona una visión más profunda sobre la contribución de cada variable:

- **Significancia de las variables:** Las variables Numero Clientes (p=0.000), Provincia (p=0.000) y Canton (p=0.009) resultaron ser estadísticamente significativas para predecir el consumo[8]. Por otro lado, las variables

Mes (p=0.753) y Parroquia (p=0.176) no mostraron una influencia significativa en este modelo lineal específico. Esto sugiere que, bajo un supuesto lineal, la localización a nivel de parroquia o el mes del año no aportan un poder predictivo relevante una vez que se considera el cantón, la provincia y el número de clientes.

- **Multicolinealidad:** El análisis arroja una advertencia sobre el alto número de condición (7.19e+03), lo que puede indicar la presencia de multicolinealidad. Esto

significa que algunas de las variables predictoras podrían estar correlacionadas entre sí (ej., Cantón y Provincia), lo que podría afectar la estabilidad e interpretación de los coeficientes individuales, aunque no necesariamente la capacidad predictiva general del modelo.

En resumen, la Regresión Lineal funciona como un modelo base aceptable, pero sufre de limitaciones importantes como la heterocedasticidad y su incapacidad para capturar relaciones más complejas, lo que justifica la exploración de modelos más avanzados.

## 4.3. Evaluación del Modelo Random Forest

Al tratarse de un modelo de conjunto (ensemble) no lineal, se esperaba que el Random Forest superara al modelo de Regresión Lineal. Su capacidad para combinar las predicciones de múltiples árboles de decisión le permite capturar interacciones complejas entre variables que un modelo lineal no puede modelar.

### *4.3.1. Rendimiento Cuantitativo*

La evaluación del modelo en el conjunto de prueba arrojó resultados notablemente superiores, demostrando un salto cualitativo en la capacidad de predicción:

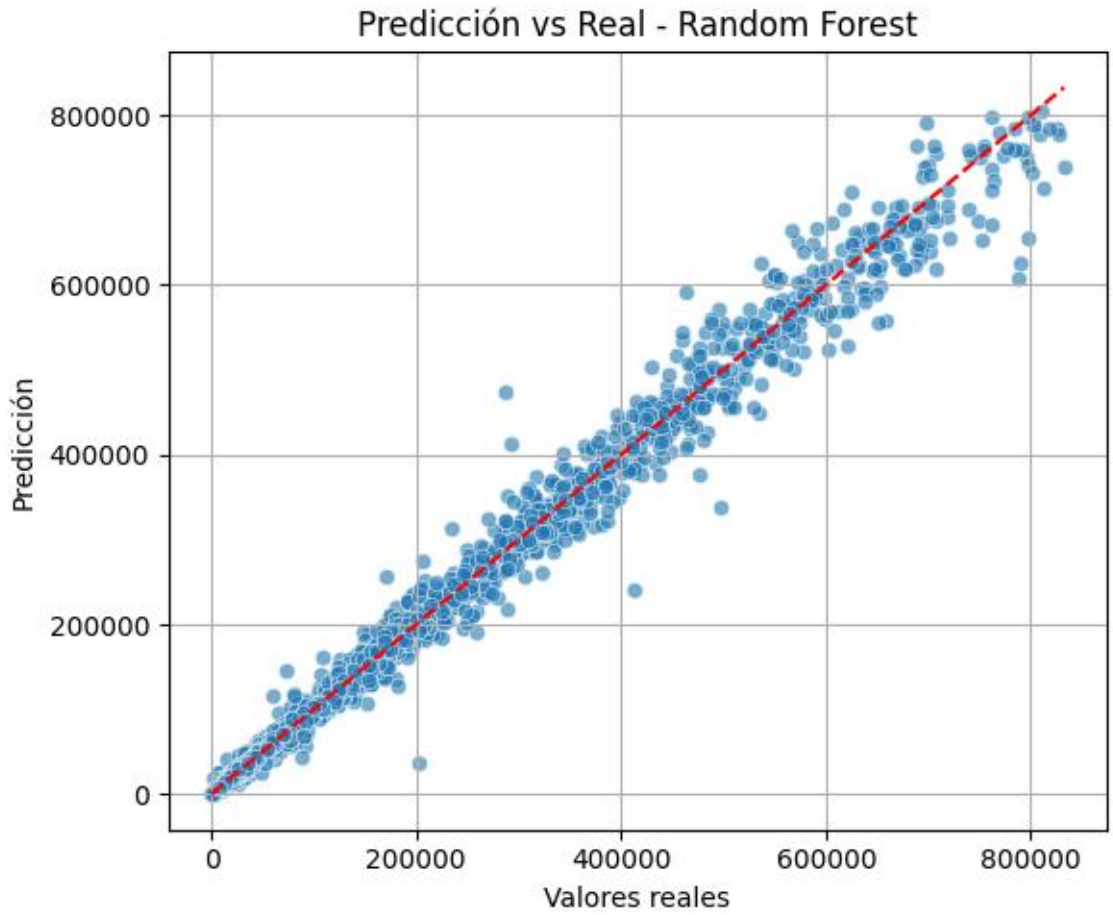
- **Error Absoluto Medio (MAE):** 7,884.33 kWh.
- **Raíz del Error Cuadrático Medio (RMSE):** 16,415.11 kWh.
- **Coeficiente de Determinación ($R^2$):** 0.9906.

