UNIVERSIDAD CATÓLICA BOLIVIANA
“SAN PABLO”

Caso 5

Integrantes:

Álvaro Ariel Torres Calle

Kael Alessandro Lopez Castillo

Esmeralda Paula Medina Paredes

Camila Nicole Rodas Miranda

Franco Jhoel Parra Aguilar

Docente: Paton Gutierrez Ovidio Roger

Fecha: 24 / 02 / 2026

Materia: Machine Learning

1. Introducción

El presente análisis se fundamenta en el conjunto de datos "Energy Efficiency", orientado al estudio del desempeño térmico en edificaciones. Los datos se centran en la evaluación de la demanda energética mediante la simulación de 12 configuraciones volumétricas distintas, procesadas a través del software especializado Ecotect. El núcleo del estudio radica en entender cómo la geometría y las propiedades físicas de una estructura impactan directamente en sus requerimientos de climatización. Este enfoque permite optimizar el diseño arquitectónico desde las etapas preliminares para garantizar la sostenibilidad y la reducción del consumo energético operativo.

En este informe se busca crear modelos basados en la regresión lineal para predecir un valor continuo de la carga de calefacción (Heating Load) y un valor categórico basándonos en los resultados de Heating Load, de tal manera que se pueda clasificar una construcción como “Eficiente” o “Ineficiente”.

2. Contextualización del Conjunto de Datos

2.2. Metodología y Descripción del Dataset

Desde la página de descarga del dataset podemos obtener la siguiente información sobre la generación de los datos:

"Se llevó a cabo un análisis energético empleando 12 geometrías de edificios simuladas en el entorno Ecotect. Las estructuras presentan variaciones respecto al área de acristalamiento, su distribución y orientación, entre otros parámetros. Mediante la simulación de diversos ajustes en función de las características mencionadas, se generaron 768 muestras. El conjunto de datos resultante integra estas muestras y 8 variables predictoras con el propósito de estimar dos respuestas de valor real. Adicionalmente, el dataset permite su aplicación en problemas de clasificación multiclase mediante el redondeo de las respuestas al entero más cercano."

Los datos no fueron recolectados mediante mediciones físicas en edificios reales, sino que fueron generados a través de simulaciones informáticas utilizando el software Ecotect. El proceso metodológico para construir el dataset fue el siguiente:

Se partió de un "cubo elemental" con dimensiones de 3.5 × 3.5 × 3.5.

• A partir de este cubo, se crearon 12 formas de edificios distintas, donde cada forma está compuesta por 18 de estos cubos elementales.

• Se asumió el uso de los mismos materiales de construcción para todas las variaciones de edificios, eligiendo los materiales más comunes y nuevos de la industria.

• Se simularon diferentes configuraciones variando los parámetros (como el área de superficie y las dimensiones de los componentes), pero manteniendo un volumen constante para todos los edificios.

• A través de estas variaciones, se realizaron simulaciones exhaustivas que resultaron en un conjunto de datos de 768 edificios residenciales diversos.

Por tanto, sabemos que cada muestra representa un modelo de edificación único derivado de la variación de los siguientes factores: compacidad relativa, el área de superficie, el área de muros y techos, la altura total, la orientación cardinal, así como el área y la distribución del acristalamiento.

2.2.1. Definición de variables

Marcadas en verde estan las features, en amarillo las variables objetivo.

Nombre	Tipo	Descripción		¿Qué indica?
X1	Continuous	Relative Compactness Compacidad Relativa		Indica qué tan compacta es. Baja compacidad indica que está “esparcida”. Las casas compactas guardan mejor el calor.
X2	Continuous	Surface Area Área de Superficie		Longitud del área externa. A mayor área expuesta, más afecta el clima
X3	Continuous	Wall Area Área de Pared		Cuántos metros cuadrados de muros tiene la construcción
X4	Continuous	Roof Area Área de Techo		El tamaño del techo. El calor tiende a subir y escapar por arriba, o el sol directo puede calentar mucho la casa a través de él.
X5	Continuous	Overall Height Altura Total		Qué tan alta es la construcción. Cambia cómo circula el aire y el calor.
X6	Integer	Orientation Orientación		Hacia qué punto cardinal mira la fachada principal (norte, sur, este, oeste). Así se calcula cuántas horas de sol recibe la edificación en un día.

X7	Continuous	Glazing Area Área de Ventanales		Indica qué tanta parte de la casa son ventanas de vidrio. El vidrio deja pasar la luz, pero no aísla tan bien como una pared de ladrillo.
X8	Integer	Glazing Area Distribution Distribución de Ventanas		Indica si las ventanas están repartidas por igual en todos lados o si hay más ventanas en un lado específico
Y1	Continuous	Heating Load Carga de Calefacción		Energía necesaria para mantener la casa a una temperatura agradable cuando hace frío. A mayor carga de calefacción, mayor gasto.
Y2	Continuous	Cooling Load Carga de Refrigeración		Energía necesaria para mantener la casa a una temperatura agradable cuando hace calor. A mayor carga de refrigeración, mayor gasto.

