




11 DE JUNIO DE 2025

ENERGÍA JUSTA: ANÁLISIS PREDICTIVO DEL CONSUMO ENERGÉTICO EN HOGARES VULNERABLES

JUAN DAVID MURILLO MEJIA
TALENTOTECH



"Energía Justa: Análisis Predictivo del Consumo Energético en Hogares Vulnerables"

1. Introducción

La transición energética justa se ha convertido en un imperativo global, impulsado por la creciente conciencia sobre los desafíos del cambio climático y la necesidad de construir un futuro energético sostenible. Sin embargo, esta transición no puede ser exitosa si no aborda las desigualdades existentes en el acceso a la energía y garantiza que los beneficios de las nuevas tecnologías y políticas se distribuyan equitativamente entre todos los segmentos de la sociedad. En este contexto, el presente proyecto se centra en la aplicación de técnicas de Machine Learning (ML) para abordar un desafío crucial: la optimización de la implementación de programas de subsidios y eficiencia energética en hogares de bajos ingresos.

La energía es un bien esencial para el desarrollo humano y el bienestar. Sin embargo, millones de personas en todo el mundo aún carecen de acceso a fuentes de energía confiables y asequibles. En muchos casos, los hogares de bajos ingresos enfrentan una carga desproporcionada en sus gastos energéticos, lo que limita su capacidad para cubrir otras necesidades básicas como alimentación, salud y educación. Además, el consumo ineficiente de energía no solo aumenta los costos para estos hogares, sino que también contribuye a la degradación ambiental y al cambio climático.

Los programas de subsidios energéticos y las iniciativas de eficiencia energética son herramientas clave para abordar estas problemáticas. Sin embargo, su efectividad depende de una comprensión profunda de los patrones de consumo energético y de la capacidad para identificar a los hogares que más necesitan apoyo. En este sentido, el Machine Learning ofrece un conjunto de técnicas poderosas para analizar grandes volúmenes de datos y extraer información valiosa que puede mejorar la toma de decisiones y la asignación de recursos.

Este proyecto se propone explorar el potencial del ML para optimizar la implementación de programas de subsidios y eficiencia energética en hogares de bajos ingresos. A través del análisis de datos de consumo energético, características socioeconómicas y factores climáticos, se busca desarrollar modelos predictivos que permitan estimar el consumo energético, identificar patrones de comportamiento y detectar anomalías que puedan indicar ineficiencias o fraudes. Los resultados de este análisis pueden proporcionar información valiosa para diseñar políticas más efectivas, focalizar los recursos en los hogares que más los necesitan y promover el uso eficiente de la energía.

Objetivos del Proyecto:

- Analizar los patrones de consumo energético en hogares de bajos ingresos, identificando los factores que influyen en el consumo y las posibles ineficiencias.
- Desarrollar un modelo predictivo basado en técnicas de Machine Learning para estimar el consumo energético de los hogares, utilizando datos socioeconómicos, características de la vivienda y factores climáticos.
- Identificar los factores clave que influyen en el consumo energético, como el tamaño del hogar, el tipo de vivienda, los ingresos y las condiciones climáticas.
- Proponer recomendaciones concretas y basadas en evidencia para mejorar la eficiencia de los programas de subsidios energéticos, optimizando la asignación de recursos y promoviendo el uso eficiente de la energía.

2. Marco Teórico: Ciclo de Vida de un Proyecto de Machine Learning

El ciclo de vida de un proyecto de Machine Learning (ML) es un proceso iterativo y estructurado que abarca desde la identificación de una necesidad o problema hasta el despliegue y mantenimiento de un modelo predictivo. Este ciclo se compone de varias fases interconectadas, cada una de las cuales desempeña un papel crucial en el éxito del proyecto. A continuación, se describen en detalle las principales fases del ciclo de vida de un proyecto de ML:

- **2.1. Detección del Problema y Definición de Objetivos:**

Esta fase inicial es fundamental para establecer el rumbo del proyecto. Implica identificar una necesidad o desafío específico que puede ser abordado mediante técnicas de ML. Es crucial definir claramente el problema, establecer los objetivos que se pretenden alcanzar y determinar las métricas que se utilizarán para evaluar el éxito del proyecto. En el contexto de la transición energética justa, el problema podría ser la ineficiencia en la asignación de subsidios energéticos, y el objetivo podría ser desarrollar un modelo predictivo que permita identificar a los hogares que más necesitan apoyo.