El $R^2$ de 0.9906 es un resultado excepcional e indica que el modelo Random Forest es capaz de explicar el 99.06% de la variabilidad en el consumo energético residencial. La comparación con el $R^2$ de 0.8049 del modelo lineal evidencia una mejora drástica. Asimismo, el MAE y el RMSE se han reducido en un 82% y un 78% respectivamente, lo que se traduce

en un error de predicción mucho menor y, por tanto, en una fiabilidad significativamente mayor.

***Figura 7***

*Predicción vs Real, Random Forest*




*Nota:* Variabilidad en el consumo energético residencial (Defaz, Nuñez, Ortiz, & Reyes, 2025)

### 4.3.2. Análisis de Residuos

Los gráficos de diagnóstico para el Random Forest confirman su superioridad y la corrección de las deficiencias observadas en el modelo lineal:

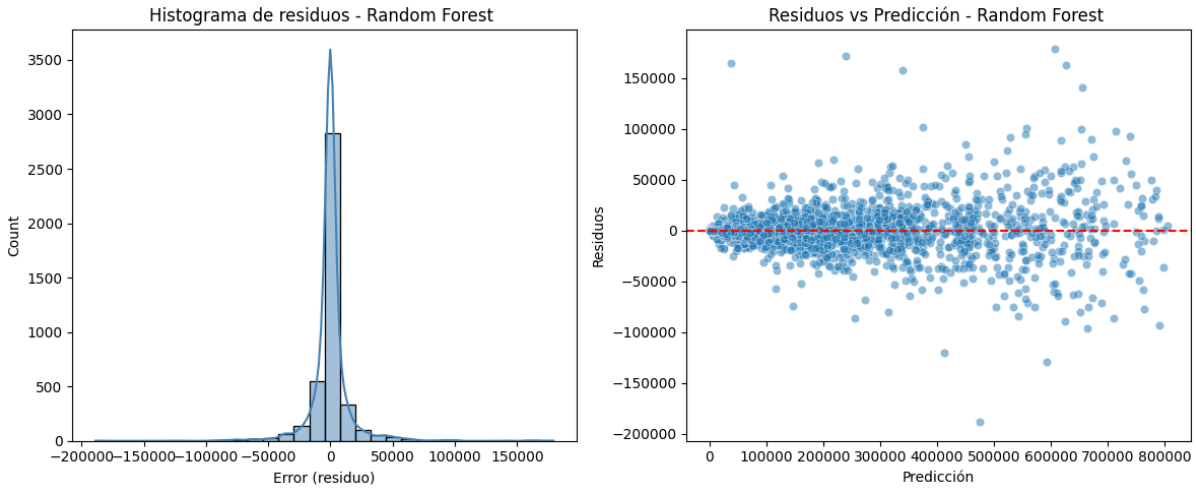
- **Gráfico de Predicción vs. Real:** La nube de puntos se alinea de forma mucho más precisa y compacta a lo largo de la línea de identidad. La dispersión es mínima en

todo el rango de valores, lo que demuestra que el modelo es altamente preciso tanto para consumos bajos como altos.

- **Gráficos de Residuos:** El histograma muestra que los residuos están fuertemente concentrados en torno al cero, con una varianza mucho menor que en el modelo lineal. De forma crucial, el gráfico de "Residuos vs. Predicción" no presenta el patrón de cono (heterocedasticidad) anterior. En su lugar, se observa una distribución de errores mucho más aleatoria y uniforme (homocedástica), lo que indica que la fiabilidad del modelo es consistente en todo el espectro de predicciones.