2.2.2. Aclaraciones de las variables

La variable **Glazing_Area (X7)** se expresa como una proporción decimal del área total (Surface area), y no en unidades de superficie absolutas. Asimismo, los registros con valor cero representan modelos de control para el análisis de transferencia térmica base. En un caso real, las construcciones no podrían tener una área de ventanales igual a 0 (sin ventanales).

La variable **Glazing_Area_Distribution (X8)** tiene 6 escenarios posibles. Uno es sin acristalamiento (0); los otros 5 indican cómo se reparte el porcentaje de ventanas anterior: (1) Uniforme: 25% en cada fachada. (2) Norte: 55% en la fachada norte y 15% en las demás. Lo mismo aplica para las distribuciones (3) Este, (4) Sur y (5) Oeste. Esta variable posee una naturaleza categórica nominal.

La variable **Surface_Area (X8)** se calcula para entender la envolvente térmica de un edificio. En sí, un edificio básico es como una caja, tiene 6 caras: 4 paredes, 1 techo y 1 piso. Así es que se calcula Surface Area con la sumatoria del área del techo, de las paredes y del suelo.

La variable **Orientation (X6)** aunque codificada numéricamente en el dataset original, posee una naturaleza categórica nominal.

Relative_Compactness (X1) y Surface_Area (X2): Tienen 12 valores posibles cada una y son perfectamente inversamente proporcionales en este dataset, ya que todas las simulaciones mantuvieron un volumen constante de 771.75 m^3

Las variables **Altura total (X5) y Área del techo (X4):** La altura del edificio solo varió en 2 valores posibles, mientras que el área del techo tuvo 4. El análisis estadístico revela que ambas variables son casi inversamente proporcionales (con una altísima correlación negativa de -0.937)

Las variables **Compacidad relativa (X1) y Área de superficie (X2):** Tienen 12 valores posibles cada una y son perfectamente inversamente proporcionales en este dataset, ya que todas las simulaciones mantuvieron un volumen constante de 771.75 m^3

2.2.3. Condiciones físicas y climáticas de la simulación

Para generar los datos, no se asumió un cubo vacío y abstracto; se configuraron parámetros ambientales, geográficos y de uso muy específicos para representar un entorno residencial real:

• Ubicación y Clima: Todos los edificios fueron simulados en Atenas, Grecia.

• Ocupación y Actividad: Se asumió una vivienda con 7 personas realizando actividad sedentaria (70 W).

• Condiciones interiores de confort: Ropa estimada en 0.6 clo, humedad relativa del 60%, velocidad del aire de 0.30 m/s y un nivel de iluminación de 300 Lux.

• Climatización (HVAC): Modo mixto con 95% de eficiencia. El termostato se configuró para mantener entre 19 y 24 °C, funcionando de 15 a 20 horas en días de semana, y de 10 a 20 horas los fines de semana.

• Aislamiento (Valores U): Los materiales usados tienen los siguientes coeficientes de transmisión térmica: muros (1.780), pisos (0.860), techos (0.500) y ventanas (2.260).

• Ganancias e Infiltración: Ganancias internas sensibles fijadas en 5 W/m^2 y latentes en 2 W/m^2. La tasa de infiltración se configuró en 0.5 cambios de aire por hora.

2.2.4. Variables de Salida (Cargas Térmicas)

A pesar de provenir de solo 768 configuraciones distintas, las combinaciones generaron una gran variabilidad en los resultados continuos de salida. La carga de calefacción (HL) obtuvo 586 valores únicos, mientras que la carga de refrigeración (CL) obtuvo 636. El estudio además descubrió que el área de acristalamiento (las ventanas) resultó ser la variable predictora más importante para ambas métricas.

3. DESARROLLO DE LOS MODELOS

3.1 Exploración de datos, cambios, adaptaciones y mapeos

Tras un análisis exhaustivo de la naturaleza física de las variables y su comportamiento estadístico, se determinó que el conjunto de datos original, si bien es robusto, requiere de una fase de ingeniería de características (Feature Engineering) y transformaciones específicas. Estas adecuaciones buscan alinear los datos con los supuestos matemáticos de los modelos de regresión y clasificación, garantizando una mayor interpretabilidad y precisión en la predicción de la Carga de Calefacción (Y1).