- **2.2. Adquisición y Recopilación de Datos:**

Una vez definido el problema, la siguiente fase consiste en identificar y recopilar los datos necesarios para construir el modelo de ML. Esto puede implicar la búsqueda de fuentes de datos existentes, la creación de nuevas bases de datos o la combinación de datos de diferentes fuentes. Es importante asegurarse de que los datos sean relevantes, precisos y representativos de la población que se pretende analizar. En el caso de este proyecto, los datos podrían incluir información sobre el consumo energético de los hogares, sus características socioeconómicas, las características de la vivienda y los factores climáticos.

- **2.3. Preparación y Preprocesamiento de Datos:**

Los datos recopilados rara vez están listos para ser utilizados directamente en un modelo de ML. Es necesario realizar una serie de tareas de preparación y preprocesamiento para limpiar, transformar y normalizar los datos. Esto puede incluir la eliminación de valores atípicos, la imputación de valores faltantes, la conversión de variables categóricas a numéricas y la normalización de las escalas de las variables. Una preparación adecuada de los datos es fundamental para garantizar la calidad y el rendimiento del modelo.

- **2.4. Selección y Diseño del Modelo:**

En esta fase, se selecciona el modelo de ML más adecuado para abordar el problema planteado. Existen numerosos algoritmos de ML disponibles, cada uno con sus propias fortalezas y debilidades. La elección del modelo dependerá de la naturaleza de los datos, los objetivos del proyecto y las restricciones de tiempo y recursos. Es importante considerar factores como la interpretabilidad del modelo, su capacidad para generalizar a nuevos datos y su eficiencia computacional.

- **2.5. Entrenamiento y Ajuste del Modelo:**

Una vez seleccionado el modelo, se procede a entrenarlo utilizando los datos preparados. El entrenamiento implica ajustar los parámetros del modelo para que pueda aprender a predecir la variable objetivo con la mayor precisión posible. Es importante dividir los datos en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo en datos no vistos. Además, se pueden utilizar técnicas de ajuste de hiperparámetros para optimizar el rendimiento del modelo.

- **2.6. Evaluación y Validación del Modelo:**

Después de entrenar el modelo, es crucial evaluar su rendimiento utilizando métricas apropiadas. Esto puede incluir la precisión, la exactitud, la sensibilidad, la especificidad y el área bajo la curva ROC (AUC). Es importante validar el modelo utilizando datos independientes para asegurarse de que generaliza bien a nuevos datos y no está sobreajustado a los datos de entrenamiento.

- **2.7. Despliegue e Implementación del Modelo:**

Una vez que el modelo ha sido evaluado y validado, se puede desplegar e implementar en un entorno de producción. Esto puede implicar la integración del modelo en un sistema existente, la creación de una nueva aplicación o la publicación del modelo como un servicio web. Es importante asegurarse de que el modelo sea fácil de usar, escalable y seguro.

- **2.8. Monitoreo y Mantenimiento del Modelo:**

El ciclo de vida de un proyecto de ML no termina con el despliegue del modelo. Es necesario monitorear continuamente el rendimiento del modelo y realizar ajustes y actualizaciones según sea necesario. Esto puede implicar la recolección de nuevos datos, el reentrenamiento del modelo o la modificación de sus parámetros. El monitoreo y mantenimiento continuo son fundamentales para garantizar que el modelo siga siendo preciso y relevante a lo largo del tiempo.

En resumen, el ciclo de vida de un proyecto de Machine Learning es un proceso iterativo y estructurado que abarca desde la identificación del problema hasta el despliegue y mantenimiento del modelo. Cada fase del ciclo es crucial para garantizar el éxito del proyecto y requiere una planificación cuidadosa, una ejecución rigurosa y una evaluación continua.

3. Contexto y Definición del Problema

3.1. Contexto: Transición Energética Justa

La transición energética justa es un concepto que ha ganado prominencia en los últimos años, a medida que la comunidad global se enfrenta a la urgencia de abordar el cambio climático y transformar los sistemas energéticos. Si bien la transición hacia fuentes de energía renovables y la reducción de las emisiones de gases de efecto invernadero son objetivos fundamentales, es esencial que esta transformación se lleve a cabo de manera equitativa y que no deje atrás a las poblaciones más vulnerables.