**Figura 8**

*Gráfico de residuos, Random Forest*



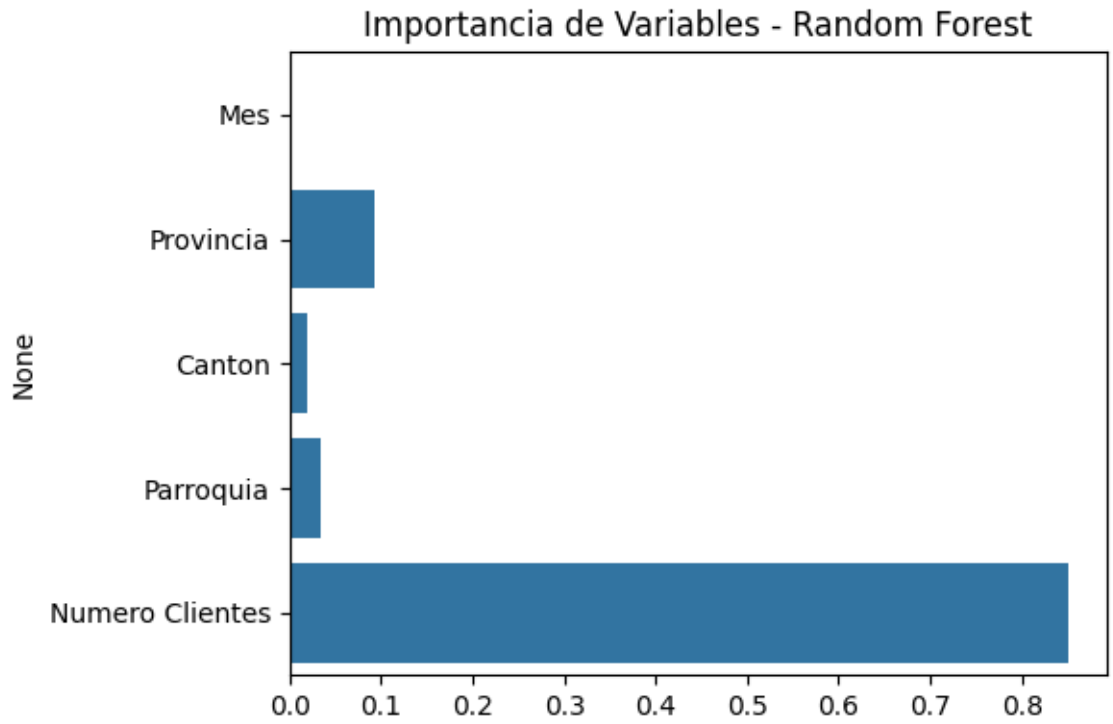

*Nota:* Distribución de errores mucho más aleatoria y uniforme, (Defaz, Nuñez, Ortiz, & Reyes, 2025)

### *4.3.3. Importancia de las Variables (Feature Importance)*

El modelo Random Forest permite, además, evaluar la contribución de cada variable al proceso de predicción. El gráfico de importancia de variables revela una jerarquía clara:

**Figura 9**

*Importancia de variables Random Forest*




*Nota:* Evaluación la contribución de cada variable al proceso de predicción, (Defaz, Nuñez, Ortiz, & Reyes, 2025)

- **Variable dominante:** La variable Numero Clientes es, con diferencia, el predictor más importante, con una puntuación de importancia superior a 0.8. Este hallazgo es intuitivo y coherente con la lógica del problema: el principal factor que determina el consumo total de una zona es, directamente, la cantidad de consumidores que residen en ella.
- **Variables secundarias:** Las variables geográficas (Provincia, Parroquia, Canton) y la temporal (Mes) tienen puntuaciones de importancia mucho menores, pero no nulas. Esto significa que, si bien el número de clientes explica la mayor parte del consumo, la ubicación específica y el mes del año aportan información valiosa que el modelo utiliza para refinar y mejorar la precisión de sus predicciones.

En conclusión, el Random Forest no solo ofrece una precisión predictiva muy alta, sino que también soluciona las limitaciones estadísticas del modelo lineal, posicionándose como un candidato robusto y fiable para la tarea propuesta.

## 4.4. Evaluación del Modelo XGBoost

El modelo XGBoost (Extreme Gradient Boosting) representa otro enfoque de ensemble basado en árboles, pero utiliza una técnica de optimización secuencial (boosting) en lugar del promediado de modelos independientes (bagging) de Random Forest. Se evalúa para determinar si puede ofrecer una mejora de rendimiento adicional.

### 4.4.1. Rendimiento Cuantitativo

Los resultados del modelo XGBoost en el conjunto de prueba son de muy alta calidad y se sitúan a la par con los de Random Forest:

- **Error Absoluto Medio (MAE):** 9,267.02 kWh.
- **Raíz del Error Cuadrático Medio (RMSE):** 16,819.74 kWh.
- **Coeficiente de Determinación ($R^2$):** 0.9901.

Con un $R^2$ de 0.9901, el modelo XGBoost también explica más del 99% de la varianza en los datos de consumo. Su rendimiento es prácticamente idéntico al de Random Forest, aunque este último presenta un error ligeramente inferior (MAE de 7,884.33 para Random Forest frente a 9,267.02 para XGBoost). Esta similitud en el rendimiento sugiere que ambos modelos han alcanzado un nivel cercano al máximo potencial predictivo con las variables disponibles.