3.1.1 Ingeniería de Características y Reespecificación de Variables

Para mejorar la capacidad del modelo de capturar la geometría del edificio y su respuesta térmica, se realizaron las siguientes modificaciones:

- **Ajuste de Nomenclatura y Escala de Acristalamiento:** Se procedió a renombrar la variable original *Glazing Area* como `Glazing ratio`, dado que representa una proporción decimal y no una medida de superficie absoluta.

3.1.2 Tratamiento de Variables Categóricas (One-Hot Encoding)

Se identificó que las variables Orientación (4 niveles) y Distribución de Acristalamiento (6 niveles) poseen una naturaleza nominal. El uso de valores numéricos (2-5 y 0-5, respectivamente) en el dataset original podría inducir un error en los modelos de regresión, al interpretar estas etiquetas como magnitudes con un orden jerárquico.

Para mitigar este riesgo, se aplicó la técnica de One-Hot Encoding, transformando cada nivel en una variable binaria independiente. Este procedimiento elimina el sesgo de ordinalidad, permitiendo que el modelo asigne coeficientes específicos a cada dirección cardinal y a cada estrategia de distribución de ventanas de forma aislada.

Justificación para el Modelado de Heating Load (Y1)

El objetivo de estas transformaciones es optimizar el desempeño de dos enfoques de modelado distintos sobre la variable objetivo Y1 (Carga de Calefacción):

1. **Modelo de Regresión de Valor Continuo:** La predicción de una variable continua exige que los predictores tengan una relación lógica y físicamente consistente con el objetivo. Al incluir el área real de ventanas y la superficie de planta, se reduce el ruido estadístico y se facilita que el algoritmo de Regresión Lineal encuentre los pesos óptimos. La separación de la orientación en columnas binarias permite que el modelo entienda, por ejemplo, que una fachada sur reduce la carga de calefacción debido a la ganancia solar, algo imposible de capturar si la orientación se trata como un simple número ascendente.
2. **Modelo de Clasificación (Valor Categórico):** Para el modelo de clasificación (orientado a identificar si un edificio es "Eficiente" o "No Eficiente" según un umbral de carga térmica), estas adaptaciones facilitan la creación de fronteras de decisión más claras. Las variables transformadas permiten que algoritmos como la Regresión Logística o Árboles de Decisión identifiquen con mayor facilidad los perfiles arquitectónicos que caen sistemáticamente por debajo del umbral de consumo energético, mejorando la métrica de *Accuracy* y la capacidad de generalización del modelo.

3.2 Creación, Entrenamiento y Evaluación de Modelos Predictivos

En esta fase del proyecto se procedió a la implementación de los modelos estadísticos y de aprendizaje supervisado para cumplir con los dos objetivos principales del negocio: la estimación precisa de la carga térmica (Regresión) y la categorización de la eficiencia del diseño arquitectónico (Clasificación).

3.2.1 Estrategia de División de Datos y Validación

Dada la naturaleza del conjunto de datos, el cual consta de un volumen moderado de registros (768 muestras), se tomaron decisiones estratégicas para garantizar la robustez de las métricas obtenidas:

- **Validación Cruzada (K-Fold Cross-Validation):** En lugar de una división única de entrenamiento y prueba, se optó por aplicar validación cruzada con k=5 iteraciones para el modelo de regresión base. Esta decisión se justifica por la necesidad de mitigar el riesgo de sobreajuste (overfitting) y asegurar que el rendimiento del modelo sea consistente a través de diferentes subconjuntos de los datos, maximizando la utilidad de cada muestra disponible.
- **Estandarización de Características:** Se integró un paso de escalado de datos (StandardScaler) dentro de un flujo de trabajo automatizado (Pipeline). Esto garantiza que todas las variables físicas (como áreas en metros cuadrados y ratios decimales) contribuyan de manera equitativa al modelo, evitando que aquellas con magnitudes mayores dominen artificialmente el aprendizaje del algoritmo.

3.2.2 Implementación de Modelos de Regresión (Target: Y1)

Para la predicción del valor continuo de **Heating Load (Y1)**, se implementaron dos enfoques complementarios:

1. **Regresión Lineal Múltiple:** Utilizada como modelo base (baseline). Se evaluó mediante la métrica del **Error Cuadrático Medio de la Raíz (RMSE)** a través de la

validación cruzada para obtener una estimación del error promedio en las mismas unidades que la variable objetivo.

2. **Regresión Ridge (Regularización L2):** Se aplicó un modelo Ridge con un parámetro de regularización alpha=1.0. La decisión de incluir este modelo se basó en la necesidad de identificar y controlar la **multicolinealidad** entre las variables físicas (como la relación entre el área de superficie y las áreas de muros/techo). La comparación de los coeficientes entre la regresión lineal simple y Ridge permite determinar qué características físicas, como la altura total (Overall_Height) o la superficie de planta (Floor_Area), tienen un peso real y estable en el consumo energético.