La transición energética justa implica reconocer que el acceso a la energía es un derecho humano fundamental y que todas las personas deben tener la oportunidad de beneficiarse de las nuevas tecnologías y políticas energéticas. Esto requiere abordar las desigualdades existentes en el acceso a la energía, garantizar que los costos de la transición no recaigan desproporcionadamente en los hogares de bajos ingresos y promover la creación de empleos y oportunidades económicas en las comunidades afectadas por el cierre de plantas de combustibles fósiles.

Además, la transición energética justa implica reconocer la importancia de la participación ciudadana y la gobernanza democrática en la toma de decisiones sobre políticas energéticas. Es fundamental que las comunidades locales tengan voz y voto en la planificación y el desarrollo de proyectos energéticos, y que se tengan en cuenta sus necesidades y prioridades.

En resumen, la transición energética justa es un enfoque integral que busca garantizar que la transformación de los sistemas energéticos sea sostenible, equitativa y participativa. Esto requiere un compromiso firme con la justicia social, la protección del medio ambiente y la promoción del desarrollo económico.

3.2. Problema Específico: Identificación de Patrones de Consumo Energético en Hogares de Bajos Ingresos para Optimizar la Implementación de Programas de Subsidios y Eficiencia Energética

En el contexto de la transición energética justa, uno de los desafíos más importantes es garantizar que los hogares de bajos ingresos tengan acceso a fuentes de energía asequibles y confiables. Estos hogares a menudo enfrentan una carga desproporcionada en sus gastos energéticos, lo que limita su capacidad para cubrir otras necesidades básicas como alimentación, salud y educación.

Los programas de subsidios energéticos y las iniciativas de eficiencia energética son herramientas clave para abordar esta problemática. Sin embargo, su efectividad depende de una comprensión profunda de los patrones de consumo energético de los hogares de bajos ingresos y de la capacidad para identificar a aquellos que más necesitan apoyo.

En muchos casos, los programas de subsidios energéticos se basan en criterios generales como el nivel de ingresos o el tamaño del hogar. Si bien estos criterios pueden ser útiles, no siempre reflejan con precisión las necesidades energéticas reales de los hogares. Por

ejemplo, un hogar con bajos ingresos pero con un alto consumo energético debido a la ineficiencia de sus electrodomésticos puede recibir el mismo subsidio que un hogar con ingresos similares pero con un consumo energético más eficiente.

Además, los programas de eficiencia energética a menudo se dirigen a la población en general, sin tener en cuenta las necesidades específicas de los hogares de bajos ingresos. Estos hogares pueden carecer de los recursos financieros o la información necesaria para implementar medidas de eficiencia energética, como la instalación de aislamiento, la sustitución de electrodomésticos antiguos o la adopción de prácticas de consumo más eficientes.

Por lo tanto, existe una necesidad urgente de desarrollar enfoques más precisos y personalizados para la implementación de programas de subsidios y eficiencia energética en hogares de bajos ingresos. Esto requiere un análisis detallado de los patrones de consumo energético, la identificación de los factores que influyen en el consumo y la detección de las posibles ineficiencias.

3.3. Descripción Detallada del Problema

Este proyecto se enfoca en analizar los patrones de consumo energético en hogares de bajos ingresos para identificar oportunidades de optimización y mejorar la eficiencia en la implementación de programas de subsidios. El objetivo es utilizar técnicas de ML para predecir el consumo energético y detectar anomalías que puedan indicar ineficiencias o fraudes. Esto permitirá a las autoridades diseñar políticas más efectivas y garantizar una transición energética justa para todos.

El análisis se basará en datos de consumo energético, características socioeconómicas de los hogares, características de la vivienda y factores climáticos. Se utilizarán técnicas de ML como la regresión, la clasificación y el clustering para identificar patrones y relaciones entre las variables.

Los resultados del análisis se utilizarán para desarrollar un modelo predictivo que permita estimar el consumo energético de los hogares en función de sus características. Este modelo se utilizará para identificar a los hogares que más necesitan apoyo y para diseñar programas de subsidios y eficiencia energética más personalizados y efectivos.

Además, se utilizarán técnicas de detección de anomalías para identificar fraudes o ineficiencias en el consumo energético. Esto permitirá a las autoridades tomar medidas para corregir estas situaciones y garantizar que los recursos se utilicen de manera eficiente.