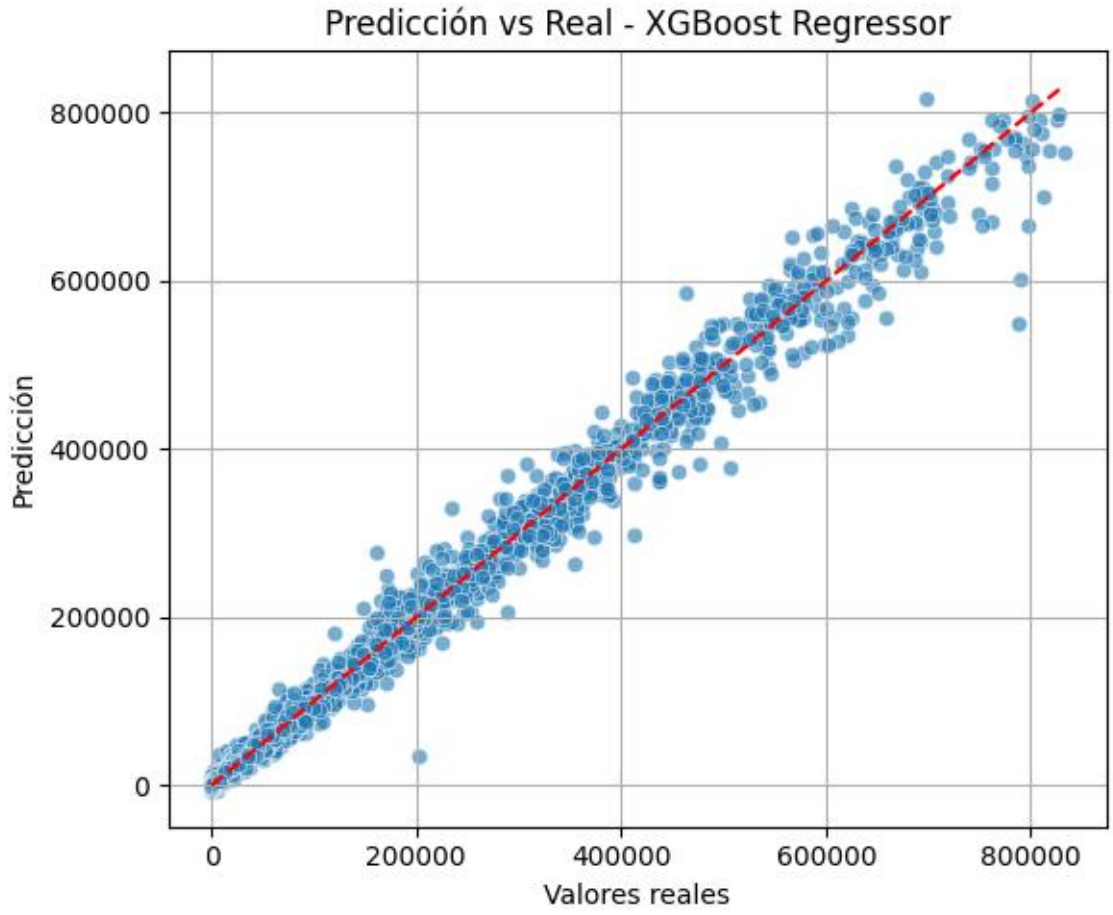
#### *4.4.2. Análisis de Residuos*

Los gráficos de diagnóstico del modelo XGBoost son muy similares a los de Random Forest, validando su robustez y precisión:

- **Gráfico de Predicción vs. Real:** La representación visual muestra una alineación casi perfecta de los puntos a lo largo de la línea de identidad, confirmando la alta precisión del modelo en todo el espectro de valores de consumo.

***Figura 10***

*Gráfico de Predicción vs. Real XGBoost*

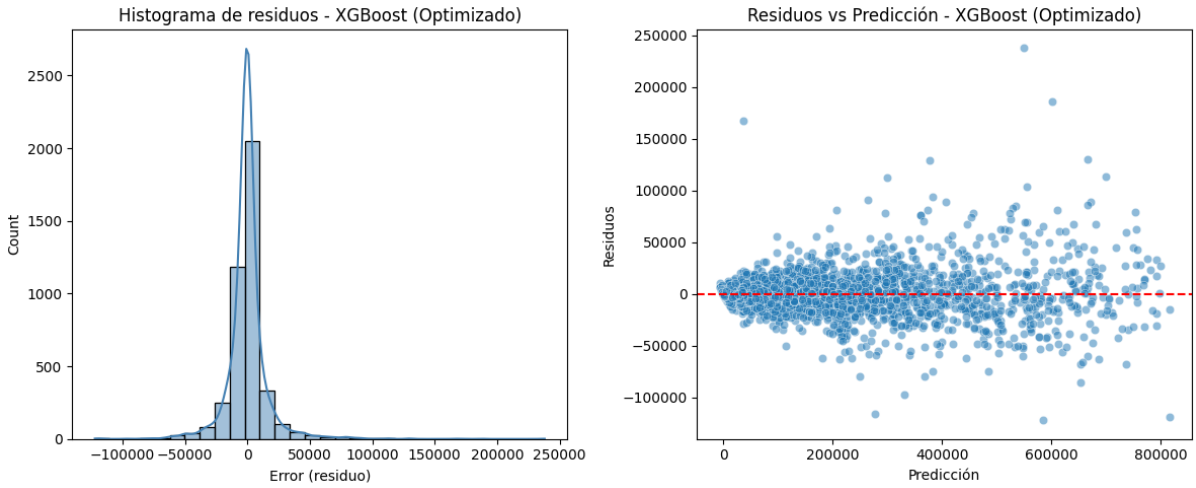



*Nota:* Alta precisión del modelo XGBoost, (Defaz, Nuñez, Ortiz, & Reyes, 2025)

- **Gráficos de Residuos:** El análisis de los residuos muestra una distribución homocedástica (varianza constante) y centrada en cero, sin los patrones problemáticos

observados en la Regresión Lineal. Esto confirma que el modelo XGBoost, al igual que Random Forest, es fiable y sus errores son pequeños y aleatorios.