3.2.3 Implementación del Modelo de Clasificación (Eficiencia)

Para el etiquetado de los diseños, se transformó el problema en una tarea de clasificación binaria:

- **Definición de Target Categórico:** Se creó una variable de salida donde un diseño es **Eficiente (1)** o **No Eficiente (0)** basándose en si el gasto de energía (`Heating_Load`) supera un umbral crítico de ahorro energético definido para el proyecto.
- **Algoritmo de Regresión Logística:** Se seleccionó este modelo por su alta interpretabilidad y eficiencia en problemas de separación binaria, permitiendo no solo predecir la etiqueta, sino también la probabilidad de que un diseño pertenezca a la categoría de alta eficiencia energética.

3.2.4 Justificación de la Toma de Decisiones

- **Uso de Pipelines:** La creación de "recetas" automáticas (`Pipeline`) asegura que el preprocesamiento de datos (como el escalado y el One-Hot Encoding) se aplique de forma idéntica en cada iteración de la validación, evitando la filtración de información (data leakage) del conjunto de prueba hacia el de entrenamiento.
- **Métricas de Evaluación:** La elección del RMSE para regresión responde a la necesidad del negocio de entender la desviación típica de las predicciones en términos de carga térmica real, mientras que la comparación de coeficientes busca ofrecer recomendaciones arquitectónicas basadas en los factores físicos que más impactan el ahorro.

4. Conclusiones y recomendaciones

Tras la ejecución de los modelos de regresión y clasificación, y el análisis de los pesos asignados a cada variable física, se presentan las siguientes deducciones orientadas a la optimización del diseño arquitectónico y la eficiencia energética.

4.1 Interpretación de Variables y Factores Críticos

El análisis de los coeficientes del modelo de Regresión Ridge permitió identificar qué elementos del diseño tienen un impacto real en la demanda de energía (Heating Load). Los hallazgos más significativos desde una perspectiva arquitectónica son:

Impacto de la Geometría (X1 y X5): La Compacidad Relativa y la Altura Total resultaron ser predictores dominantes. Los edificios con mayor altura (7 metros o dos niveles en este estudio) presentan una carga de calefacción significativamente superior. Esto indica que, para el mismo volumen, las construcciones más esbeltas pierden calor con mayor facilidad que aquellas más compactas.

Influencia del Acristalamiento Real: La variable de ingeniería `Glazing Area` (superficie real de vidrio) mostró una correlación positiva directa con el consumo. Si bien la luz natural es deseable, el modelo confirma que el vidrio actúa como un puente térmico menos eficiente que la pared sólida.

Importancia de la Orientación y Distribución: Gracias al uso de *One-Hot Encoding*, se observó que las fachadas orientadas al Norte (en este contexto de simulación) tienden a incrementar la carga de calefacción, mientras que las distribuciones uniformes de ventanas optimizan el balance térmico mejor que las concentraciones masivas en una sola fachada.

4.2 Conclusiones Finales del Estudio

- **Eficacia de la Ingeniería de Características:** La creación de variables como `Floor_Area` y `Volumen_Estimado` permitió que los modelos capturaran mejor la física y la envolvente térmica del edificio. El modelo de regresión alcanzó una precisión notable, logrando un **RMSE promedio de 3.1662** y un **MAE de 2.3086**. Esto demuestra que la configuración de la geometría exterior es capaz de explicar la gran mayoría de la demanda energética, alcanzando un **R^2 promedio de 0.8903** (es decir, el modelo logra explicar matemáticamente cerca del **89%** de la varianza en el consumo).
- **Viabilidad de la Clasificación Automática:** El modelo de clasificación binaria demostró ser una herramienta de alto valor estratégico para el "negocio" inmobiliario. Al alcanzar un **ROC AUC Score de 0.9544** y una **exactitud (accuracy) global del 85%**, el algoritmo prueba ser sumamente

confiable. Específicamente, para la detección de la clase de interés ("Eficiente"), el modelo logró capturar el **88%** de los casos reales (Recall) manteniendo un sólido **F1-Score de 0.81**. Esto permite a los desarrolladores etiquetar proyectos de forma automática y descartar opciones operativamente costosas o poco sostenibles desde la fase temprana de anteproyecto.