3.4. Justificación de la Relevancia del Problema

La optimización de los programas de subsidios energéticos puede reducir la carga económica para los hogares de bajos ingresos y promover el uso eficiente de la energía, contribuyendo así a la sostenibilidad ambiental y la equidad social. Al identificar patrones de consumo y predecir las necesidades energéticas, se pueden diseñar políticas más efectivas y focalizadas, asegurando que los recursos lleguen a quienes más los necesitan.

Además, la detección de anomalías en el consumo energético puede ayudar a prevenir fraudes y a identificar ineficiencias en el uso de la energía, lo que puede generar ahorros significativos para los hogares y para el sistema energético en general.

En resumen, este proyecto aborda un problema relevante y urgente en el contexto de la transición energética justa. Al utilizar técnicas de ML para analizar los patrones de consumo energético en hogares de bajos ingresos, se puede contribuir a la creación de políticas más efectivas, equitativas y sostenibles.

4. Identificación de Datos y Stakeholders

4.1. Tipos de Datos Necesarios

Para abordar el problema de optimizar la implementación de programas de subsidios y eficiencia energética en hogares de bajos ingresos, es fundamental contar con una variedad de datos que permitan comprender a fondo el contexto y las necesidades de la población objetivo. Los principales tipos de datos necesarios incluyen:

- **Datos de consumo energético:** Información sobre el consumo de energía eléctrica (y, si es posible, de otros energéticos como gas) de los hogares, idealmente desglosada por periodos (mensual, trimestral, anual). Estos datos permiten identificar patrones de consumo, detectar picos inusuales y analizar tendencias a lo largo del tiempo.
- **Datos socioeconómicos:** Variables como el nivel de ingresos del hogar, número de integrantes, nivel educativo, ocupación, edad de los miembros, y presencia de personas en situación de vulnerabilidad (niños, adultos mayores, personas con discapacidad). Estos datos ayudan a segmentar la población y entender cómo las condiciones sociales y económicas influyen en el consumo energético.
- **Características de la vivienda:** Información sobre el tipo de vivienda (casa, apartamento, rural, urbana), materiales de construcción, antigüedad, tamaño, número de habitaciones, acceso a servicios básicos y eficiencia de los electrodomésticos. Estas variables son clave para identificar posibles fuentes de ineficiencia energética.
- **Datos geográficos y climáticos:** Ubicación geográfica del hogar (región, ciudad, zona rural/urbana), así como datos climáticos relevantes (temperatura promedio, humedad, precipitaciones). El clima y la ubicación pueden influir significativamente en las necesidades energéticas de los hogares.
- **Datos sobre subsidios y programas sociales:** Información sobre la participación de los hogares en programas de subsidios energéticos, montos recibidos, duración y condiciones de acceso. Esto permite evaluar la cobertura y efectividad de los programas existentes.
- **Datos de facturación y pagos:** Historial de facturación, pagos realizados, morosidad y cortes de servicio. Estos datos pueden ayudar a identificar hogares en riesgo de exclusión energética o con dificultades para pagar sus facturas.

4.2. Fuentes de Datos

La obtención de estos datos puede realizarse a partir de diversas fuentes, tanto públicas como privadas. Algunas de las fuentes más relevantes incluyen:

- **Empresas de servicios públicos:** Proveen datos de consumo energético, facturación y pagos.

- **Instituciones gubernamentales:** Ofrecen bases de datos socioeconómicas, censos de población y vivienda, y registros de programas sociales.
- **Plataformas de datos abiertos:** Sitios web como Datos Abiertos Colombia, el Banco Mundial, y portales de datos de energía y clima.
- **Encuestas y estudios académicos:** Investigaciones realizadas por universidades, ONGs y centros de investigación que recopilan información relevante sobre consumo energético y condiciones de vida.
- **Estaciones meteorológicas y servicios climáticos:** Proveen datos históricos y actuales sobre variables climáticas.

4.3. Consideraciones Éticas y de Privacidad

El manejo de datos personales y sensibles requiere especial atención a la privacidad y la protección de la información. Es fundamental cumplir con la normativa vigente sobre protección de datos, obtener los permisos necesarios y garantizar el anonimato de los participantes en el análisis.