**Figura 11** *Gráfico de residuos XGBoost*



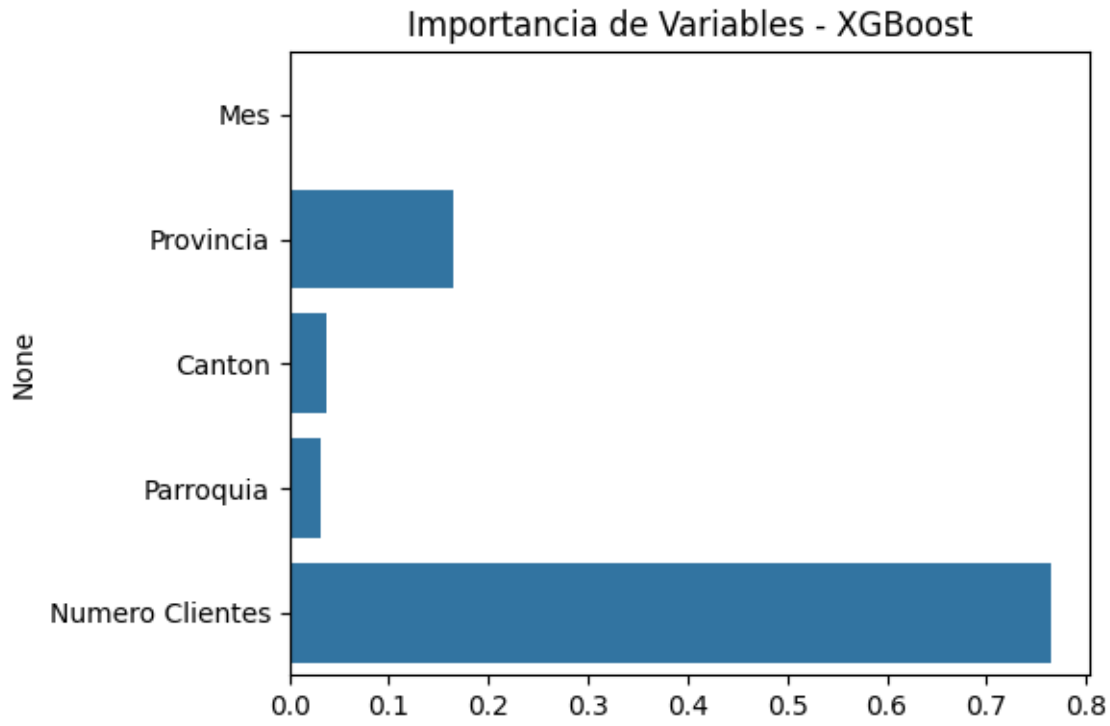

*Nota:* residuos muestra una distribución homocedástica, (Defaz, Nuñez, Ortiz, & Reyes, 2025)

### *4.4.3. Importancia de las Variables*

El análisis de importancia de variables en XGBoost refuerza los hallazgos obtenidos con Random Forest:

**Figura 12**

*Importancia de variables*




*Nota:* variable con mayor poder predictivo y variables geográficas, (Defaz, Nuñez, Ortiz, & Reyes, 2025)

- **Variable dominante:** Nuevamente, Numero Clientes se identifica como la variable con mayor poder predictivo, destacando muy por encima de las demás.
- **Variables secundarias:** Las variables geográficas como Provincia, Parroquia y Canton muestran una contribución menor pero relevante, al igual que la variable Mes. Es interesante notar que Provincia parece tener una importancia relativa ligeramente mayor en XGBoost en comparación con Random Forest, aunque la jerarquía general se mantiene.

En conclusión, XGBoost se presenta como otro modelo de excelente rendimiento, confirmando que los métodos basados en árboles de decisión son extraordinariamente efectivos para esta tarea de predicción. Su desempeño, casi indistinguible del de Random Forest, proporciona una fuerte evidencia de la fiabilidad de los resultados.

## 4.5. Análisis Comparativo y Discusión General

Una vez evaluado el rendimiento de cada modelo de forma individual, es fundamental realizar una comparación directa para seleccionar el algoritmo más adecuado y discutir las implicaciones de los hallazgos.

### *4.5.1. Tabla Comparativa de Métricas*

Para facilitar la comparación, la siguiente tabla resume las métricas de rendimiento clave obtenidas para cada uno de los tres modelos en el conjunto de datos de prueba.

**Tabla 3**

*Comparación de métricas*

| **Métrica** | **Regresión Lineal** | **Random Forest** | **XGBoost Regressor** |
|---|---|---|---|
| **MAE (kWh)** | 44,309.91 | 7,884.33 | 9,267.02 |
| **RMSE (kWh)** | 74,643.14 | 16,415.11 | 16,819.74 |
| **$R^2$ (Coef. de Determinación)** | 0.8049 | 0.9906 | 0.9901 |

*Nota:* Detalle métricas de rendimiento, (Defaz, Nuñez, Ortiz, & Reyes, 2025)

### *4.5.2. Selección del Mejor Modelo*

La evidencia cuantitativa es concluyente. Tanto Random Forest como XGBoost superan de manera abrumadora al modelo de Regresión Lineal, demostrando la naturaleza no lineal del problema.

Al comparar los dos modelos de conjunto, se observa que ambos ofrecen un rendimiento excepcional. Sin embargo, el modelo Random Forest presenta métricas

marginalmente superiores en todos los aspectos: su Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE) son ligeramente más bajos, y su coeficiente de determinación ($R^2$) es mínimamente más alto. Aunque la diferencia es pequeña, basándose estrictamente en la evidencia empírica de este estudio, el Random Forest se selecciona como el modelo con el mejor rendimiento predictivo.

### *4.5.3. Discusión de los Hallazgos*

El análisis comparativo revela varias conclusiones importantes:

- **Superioridad de los modelos no lineales:** La drástica mejora en el rendimiento al pasar del modelo lineal a los modelos basados en árboles (Random Forest y XGBoost) indica que las relaciones entre las variables geográficas, el número de clientes y el consumo energético no son puramente lineales. Los modelos de conjunto son capaces de capturar interacciones complejas y patrones que el modelo lineal ignora, como, por ejemplo, que el consumo *per cápita* puede variar sistemáticamente entre diferentes provincias o cantones.