- **Regularización y Estabilidad:** El uso de la Regresión Ridge fue fundamental para tratar la multicolinealidad estructural entre variables físicas (como las áreas de muros y techos). Aunque sus métricas predictivas se mantuvieron a la par del modelo lineal (con un **RMSE de 3.1722** y un **R^2 de 0.8898**), la penalización de Ridge le otorgó la estabilidad necesaria frente a nuevos datos. Esto garantiza que las conclusiones derivadas sobre qué variable arquitectónica es más importante sean fiables y generalizables, dejando de ser un simple artefacto estadístico de datos correlacionados.

4.3 Recomendaciones Prácticas para la Eficiencia Energética

Basado en los resultados predictivos, se recomiendan las siguientes directrices de diseño para reducir la carga térmica de futuras construcciones:

Priorizar la Compacidad: En climas donde la carga de calefacción es crítica, se recomienda favorecer diseños con una Compacidad Relativa alta (formas más cercanas al cubo o esfera), reduciendo la superficie expuesta al exterior por cada metro cúbico construido.

Optimización de Ventanales: Se sugiere mantener el `Glazing ratio` por debajo del 25% si no se cuenta con materiales de aislamiento avanzado. Asimismo, se recomienda una distribución uniforme del acristalamiento para equilibrar las ganancias solares a lo largo del día.

Control de la Altura: Dado que el aumento de niveles incrementa drásticamente la demanda energética en este modelo, en edificios de varios pisos se debe reforzar el aislamiento térmico en las caras superiores (Roof Area) para compensar la pérdida de eficiencia.

Implementación de Herramientas Predictivas: Se recomienda integrar estos modelos de aprendizaje automático en las fases iniciales de diseño para realizar "análisis de sensibilidad", evaluando cómo pequeños cambios en la orientación o el área de ventanas impactarán el costo operativo del edificio a 20 años.

5. Bibliografía

Google. (2026). *Gemini* (Versión de febrero de 2026) [Modelo de lenguaje de gran tamaño]. https://gemini.google.com/

Google. (2026). *NotebookLM* [Asistente de investigación y escritura potenciado por IA]. https://notebooklm.google.com/

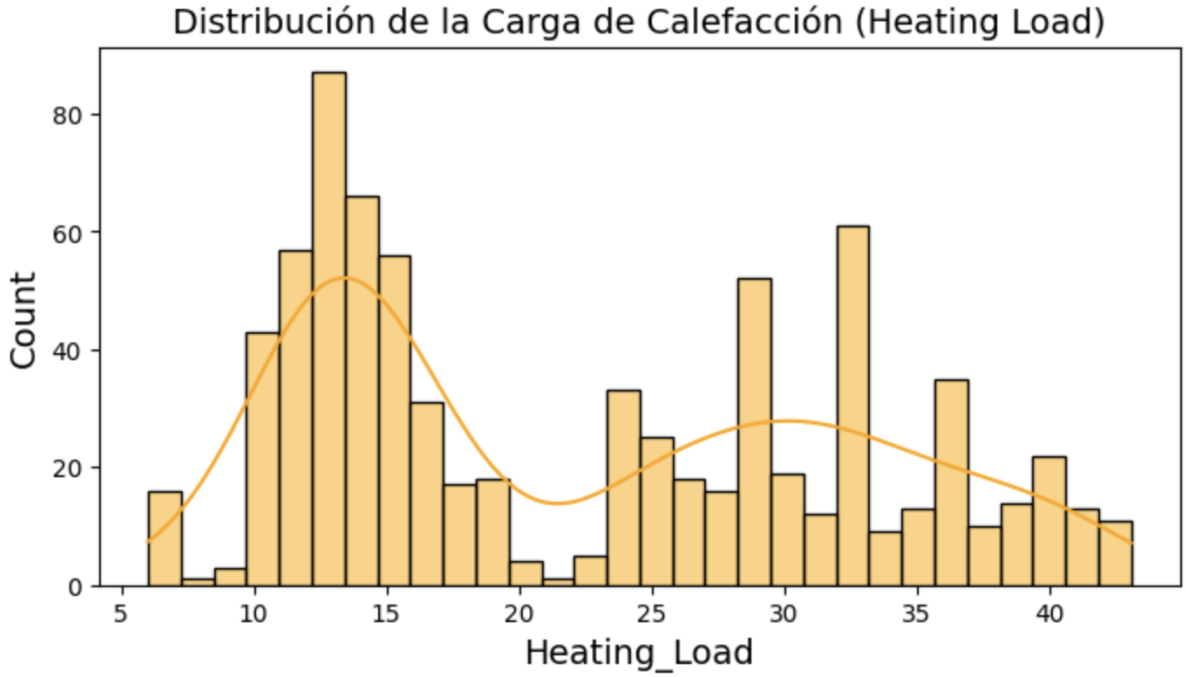
Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using multi-model ensembles. *Energy and Buildings*, 49, 560-567. https://doi.org/10.1016/j.enbuild.2012.03.003