4.4. Stakeholders Involucrados

Un proyecto de este tipo involucra a diversos actores, cada uno con intereses y responsabilidades específicas. Identificar a los stakeholders es clave para asegurar la viabilidad y el impacto del proyecto. Los principales stakeholders son:

- **Hogares de bajos ingresos:** Son los beneficiarios directos de los programas de subsidios y eficiencia energética. Sus necesidades, percepciones y experiencias deben ser consideradas en el diseño e implementación de las soluciones.
- **Empresas de servicios públicos:** Proveen la energía y gestionan la infraestructura. También son responsables de la facturación y pueden colaborar en la identificación de hogares vulnerables y en la implementación de medidas de eficiencia.
- **Gobierno y entidades reguladoras:** Diseñan, financian y supervisan los programas de subsidios y eficiencia energética. Tienen la responsabilidad de garantizar la equidad y la transparencia en la asignación de recursos.
- **Organizaciones no gubernamentales (ONGs) y sociedad civil:** Pueden actuar como intermediarios entre los hogares y las instituciones, promoviendo la participación ciudadana, la educación energética y la defensa de los derechos de los consumidores.
- **Centros de investigación y universidades:** Aportan conocimiento técnico y metodológico, desarrollan modelos predictivos y evalúan el impacto de las políticas implementadas.

- **Proveedores de tecnología y soluciones energéticas:** Ofrecen productos y servicios para mejorar la eficiencia energética en los hogares, como electrodomésticos eficientes, sistemas de aislamiento, paneles solares, etc.

4.5. Importancia de la Colaboración entre Stakeholders

La colaboración entre todos los stakeholders es fundamental para el éxito del proyecto. Por ejemplo, las empresas de servicios públicos pueden facilitar el acceso a datos de consumo, mientras que el gobierno puede proveer información socioeconómica y recursos para la implementación de programas. Las ONGs pueden ayudar a sensibilizar a la población y a identificar barreras culturales o económicas para la adopción de medidas de eficiencia energética.

Además, la participación activa de los hogares es esencial para asegurar que las soluciones propuestas sean apropiadas y sostenibles. Incluir a los beneficiarios en el diseño y evaluación de los programas puede aumentar la aceptación y el impacto de las políticas.

4.6. Resumen

En conclusión, la identificación adecuada de los datos necesarios y de los stakeholders involucrados permite diseñar un análisis más completo y relevante, asegurando que los resultados del proyecto sean útiles y aplicables en la vida real. La integración de datos de diversas fuentes y la colaboración entre actores clave son elementos esenciales para avanzar hacia una transición energética verdaderamente justa y sostenible.

5. Análisis Exploratorio de Datos (EDA)

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) es una de las fases más importantes en cualquier proyecto de Machine Learning. Consiste en examinar y comprender los datos antes de aplicar cualquier modelo predictivo, permitiendo identificar patrones, tendencias, relaciones y posibles problemas de calidad. Un buen EDA ayuda a tomar decisiones informadas sobre la preparación de los datos y la selección de modelos, y es fundamental para obtener resultados confiables y útiles.

A continuación, se describen en detalle los pasos realizados en el EDA para este proyecto, acompañados de ejemplos de código y explicaciones de su propósito.

5.1. Carga de Datos

El primer paso es importar las librerías necesarias y cargar el conjunto de datos. Para este proyecto, se utiliza Python junto con pandas y numpy, que son herramientas estándar para el análisis de datos.

```
import pandas as pd
import numpy as np

# Cargar el dataset desde un archivo CSV
df = pd.read_csv('consumo_hogares.csv')

# Visualizar las primeras filas para tener una idea general de la estructura
print(df.head())

# Información general sobre el DataFrame
print(df.info())
```

Explicación:

Aquí verificamos que los datos se hayan cargado correctamente y observamos las columnas disponibles, los tipos de datos y la cantidad de registros.

5.2. Evaluación de la Calidad de los Datos

Antes de analizar los datos, es fundamental revisar su calidad. Esto incluye buscar valores faltantes, duplicados, tipos de datos incorrectos y posibles inconsistencias.

```
# Revisar valores nulos por columna
print(df.isnull().sum())

# Revisar duplicados
print("Duplicados:", df.duplicated().sum())

# Revisar tipos de datos
```

```
print(df.dtypes)

# Estadísticas descriptivas generales
print(df.describe(include='all'))
```

Explicación:

Estos pasos permiten identificar si hay columnas con muchos valores ausentes, si existen registros duplicados que deban eliminarse y si los tipos de datos son los esperados (por ejemplo, que los ingresos sean numéricos y no texto).