- **Importancia crítica del "Número de Clientes":** Ambos modelos de alto rendimiento, Random Forest y XGBoost, coinciden inequívocamente en que la variable Numero Clientes es el predictor más influyente por un amplio margen. Este hallazgo, aunque intuitivo, tiene una implicación práctica fundamental: para las empresas de distribución eléctrica y los planificadores energéticos, disponer de un censo preciso y actualizado del número de clientes por parroquia es el dato más valioso para realizar pronósticos de demanda a corto y medio plazo.

- **Relevancia de las variables geográficas:** A pesar del dominio del número de clientes, las variables geográficas (Provincia, Cantón, Parroquia) y la temporal (Mes) demuestran tener una importancia predictiva no nula. Actúan como factores de ajuste

que permiten al modelo refinar sus predicciones. Esto sugiere que existen patrones de consumo endémicos de cada región que van más allá del simple número de usuarios, probablemente relacionados con factores socioeconómicos, climáticos o culturales propios de cada localidad.

## 4.6. Conclusión del capitulo

El análisis de resultados ha permitido evaluar y comparar de forma sistemática tres modelos de aprendizaje automático para la predicción del consumo eléctrico residencial. Se ha demostrado que los modelos de conjunto, Random Forest y XGBoost, son extraordinariamente efectivos para esta tarea, alcanzando un $R^2$ superior a 0.99.

Con base en una ligera superioridad en las métricas de error y en el coeficiente de determinación, se ha identificado al Random Forest como el modelo óptimo entre los evaluados. Más allá de la selección del modelo, el análisis de importancia de variables ha confirmado que el número de clientes es el principal motor del consumo energético, si bien las características geográficas aportan un valor predictivo secundario pero significativo.

Estos hallazgos no solo validan la metodología propuesta, sino que también proporcionan una base sólida para las conclusiones generales del presente trabajo, sus limitaciones y las futuras líneas de investigación que se abordarán en el capítulo final.

# CAPITULO 5: CONCLUSIONES Y RECOMENDACIONES

## 5.1. CONCLUSIONES

Se pudo recopilar un datset de datos historicos sobre sobre el consumo de energía en los hogares que incluye valores del consumo en kWh como detalles sobre lugar (provincia, cantón y parroquia), mes del año y número de clientes por área. La elección de este conjunto resultó excelente, ya que tiene los elementos importantes para entender la necesidad de energía en distintos lugares y tiempos, permitiendo un análisis claro de los hábitos del consumo.

El preprocesamiento del dataset se basó con la metodología CRISP-DM, en el cual se ejecutó la limpieza, transformación y carga de los datos proceso ETL. Con este proceso se depuro las inconsistencias, como: duplicidad de datos, valores faltantes, nulos, etc. Además, la conversión de variables categóricas para los algoritmos de aprendizaje automático. Este proceso es crucial para obtener datos de calidad y obtener resultados reales.

El análisis del rendimiento de los modelos se aplicaron tres modelos con enfoques predictivos: Regresión Lineal, Random Forest y XGBoost. La regresión lineal fue el modelo base con un 80.5%, $R^2 = 0.8049$. Con errores de predicción (MAE de 44,309.91 kWh y RMSE de 74,643.14 kWh) fueron elevados. Además, el análisis de residuos evidenció problemas de heterocedasticidad, lo que limita su uso en escenarios donde se requiere alta precisión, especialmente para consumos elevados. Sin embargo, el Random Forest superó modelo lineal, alcanzando un $R^2$ de 0.9906, reduciendo los errores de predicción (MAE de 7,884.33 kWh y RMSE de 16,415.11 kWh). El resultado de los gráficos de residuos presentó homocedasticidad, y finalmente, XGBoost se obtuvo un rendimiento aceptable ($R^2$ de 0.9901). De acuerdo a la evaluación de las métricas se establece que el modelo Random

Forest ofrece el mejor desempeño predictivo para este caso de estudio. Además, se se podría decir que el número de Clientes es la variable principal en la predicción del consumo de energía en los hogares-.

## 5.2. RECOMENDACIONES

Se recomienda actualizar el dataset con variables externas, como clima, indicadores socioeconómicos, indicadores geográficos, son factores pueden influir en consumo energético de los hogares. Añadir estas variables se podría obtener una mejora en la precisión de predicción de los modelos.

Tomando en cuenta que el modelo Random Forest mostró mejor rendimiento predictivo, se sugiere implementar un demo en una entidad de servicio de energía, para evaluar el rendimiento en la situación real. Y evaluar como mejora la demanda energética y permita diseñar estrategias eficientes para la distribución de energía, evitando los escases del servicio o sobrecargas de consumo.

Se recomienda establecer mecanismos de forma periódica la verificación y actualización del número de clientes por parroquia, cantón y provincia, al contar con datos precisos garantizaría la efectividad de los modelos de predicción que en este estudio se ha evaluado.