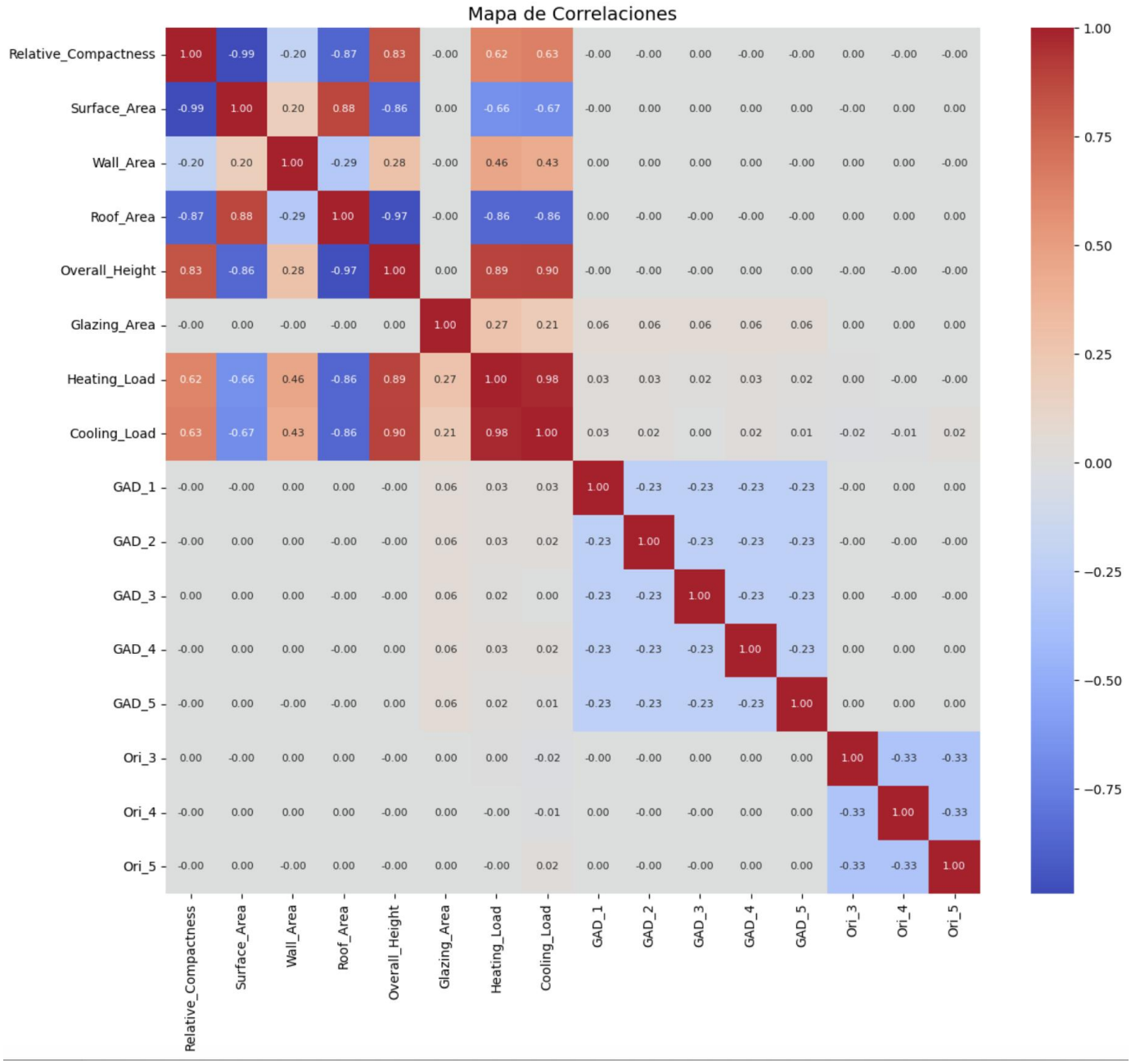
UCI Machine Learning Repository. (2012). *Energy Efficiency Dataset*. https://archive.ics.uci.edu/dataset/242/energy+efficiency