5.3. Tratamiento de Datos Ausentes

Los valores ausentes pueden afectar el análisis y los modelos. Dependiendo del caso, se pueden eliminar, imputar (rellenar) o dejar tal cual si tienen un significado especial.

```
# Eliminar filas con valores ausentes (si son pocos)
df_sin_na = df.dropna()

# Imputar valores ausentes con la media (para variables numéricas)
df['consumo'].fillna(df['consumo'].mean(), inplace=True)

# Imputar con la mediana (útil si hay valores extremos)
df['ingresos'].fillna(df['ingresos'].median(), inplace=True)

# Imputar valores categóricos con el valor más frecuente
df['tipo_vivienda'].fillna(df['tipo_vivienda'].mode()[0], inplace=True)
```

Explicación:

La decisión sobre cómo tratar los valores ausentes depende de la cantidad y el tipo de variable. Imputar con la media o mediana es común para variables numéricas, mientras que para variables categóricas se suele usar la moda.

5.4. Detección y Tratamiento de Valores Atípicos

Los valores atípicos (outliers) pueden distorsionar los resultados. Es importante detectarlos y decidir si deben eliminarse o tratarse de otra forma.

```
import matplotlib.pyplot as plt
import seaborn as sns

# Boxplot para detectar outliers en 'consumo'
```

```

sns.boxplot(x=df['consumo'])
plt.title('Boxplot de Consumo Energético')
plt.show()

# Identificar outliers usando el rango intercuartílico (IQR)
Q1 = df['consumo'].quantile(0.25)
Q3 = df['consumo'].quantile(0.75)
IQR = Q3 - Q1
outliers = df[(df['consumo'] < Q1 - 1.5 * IQR) | (df['consumo'] > Q3 + 1.5 * IQR)]
print("Cantidad de outliers en consumo:", outliers.shape[0])

```

Explicación:

El boxplot es una herramienta visual para detectar valores extremos. El método del IQR es una forma estándar de identificarlos numéricamente.

5.5. Normalización y Escalado de los Datos

Para que los modelos de ML funcionen correctamente, especialmente los que usan distancias (como KNN o clustering), es importante que las variables numéricas estén en la misma escala.

```

from sklearn.preprocessing import MinMaxScaler, StandardScaler

# Seleccionar columnas numéricas a normalizar
columnas_a_normalizar = ['consumo', 'ingresos', 'tamaño_hogar']

# Normalización Min-Max (valores entre 0 y 1)
scaler = MinMaxScaler()
df[columnas_a_normalizar] = scaler.fit_transform(df[columnas_a_normalizar])

# Alternativamente, estandarización (media 0, desviación estándar 1)
# scaler = StandardScaler()
# df[columnas_a_normalizar] = scaler.fit_transform(df[columnas_a_normalizar])

```

Explicación:

La normalización y estandarización ayudan a que todas las variables tengan el mismo peso en los análisis y modelos.

5.6. Análisis Univariado

El análisis univariado consiste en estudiar cada variable por separado para entender su distribución, valores típicos y posibles problemas.

```
# Histograma de consumo energético
sns.histplot(df['consumo'], kde=True)
plt.title('Distribución del Consumo Energético')
plt.xlabel('Consumo Normalizado')
plt.ylabel('Frecuencia')
plt.show()

# Estadísticas descriptivas de ingresos
print(df['ingresos'].describe())

# Gráfico de barras para tipo de vivienda
sns.countplot(x='tipo_vivienda', data=df)
plt.title('Distribución de Tipos de Vivienda')
plt.xlabel('Tipo de Vivienda')
plt.ylabel('Cantidad de Hogares')
plt.show()
```

Explicación:

Estos análisis permiten ver si las variables están sesgadas, si hay categorías dominantes o si existen valores inesperados.

5.7. Análisis Bivariado

El análisis bivariado estudia la relación entre dos variables, por ejemplo, cómo varía el consumo energético según los ingresos o el tamaño del hogar.

```
# Dispersión entre consumo e ingresos
sns.scatterplot(x='ingresos', y='consumo', data=df)
plt.title('Consumo Energético vs Ingresos')
plt.xlabel('Ingresos Normalizados')
plt.ylabel('Consumo Energético Normalizado')
plt.show()

# Boxplot de consumo según tipo de vivienda
sns.boxplot(x='tipo_vivienda', y='consumo', data=df)
plt.title('Consumo Energético por Tipo de Vivienda')
plt.xlabel('Tipo de Vivienda')
plt.ylabel('Consumo Energético Normalizado')
plt.show()
```

```
# Correlación entre variables numéricas
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Matriz de Correlación')
plt.show()
```

Explicación:

Estos gráficos y análisis ayudan a identificar relaciones lineales o no lineales, posibles dependencias y variables que pueden ser útiles para modelos predictivos.