## 5.3 Bibliografía

Aco, A., Hancco, B., & Pérez, Y. (2023). Análisis comparativo de Técnicas de Machine Learning para la predicción de casos de deserción universitaria. *Revista lbérica de Sistemas e Tecnologias de Informatica RISTI*. doi: 10.17013/risti.51.84–98

Ahmad, I. (2024). *50 Algoritmos que todo programador debe conocer* (Segunda ed.). Marcombo. Obtenido de

https://www.google.com.ec/books/edition/50_algoritmos_que_todo_programador_debe/bxElEQAAQBAJ?hl=es&gbpv=1&dq=StandardScaler,+escalado&pg=PT246&printsec=frontcover

aichallenge. (24 de 06 de 2021). *https://aichallenge.utec.edu.uy.* Obtenido de https://aichallenge.utec.edu.uy/community/machine-learning/eligiendo-buenas-variables-a-traves-de-coeficientes-de-correlacion/

Amézquita, J., & Eslava, E. (2022). Supervised Learning for data cleaning in the coherence and completeness dimensions. *Scielo, 24*(2). doi:https://doi.org/10.25100/iyc.v24i2.11361

Amor, N., Tayyab, M., Petru, M., & Neethu, S. (2023). A review on computational intelligence methods for modeling of light weight composite materials. *ELSEVIER*. doi:https://doi.org/10.1016/j.asoc.2023.110812

Arellano, A., & Peña, D. (2020). Linear regression models for predicting drinking water consumption. *NOVASINERGIA, 3*(1), 27-36. doi: 10.37135/ns.01.05.03

BRZOZOWSKA, J., PIZOŃ, J., BAYTIKENOVA, G., GOLA, A., ZAKIMOVA, .. A., & PIOTROWSKA, K. (2023). INGENIERÍA DE DATOS EN DATOS DE PRODUCCIÓN DEL PROCESO CRISP-DM – ESTUDIO DE CASO. *Informática Aplicada, 19*(3), 83–95. doi:https://doi.org/10.35784/acs-2023-26

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. doi:http://dx.doi.org/10.1145/2939672.2939785

Comisión Nacional de los Mercados y la Competencia. (12 de 2024). *https://www.cnmc.es*. Obtenido de https://www.cnmc.es/prensa/panel-hogares-servicios-electricidad-gas-20241205: https://www.cnmc.es/prensa/panel-hogares-servicios-electricidad-gas-20241205

Contreras Bravo, L. E., & Padilla Beltrán, J. E. (2024). *Ciencia de datos con Python* (Primera ed.). Bogota, Colombia: Ediciones de la U. Obtenido de https://www.google.com.ec/books/edition/Ciencia_de_datos_con_python/laY3EQAAQBAJ?hl=es&gbpv=1&dq=xgboost&pg=PA271&printsec=frontcover

datascientest. (01 de 09 de 2022). *https://datascientest.com/es/scikit-learn-decubre-la-biblioteca-python.*

Enerdata. (2024). *https://datos.enerdata.net*. Obtenido de https://datos.enerdata.net/energia-total/datos-consumo-internacional.html: https://datos.enerdata.net/energia-total/datos-consumo-internacional.html

Espinosa-Zúñiga, J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *redalyc, XXI*(3). doi:https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Espinosa-Zúñiga, J. J. (17 de 1 de 2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, investigación y*

*tecnología, XXI*(1), 1-18. Obtenido de https://www.revistaingenieria.unam.mx/numeros/v21n1-08.php

Espinoza, J. (2020). Application of Random Forest and XGBoost algorithms based on a credit card applications database. *Ing. invest. y tecnol. Scielo, 21*(3). doi:https://doi.org/10.22201/fi.25940732e.2020.21.3.022

fastercapital. (12 de 05 de 2025). *http://fastercapital.com.* Obtenido de http://fastercapital.com/es/contenido/Homoskedasticidad-y-OLS--la-columna-vertebral-de-la-regresion-lineal.html

Forero Corba, W., & Negre Bennasar, F. (2024). Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: una revisión sistemática. *RIED-Revista Iberoamericana de Educación a Distancia, 27*(1). doi:https://doi.org/10.5944/ried.27.1.37491

Gatica Oyarzún, N. D. (2024). Metodologías de división de datos selectiva mediante optimización para modelos de aprendizaje de máquinas. Chile. Obtenido de https://repositorio.uchile.cl/handle/2250/203462

Gonzalez, G. (2023). *Modelo de Machine Learning para la eficiencia operativa del portafolio de productos y servicios del área de Cultura, Recreación y deporte de una Caja de Compensación en Colombia.* Maestria en Inteligencia de Negocios, Universidad EAN. Obtenido de https://repository.universidadean.edu.co/server/api/core/bitstreams/0497240f-8b58-49ef-9b6b-9aaa1483c6ab/content

IBM. (17 de 08 de 2021). *https://www.ibm.com*. Obtenido de https://www.ibm.com/docs/es/spss-modeler//saas?topic=dm-crisp-help-overview#crisp_overview

Incio Puyen, J. a. (10 de 04 de 2021). Librerias de python. Obtenido de https://es.scribd.com/document/517853061/Librerias-Python

Instituto de Investigación Geológico y Energético . (17 de 12 de 2020). *https://www.geoenergia.gob.ec*. doi:https://www.geoenergia.gob.ec/category/noticias/page/25/

Jijo, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *ournal of Applied Science and Technology Trends, 2*(1), 20-28. doi:https://doi.org/10.38094/jastt20165

Khaoula, E., Amine, B., Mostafa, B., & Deifalla, A. F. (2024). Machine Learning Algorithms for Predicting Energy Consumption in Educational Buildings. *International journal of Energy Research*. doi: https://doi.org/10.1155/2024/6812425