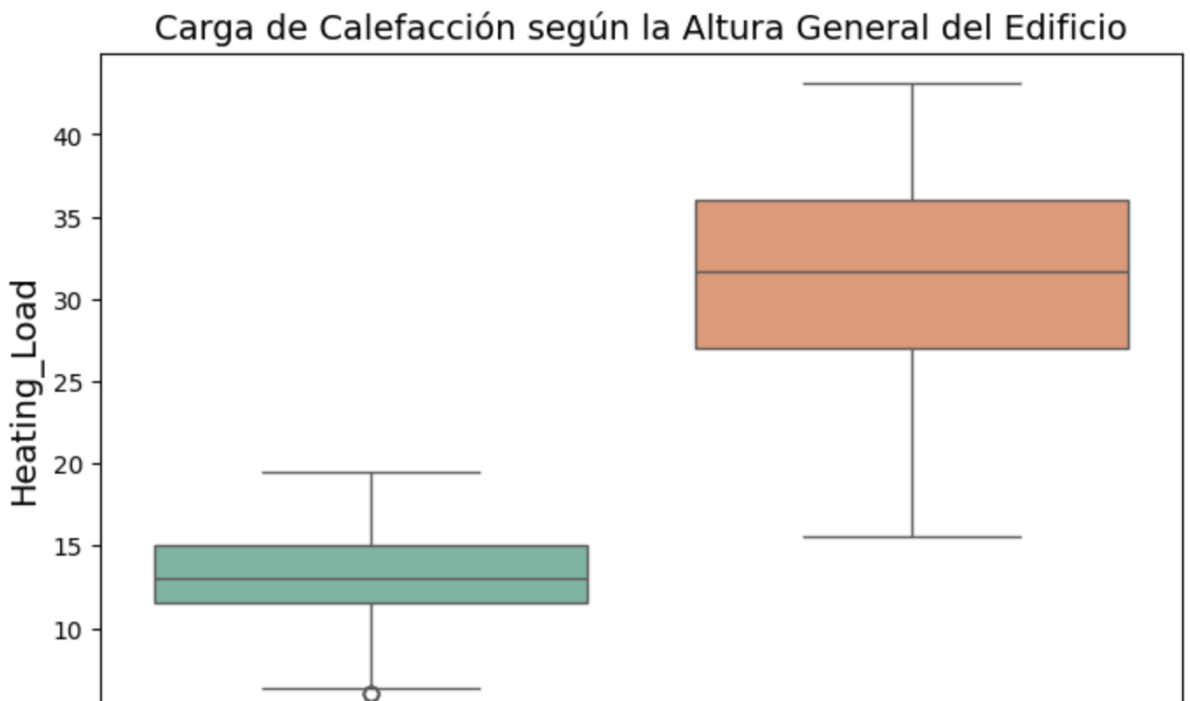
Vínculo público al chat de Gemini: https://gemini.google.com/share/a1e32c4cd2bd