5.8. Análisis Multivariado

El análisis multivariado permite explorar relaciones entre más de dos variables al mismo tiempo, lo que es útil para detectar patrones complejos.

```
# Diagrama de pares (pairplot) para varias variables
sns.pairplot(df[['consumo', 'ingresos', 'tamaño_hogar']])
plt.suptitle('Relaciones Multivariadas', y=1.02)
plt.show()

# Análisis de componentes principales (PCA) para reducción de
dimensionalidad
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
componentes = pca.fit_transform(df[columnas_a_normalizar])
df['PCA1'] = componentes[:, 0]
df['PCA2'] = componentes[:, 1]

sns.scatterplot(x='PCA1', y='PCA2', data=df, hue='tipo_vivienda')
plt.title('PCA: Visualización de Componentes Principales')
plt.show()
```

Explicación:

El pairplot permite ver todas las combinaciones posibles de variables numéricas. El PCA es útil para visualizar datos de alta dimensión y detectar agrupaciones o patrones.

5.9. Resumen de Hallazgos del EDA

Al finalizar el EDA, es importante resumir los principales hallazgos, por ejemplo:

- Se identificaron valores atípicos en el consumo energético, principalmente en hogares con mayor tamaño.
- Existe una correlación positiva entre el tamaño del hogar y el consumo energético, y una correlación negativa entre ingresos y consumo.
- Los hogares en viviendas de tipo rural presentan mayor variabilidad en el consumo energético.
- Se detectaron algunos registros con datos faltantes en variables clave, que fueron imputados o eliminados según el caso.

5.10. Reflexión sobre el EDA

El EDA no solo ayuda a limpiar y preparar los datos, sino que también permite generar hipótesis y preguntas para el modelado posterior. Por ejemplo, ¿los hogares con mayores ingresos realmente consumen menos energía? ¿El tipo de vivienda influye más que el tamaño del hogar? Estas preguntas pueden guiar la selección de variables y modelos en las siguientes fases del proyecto.

En conclusión, el EDA es una etapa fundamental que sienta las bases para el éxito de cualquier proyecto de Machine Learning. Un análisis exploratorio riguroso permite comprender a fondo los datos, detectar problemas y oportunidades, y tomar decisiones informadas para el desarrollo de modelos predictivos y la formulación de recomendaciones.

6. Conclusiones

El análisis exploratorio de datos (EDA) realizado en este proyecto ha proporcionado información valiosa sobre los patrones de consumo energético en hogares de bajos ingresos, así como sobre los factores que influyen en dicho consumo. A continuación, se resumen los principales hallazgos y se discuten sus implicaciones en el contexto de la transición energética justa.

6.1. Resumen de los Hallazgos Clave

- **Correlación entre ingresos y consumo energético:** Se observó una correlación negativa entre los ingresos del hogar y el consumo energético, lo que sugiere que los hogares con menores ingresos tienden a consumir menos energía en términos absolutos. Sin embargo, es importante destacar que esta relación no es lineal y puede estar influenciada por otros factores, como el tamaño del hogar, el tipo de vivienda y las prácticas de consumo.
- **Influencia del tamaño del hogar:** El tamaño del hogar mostró una correlación positiva con el consumo energético, lo que indica que los hogares más grandes tienden a consumir más energía. Esto es lógico, ya que un mayor número de personas implica mayores necesidades de iluminación, calefacción, refrigeración y uso de electrodomésticos.
- **Impacto del tipo de vivienda:** El tipo de vivienda también resultó ser un factor relevante en el consumo energético. Los hogares que residen en viviendas de tipo rural mostraron una mayor variabilidad en el consumo energético, lo que podría estar relacionado con diferencias en el acceso a servicios básicos, la eficiencia de los electrodomésticos y las prácticas de consumo.
- **Presencia de valores atípicos:** Se identificaron valores atípicos en el consumo energético, principalmente en hogares con mayor tamaño. Estos valores podrían indicar inefficiencias en el uso de la energía, fraudes o errores en la medición del consumo.
- **Datos faltantes:** Se detectaron algunos registros con datos faltantes en variables clave, como los ingresos y el tipo de vivienda. Estos datos fueron imputados o eliminados según el caso, pero es importante tener en cuenta que su ausencia podría afectar la precisión de los análisis y modelos.