Lozada Cabrera, C. H., Apolo, A., P., N., Paredes Moran, D. V., & Vique Salazar, G. I. (2022). Predicción de demanda eléctrica utilizando redes neuronales artificiales para un sistema de distribución de energía eléctrica. *Ciencia Latina Revista Científica Multidisciplinar, 6*(3), 3871-3894. doi: https://doi.org/10.37811/cl_rcm.v6i3.2502

Luo, T., & Chen, M. (2025). Advancements in supervised machine learning for outdoor thermal comfort: A comprehensive systematic review of scales, applications, and data types. *Elsevier ScienceDirect, 329*. doi:https://doi.org/10.1016/j.enbuild.2024.115255

Mejía Vásquez, E. J., & Gonzalez Chavez, S. (2019). Prediction of residential electric power consumption in the Cajamarca Region through Holt -Winters models. *Energética, 40*(3). doi:1815-5901

Menacho, C. (2023). Lineal regression models with neural networks. *Dialnet, 75*(2), 253-260. doi:http://dx.doi.org/10.21704/ac.v75i2.961

Meneses Díez, M., Crespo García, D., & Monzo Sanchez, C. (27 de 12 de 2021). Aplicación de Machine Learning al consumo eléctrico de edificios inteligentes. Catalunya. Obtenido de https://openaccess.uoc.edu/bitstream/10609/138247/6/mmenesesdTFM0122memoria.pdf?utm_source=chatgpt.com

Mohsin Abdulazeez, A., & Hussen, D. (2021). Una revisión sobre la regresión lineal integral en el aprendizaje automático. *Revista de tendencias en ciencia y tecnología aplicadas, 1*(4), 140-147. doi: 10.38094/jastt1457

Molina, C., & Bonilla, F. (2024). Application of CRISP-DM Methodology in the Analysis of Dissolved Gases in Dielectric oil of Electrical Transformers in the Ecuadorian Electrical sector. *Scielo Revista Técnica energía, 21*(1). doi:https://doi.org/10.37116/revistaenergia.v21.n1.2024.635

Montenegro, M., Menchaca, R., & Menchaca, R. (2023). A gently but rigorous introduction to reinforcement learning. *ReCIBE. Revista electrónica de Computación, Informática, Biomédica y Electrónica, 12*(1), 1-15. Obtenido de https://www.redalyc.org/journal/5122/512275598001/html/

Nazari, R. (2024). *Estadística para principiantes.* Obtenido de https://www.google.com.ec/books/edition/ESTAD%C3%8DSTICAS_PARA_PRINCIPIANTES_La_gu%C3%AD/zjb6EAAAQBAJ?hl=es&gbpv=1&dq=IQR+(Rango+intercuartil)&pg=PA62&printsec=frontcover

Ortega, L., & Cardenas, J. (2022). Estrategias de predicción de consumo energético en edificaciones: una revisión. *TecnoLógicas, 26*(58). doi:https://doi.org/10.22430/22565337.2650

Pedrero, V., Reynaldos, K., Ureta, J., & Cortez, E. (2021). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. *Revista médica de Chile, 49*(2). doi:http://dx.doi.org/10.4067/s0034-98872021000200248

Pineda Pertuz, C. M. (10 de 2022). *Aprendizaje automático y profundo en Python* (Primera ed.). Bogota, Colombia: Ediciones de la U. Obtenido de https://www.google.com.ec/books/edition/Aprendizaje_autom%C3%A1tico_y_profundo_en_py/mgNcEAAAQBAJ?hl=es&gbpv=1&dq=LabelEncoder+que+es&pg=PA77&printsec=frontcover

Ramérez Gil, C. M. (2023). *Programación de Inteligencia Artificial. Curso Práctico.* Bogota, Colombia: Ediciones de la U. Obtenido de https://www.google.com.ec/books/edition/Programaci%C3%B3n_de_inteligencia_artificial/fnYJEQAAQBAJ?hl=es&gbpv=1&dq=statsmodels+que+es&pg=PA90&printsec=frontcover

Salinas, J., Garciá, A., Riveros, D., Gonzalez, A., & Goanzalesz, A. (2024). Annual Daily Irradiance Analysis of Clusters in Mexico by Machine Learning Algorithms. *MPDI, 16*(4).

Sheikh, C. (15 de 05 de 2025). *https://studyeasy.org/articles/.* Obtenido de https://studyeasy.org/es/course-articles/machine-leaning-articles-es/test-and-train-data-split-and-feature-scaling-es/

Soporte de Minitab. (2025). *https://support.minitab.com.* Obtenido de https://support.minitab.com/es-mx/minitab/help-and-how-to/statistical-modeling/regression/how-to/fit-regression-model/interpret-the-results/all-statistics-and-graphs/residual-plots/

statsmodels. (03 de 10 de 2024). *https://www.statsmodels.org/stable/index.html.*

Vaish, R., Dwivedi, U., Tewari, S., & Tripathi, S. (2021). Machine learning applications in power system fault diagnosis: Research advancements and perspectives. *ELSEIVER*. doi:https://doi.org/10.1016/j.engappai.2021.104504

Yazici, İ., Ibraheem, S., & Jafi, D. (2023). A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems. *Engineering Science and Technology, an International Journal, 44*. doi:https://doi.org/10.1016/j.jestch.2023.101455

Zhou, Y. (2022). Advances of machine learning in multi-energy district communities–mechanisms, applications and perspectives. *10*. doi:https://doi.org/10.1016/j.egyai.2022.100187