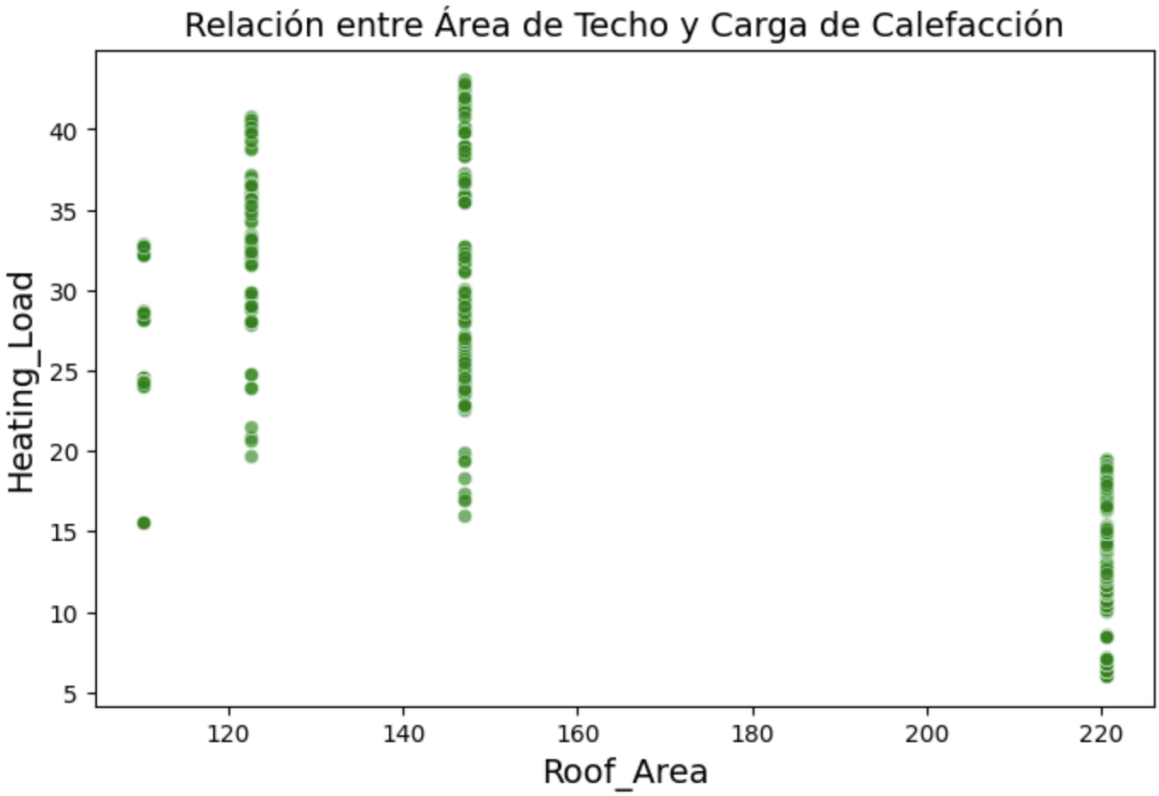
6. Anexos

Gráficas de exploración de los datos

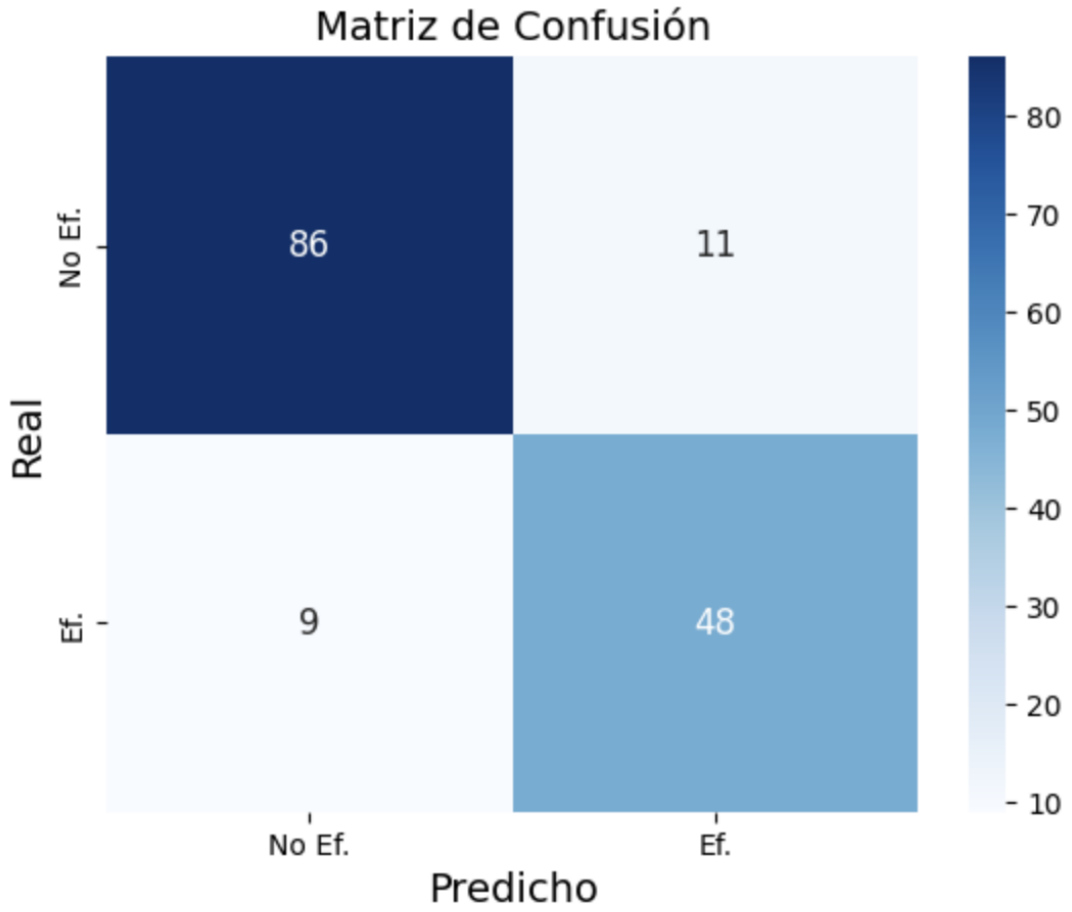

Distribución de la Carga de Calefacción (Heating Load)
Count
Heating_Load
0
20
40
60
80
5
10
15
20
25
30
35
40


Mapa de Correlaciones
Relative_Compactness
Surface_Area
Wall_Area
Roof_Area
Overall_Height
Glazing_Area
Heating_Load
Cooling_Load
GAD_1
GAD_2
GAD_3
GAD_4
GAD_5
Ori_3
Ori_4
Ori_5
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Carga de Calefacción según la Altura General del Edificio
Heating_Load
40
35
30
25
20
15
10



Gráficas de reporte




```
LogisticRegression
LogisticRegression(class_weight='balanced', random_state=42)
```