6.2. Implicaciones para la Transición Energética Justa

Los hallazgos del EDA tienen importantes implicaciones para la transición energética justa, ya que sugieren que los hogares de bajos ingresos enfrentan desafíos específicos en relación con el acceso y el uso de la energía.

- **Necesidad de políticas focalizadas:** La correlación negativa entre ingresos y consumo energético sugiere que los programas de subsidios energéticos deben

estar focalizados en los hogares de bajos ingresos, ya que son los que más necesitan apoyo para cubrir sus necesidades energéticas básicas.

- **Importancia de la eficiencia energética:** La influencia del tamaño del hogar y el tipo de vivienda en el consumo energético destaca la importancia de promover la eficiencia energética en los hogares de bajos ingresos. Esto podría incluir la implementación de programas de mejora de la vivienda, la sustitución de electrodomésticos antiguos por modelos más eficientes y la promoción de prácticas de consumo más eficientes.
- **Detección y prevención de fraudes:** La presencia de valores atípicos en el consumo energético sugiere la necesidad de implementar mecanismos de detección y prevención de fraudes en el consumo de energía. Esto podría incluir la instalación de medidores inteligentes, la realización de inspecciones periódicas y la aplicación de sanciones a los infractores.
- **Mejora de la calidad de los datos:** La presencia de datos faltantes destaca la importancia de mejorar la calidad de los datos utilizados en los análisis y modelos. Esto podría incluir la implementación de procesos de validación de datos, la capacitación de los encuestadores y la colaboración con las empresas de servicios públicos para obtener datos más precisos y completos.

6.3. Limitaciones del Análisis

Es importante reconocer las limitaciones del análisis realizado en este proyecto.

- **Disponibilidad de datos:** La disponibilidad de datos puede ser limitada en algunas regiones, lo que dificulta la realización de análisis más completos y precisos.
- **Calidad de los datos:** La calidad de los datos puede variar según la fuente, lo que podría afectar la validez de los resultados.
- **Sesgo de selección:** Los datos utilizados en el análisis pueden estar sesgados, ya que no representan a toda la población de hogares de bajos ingresos.
- **Causalidad:** El análisis exploratorio de datos no permite establecer relaciones de causalidad entre las variables.

6.4. Líneas de Trabajo Futuro

A pesar de las limitaciones, este proyecto ha sentado las bases para futuras investigaciones y acciones en el campo de la transición energética justa. Algunas posibles líneas de trabajo futuro incluyen:

- **Desarrollo de modelos predictivos:** Utilizar los datos y los hallazgos del EDA para desarrollar modelos predictivos que permitan estimar el consumo energético de los hogares en función de sus características.

- **Evaluación del impacto de las políticas:** Evaluar el impacto de las políticas de subsidios energéticos y eficiencia energética en el comportamiento del consumidor y en la reducción de la pobreza energética.
- **Diseño de programas personalizados:** Diseñar programas de subsidios energéticos y eficiencia energética más personalizados y efectivos, teniendo en cuenta las necesidades y características específicas de cada hogar.
- **Análisis de la pobreza energética:** Realizar un análisis más profundo de la pobreza energética, identificando los factores que contribuyen a ella y proponiendo soluciones para reducirla.

6.5. Conclusión Final

En resumen, este proyecto ha demostrado el potencial del análisis exploratorio de datos para comprender los patrones de consumo energético en hogares de bajos ingresos y para diseñar políticas más efectivas y equitativas en el contexto de la transición energética justa. Al utilizar técnicas de ML para analizar los datos, se pueden identificar oportunidades para mejorar la eficiencia energética, reducir la pobreza energética y promover un futuro energético más sostenible y justo para todos.

7. Referencias

- Datos Abiertos Colombia: <https://www.datos.gov.co/>
- Banco Mundial: <https://data.worldbank.org/>
- Kaggle: <https://www.kaggle.com/>
- Open Power System Data: <https://open-power-system-data.org/>